Redes complexas para classificação de dados via
conformidade de padrão, caracterização de
importância e otimização estrutural

Murillo Guimarães Carneiro

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Murillo Guimarães Carneiro

**Redes complexas para classificação de dados via
conformidade de padrão, caracterização de importância e
otimização estrutural**

Tese apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Doutor em Ciências – Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientador: Prof. Dr. Zhao Liang

USP – São Carlos
Janeiro de 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

C634r Carneiro, Murillo Guimarães
Redes complexas para classificação de dados via
conformidade de padrão, caracterização de importância
e otimização estrutural / Murillo Guimarães Carneiro;
orientador Zhao Liang. - São Carlos - SP, 2017.
161 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional)
- Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2017.

1. Redes complexas. 2. aprendizado de máquina.
3. classificação de dados. 4. classificação por
conformidade de padrão. 5. classificação por
importância. 6. otimização de redes. I. Liang, Zhao,
orient. II. Título.

Murillo Guimarães Carneiro

**Data classification in complex networks via pattern
conformation, data importance and structural optimization**

Doctoral dissertation submitted to the Instituto de
Ciências Matemáticas e de Computação – ICMC-
USP, in partial fulfillment of the requirements for the
degree of the Doctorate Program in Computer Science
and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and
Computational Mathematics

Advisor: Prof. Dr. Zhao Liang

USP – São Carlos
January 2017

*Aos meus amados pais, Carmelinda e Sérgio,
e aos seus (ainda mais) amados pais,
Celina (in memoriam) e Samuel Guimarães (in memoriam),
Geracina (in memoriam) e José Carneiro.*

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por iluminar os meus passos durante esse período tão árduo, de tanto trabalho e comprometimento, porém de inúmeras bênçãos derramadas. Agradeço a Ele por todas as pessoas que passaram pela minha vida nesse período, pelos momentos especiais e pelas conquistas alcançadas. Muito obrigado por tudo, Pai! E obrigado pela intercessão de sempre, Mãe.

Gostaria de expressar a minha mais sincera gratidão ao Prof. Zhao Liang, por sua orientação consistente e seu suporte tanto nos meus estudos quanto na minha vida pessoal. Sua orientação e sua habilidade para inovar contribuíram muito na minha formação e no meu crescimento como pesquisador. Obrigado por compartilhar comigo ideias tão inspiradoras. Espero que este seja apenas o início de muitas colaborações em conjunto.

Também gostaria de deixar aqui os meus sinceros agradecimentos a dois professores que muito me ajudaram a chegar até aqui. Ao Prof. João Luís Garcia Rosa, agradeço pelo período em que me orientou, no qual os seus valiosos conselhos e o seu zelo pela escrita contribuíram muito para a minha formação profissional. Ao Prof. Yaochu Jin, agradeço pela recepção que tive na Universidade de *Surrey*, por me conceder um ambiente de trabalho extremamente produtivo e pelas valiosas sugestões e contribuições na minha pesquisa.

Acredito que nenhuma das grandes experiências que vivi durante esse período seria possível sem o auxílio financeiro que recebi da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Sou muito agradecido à FAPESP pela minha bolsa de estudos e suporte oferecido durante o período em que estive em São Carlos. Também agradeço à CAPES, que me forneceu os recursos necessários para desenvolver parte do meu trabalho no exterior, período que concorreu muito para a minha qualificação enquanto docente e pesquisador.

Gostaria de expressar minha gratidão à Universidade Federal de Uberlândia (UFU), especificamente à Faculdade de Computação, unidade da qual faço parte. Muito obrigado pelo período de afastamento que me foi concedido no qual pude me dedicar integralmente à pesquisa, realizar doutorado sanduíche, produzir artigos e escrever a tese. Espero poder retribuir todo esse incentivo com ainda mais trabalho e dedicação, levando o nome da UFU cada vez mais à frente. Obrigado aos colegas professores e aos técnicos pela amizade, e palavras de incentivo.

Muito obrigado ao Instituto de Ciências Matemáticas e de Computação (ICMC), pela oportunidade de desenvolver meu trabalho aqui, em um ambiente repleto de grandes doutores. Agradeço aos Profs. Rodrigo Mello, Alneu Lopes, Zhao Liang, Roseli Romero, Sandra Aluisio

e Gustavo Batista, pelas disciplinas que cursei e que muito me motivaram a buscar ainda mais conhecimento e crescimento nas áreas de redes complexas e aprendizado de máquina. Além da formação científica, tive o privilégio de participar de várias palestras oferecidas pelo ICMC em que se debateram a educação, o papel da universidade, práticas pedagógicas, etc, as quais muito contribuíram para a minha formação como docente. Agradeço aos Profs. Zhao Liang, Thiago Pardo e Maria Cristina de Oliveira, pelos ensinamentos durante o período em que fui estagiário do Programa de Aperfeiçoamento do Ensino (PAE). Agradeço aos servidores administrativos, que, com muita atenção e competência, responderam de maneira eficiente às minhas dúvidas e solicitações.

Deixo o meu agradecimento aos membros da banca de defesa, Profs. Elbert Macau, Fabrício Breve, Renato Tinós e Jo Ueyama, pelos comentários e sugestões pertinentes proferidos durante a arguição.

Quando comecei o doutorado, eu estava namorando. Quando conclui, casado. Obrigado, Luana Carneiro. Obrigado por cuidar de mim durante esse período, pelo seu apoio e por trazer tanta alegria para os meus dias.

Agradeço aos meus pais, Sérgio e Carmelinda. Quanta humildade, quanto apoio, quanto carinho, quanto amor! "Obrigado" parece uma palavra pequena e simples demais quando penso em tudo o que fizeram e fazem na minha vida. Agradeço também ao meu grande amigo e irmão, Arthur, pelas nossas conversas, pelo seu apoio e palavras de incentivo.

Agradeço de maneira muito especial às minhas orientadoras do mestrado e da graduação, Profs. Gina M. B. Oliveira e Mônica S. Pais, respectivamente. Se hoje sou professor e me sinto tão realizado nessa carreira, certamente, elas possuem grande contribuição nisso.

Aos amigos que conheci em São Carlos-SP, também só posso dizer muito obrigado. A minha adaptação à cidade foi mais fácil porque conheci pessoas maravilhosas que participaram da minha vida e muito me ensinaram durante o período em que aí vivi. Quero deixar registrados meus sinceros agradecimentos a todos os amigos do Ministério Universidades Renovadas de São Carlos, aos amigos do Laboratório de Computação Bio-inspirada (Biocom) e aos demais amigos que fiz no ICMC e na Universidade de São Paulo (USP).

Sou muito grato a todos os meus familiares, especialmente, ao meu avô José Carneiro. Foi sempre revigorante visitá-lo em Ipameri-GO e passar várias tardes juntos conversando sobre futebol, clima e outras cidades que ele conheceu quando jovem. Também agradeço muito aos meus primos, tios e amigos de Ipameri e região. O carinho deles e a compreensão da minha ausência, em razão dos meus estudos, sempre me motivaram a trabalhar duro para um dia poder estar mais próximo de todos novamente. Obrigado pelas mensagens positivas e repletas de carinho, especialmente, durante o período em que estive em Inglaterra. Também deixo aqui meu muito obrigado aos amigos de Uberlândia-MG e de Guildford-UK.

*“ A fé e a razão constituem como que
duas asas pelas quais o espírito humano
se eleva para a contemplação da verdade.*

*Foi Deus quem colocou no coração
do homem o desejo de conhecer a verdade
e, em última análise, de O conhecer a Ele,
para que, conhecendo-O e amando-O,
possa chegar também à verdade plena
sobre si próprio.”*

Papa João Paulo II
Na Carta Encíclica *Fides et Ratio*

RESUMO

CARNEIRO, M. G.. **Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural**. 2017. 161 f. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

A classificação é uma tarefa do aprendizado de máquina e mineração de dados, na qual um classificador é treinado sobre um conjunto de dados rotulados de forma que as classes de novos itens de dados possam ser preditas. Tradicionalmente, técnicas de classificação trabalham por definir fronteiras de decisão no espaço de dados considerando os atributos físicos do conjunto de treinamento e uma nova instância é classificada verificando sua posição relativa a tais fronteiras. Essa maneira de realizar a classificação, essencialmente baseada nos atributos físicos dos dados, impossibilita que as técnicas tradicionais sejam capazes de capturar relações semânticas existentes entre os dados, como, por exemplo, a formação de padrão. Por outro lado, o uso de redes complexas tem se apresentado como um caminho promissor para capturar relações espaciais, topológicas e funcionais dos dados, uma vez que a abstração da rede unifica a estrutura, a dinâmica e as funções do sistema representado. Dessa forma, o principal objetivo desta tese é o desenvolvimento de métodos e heurísticas baseadas em teorias de redes complexas para a classificação de dados. As principais contribuições envolvem os conceitos de conformidade de padrão, caracterização de importância e otimização estrutural de redes. Para a conformidade de padrão, onde medidas de redes complexas são usadas para estimar a concordância de um item de teste com a formação de padrão dos dados, é apresentada uma técnica híbrida simples pela qual associações físicas e topológicas são produzidas a partir da mesma rede. Para a caracterização de importância, é apresentada uma técnica que considera a importância individual dos itens de dado para determinar o rótulo de um item de teste. O conceito de importância aqui é definido em termos do PageRank, algoritmo usado na *engine* de busca do Google para definir a importância de páginas da web. Para a otimização estrutural de redes, é apresentado um *framework* bioinspirado capaz de construir a rede enquanto otimiza uma função de qualidade orientada à tarefa, como, por exemplo, classificação, redução de dimensionalidade, etc. A última investigação apresentada no documento explora a representação baseada em grafo e sua habilidade para detectar classes de distribuições arbitrárias na tarefa de difusão de papéis semânticos. Vários experimentos em bases de dados artificiais e reais, além de comparações com técnicas bastante usadas na literatura, são fornecidos em todas as investigações. Em suma, os resultados obtidos demonstram que as vantagens e novos conceitos propiciados pelo uso de redes se configuram em contribuições relevantes para as áreas de classificação, sistemas de aprendizado e redes complexas.

Palavras-chave: Redes complexas, aprendizado de máquina, classificação de dados, classificação por conformidade de padrão, classificação por importância, otimização de redes.

ABSTRACT

CARNEIRO, M. G.. **Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural**. 2017. 161 f. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Data classification is a machine learning and data mining task in which a classifier is trained over a set of labeled data instances in such a way that the labels of new instances can be predicted. Traditionally, classification techniques define decision boundaries in the data space according to the physical features of a training set and a new data item is classified by verifying its relative position to the boundaries. Such kind of classification, which is only based on the physical attributes of the data, makes traditional techniques unable to detect semantic relationship existing among the data such as the pattern formation, for instance. On the other hand, recent works have shown the use of complex networks is a promising way to capture spatial, topological and functional relationships of the data, as the network representation unifies structure, dynamic and functions of the networked system. In this thesis, the main objective is the development of methods and heuristics based on complex networks for data classification. The main contributions comprise the concepts of pattern conformation, data importance and network structural optimization. For pattern conformation, in which complex networks are employed to estimate the membership of a test item according to the data formation pattern, we present, in this thesis, a simple hybrid technique where physical and topological associations are produced from the same network. For data importance, we present a technique which considers the individual importance of the data items in order to determine the label of a given test item. The concept of importance here is derived from PageRank formulation, the ranking measure behind the Google's search engine used to calculate the importance of webpages. For network structural optimization, we present a bioinspired framework, which is able to build up the network while optimizing a task-oriented quality function such as classification, dimension reduction, etc. The last investigation presented in this thesis exploits the graph representation and its ability to detect classes of arbitrary distributions for the task of semantic role diffusion. In all investigations, a wide range of experiments in artificial and real-world data sets, and many comparisons with well-known and widely used techniques are also presented. In summary, the experimental results reveal that the advantages and new concepts provided by the use of networks represent relevant contributions to the areas of classification, learning systems and complex networks.

Keywords: Complex networks, machine learning, data classification, classification by pattern conformation, classification by data importance, network structural optimization.

LISTA DE ILUSTRAÇÕES

Figura 1 – Ilustração da diferença entre classificação computacional e cerebral. (a) resultado de classificação considerando apenas os atributos físicos dos dados de entrada. Nesse caso, a instância de teste y - uma porta - é associada à classe porta (D), porque é similar a d_1 e d_3 , porém é fisicamente diferente de outros objetos. (b) a classificação cerebral considera a relação organizacional entre os itens de dados, o que, por vez, organiza as instâncias de dados em um sistema repleto de sentido: y pode ser associado a uma classe carro C_2	5
Figura 2 – Análise do processo de classificação de técnicas tradicionais. (a) Base de dados artificial composta por duas classes denotadas por itens de dados no formato de círculos e quadrados, enquanto itens de teste são representados na forma de triângulo; (b) k vizinhos mais próximos classifica os itens de teste verificando o rótulo dos seus k vizinhos mais próximos (circulados na figura); (c) Máquina de vetores de suporte levam em conta os vetores de suporte (circulados na figura) para aproximar cada classe com um fecho convexo e encontrar o melhor hiperplano de separação entre as classes, o qual é usado como fronteira de decisão para classificar os itens de teste. Técnicas de classificação tradicionais são incapazes de considerar a estrutura semântica dos dados e, conseqüentemente, os itens de teste são rotulados como pertencentes à classe de quadrados.	6
Figura 3 – Análise da classificação realizada pela técnica baseada em <i>importância</i> . (a) rede formada a partir da base de dados mostrada pela Fig. 2a; (b) explorando propriedade topológicas, funcionais e espaciais representadas nas conexões da rede, a técnica baseada em <i>importância</i> é capaz de detectar relações semânticas entre os dados e, conseqüentemente, classifica os itens de teste como pertencentes à classe de círculos.	8
Figura 4 – Rede “Caratê” compilada em Zachary (1977). As arestas caracterizam as relações de amizade de pessoas envolvidas no episódio do rompimento das colaborações entre dois instrutores de caratê de uma academia, representados pelos nós 1 e 34 da rede. Por conta da mudança de um dos instrutores para outra academia, vários alunos também se desligaram da academia para acompanhá-lo.	18
Figura 5 – Grau dos nós da rede “Caratê”.	19
Figura 6 – Coeficiente de agrupamento dos elementos da rede “Caratê”.	20

Figura 7 – Proximidade normalizada dos elementos da rede “Caratê”.	21
Figura 8 – Intermedialidade dos nós da rede “Caratê”.	22
Figura 9 – Exemplo ilustrativo dos problemas de convergência <i>spider-trap</i> e <i>dead-end</i>	24
Figura 10 – Valores de PageRank obtidos para os vértices da rede “Caratê” adaptada.	25
Figura 11 – Redes aleatórias geradas a partir do modelo <i>Erdős-Rényi</i> com número de vértices $n = 10$ e diferentes probabilidades.	26
Figura 12 – Distribuição de grau médio de redes aleatórias geradas a partir do modelo <i>Erdős-Rényi</i> . Os parâmetros usados foram $n = 5000$ e $p = 0.1$	27
Figura 13 – Simulação do modelo de Watts e Strogatz (1998) para geração de redes pequeno mundo com $n = 20$ vértices: (a) $p = 0$; (b) $p = 0.06$; (c) $p = 0.8$	28
Figura 14 – Exemplo de rede livre de escala obtida a partir do algoritmo proposto em Barabási e Albert (1999) com $n = 60$ e $m_0 = 5$	30
Figura 15 – Distribuição de grau médio de redes livres de escala geradas a partir do modelo <i>Barabasi-Albert</i> . Os parâmetros usados foram $n = 5000$ e $m_0 = 5$	30
Figura 16 – Rede modular com $n = 60$ vértices gerada a partir do modelo de Girvan e Newman (2002). Para gerar a rede, os vértices foram divididos em dois módulos de igual tamanho e todas as possíveis arestas do grafo simples foram consideradas com probabilidade $p_{in} = 0.3$ para nós do mesmo módulo e $p_{out} = 0.03$ caso contrário.	31
Figura 17 – Detecção de comunidades na rede “Caratê” (Fig. 4) utilizando o algoritmo proposto em Newman e Girvan (2004).	32
Figura 18 – Redes encontradas em Cancho e Solé (2003) mediante otimização estrutural: exponencial, estrela, livre de escala e densidade elevada, respectivamente. O eixo y é denotado pela entropia do grau, dada por $H = -\sum_{k=1}^{n-1} p_k \log p_k$, onde p_k é a frequência de vértices com grau k . Figura retirada de Cancho e Solé (2003).	33
Figura 19 – Ilustração da diferença básica entre os principais paradigmas do aprendizado de máquina: (a) classificação supervisionada; (b) classificação semissupervisionada; e (c) agrupamento de dados.	35
Figura 20 – Medida de <i>pureza</i> calculada sobre componentes com diferentes níveis de mistura de classes. Figura retirada de Bertini <i>et al.</i> (2011).	47
Figura 21 – Grafo intrínseco e grafo penalidade empregados pelo <i>framework</i> de redução de dimensionalidade. Retirado e adaptado de Yan <i>et al.</i> (2007).	52
Figura 22 – Representação e árvore sintática de uma sentença do PropBank-br. Retirado e adaptado de Alva-Manchego (2013).	58
Figura 23 – Rede <i>kAOG</i> formada a partir dos dados de treinamento da base de dados Art-triângulo.	71

Figura 24 – Rede k AOG obtida para a base de dados Art-reta I. A base de dados é composta por duas classes de itens de dados (cinza e preta), enquanto itens de teste são representados pela cor branca. Claramente, os itens de dado na cor cinza formam uma reta.	72
Figura 25 – Rede k AOG gerada para a base de dados Art-reta II, a qual é composta por itens de dados de duas classes (cores cinza e preta). O item de dado na cor branca representa um item de teste a ser classificado.	74
Figura 26 – Rede k AOG formada para as bases de dados Art-multiclasse. As bases de dados são compostas por itens de dados de três classes (cores cinza, preta e verde). O item de dado na cor branca representa um item de teste a ser classificado em ambas as bases: (a) um caso muito duvidoso para detectar alguma formação de padrão em relação à nova instância; (b) formação de padrão clara relacionada ao item de teste.	74
Figura 27 – Análise da variação de λ na técnica de alto nível proposta considerando duas bases de dados artificiais geradas com diferentes níveis de mistura entre as classes.	75
Figura 28 – Exemplo ilustrativo do funcionamento da abordagem de classificação proposta. (a) Uma rede $\mathcal{G}$ é formada a partir dos dados de treinamento $\mathcal{X}_{train}$; (b) o valor de <i>eficiência</i> de cada componente em $\mathcal{G}$ é calculado; (c) a <i>importância</i> de cada nó rotulado é calculada; (d) uma instância de teste y é conectada, temporariamente, a outros nós de acordo com a medida de <i>eficiência diferencial espaço-estrutural</i> ; (e) a <i>importância</i> de y para cada classe é calculada; (f) y é classificado para a classe na qual possui o maior valor de <i>importância</i> . Na figura, d significa a distância Euclidiana entre dois nós (dois itens de dados), e é a eficiência local de cada nós, ε representa a eficiência de cada componente, i denota a <i>importância</i> de um nó e $\mathcal{I}$ é a <i>importância</i> recebida por um item de teste em uma dada classe.	83
Figura 29 – Base de dados composta por itens de dados de duas classes (Azul/ $\circ$ e Vermelho/ $\square$) os quais apresentam padrões claros. (a) Branco/ $\diamond$ e Preto/ $\diamond$ são itens de teste que precisam ser classificados. (b) Preto/ $\triangle$ é uma instância de teste a ser classificada.	87
Figura 30 – Redes construídas a partir das bases de dados da Fig. 29 contendo padrões de Linha e Triângulo, respectivamente. O método de formação de rede k NN ($k = 3$) foi usado na base de dado Linha e k AOG na base de dados Triângulo. O valor de <i>importância</i> de alguns nós também são exibidos na figura.	88

Figura 31 – Fronteiras de decisão formadas por diferentes técnicas de classificação sobre as mesmas bases de dados mostradas na Fig. 29. Itens de dados Azul/◦ e Vermelho/◻ denotam dados de treinamento enquanto Branco/◊, Preto/◊ e Preto/△ indicam que y_1 , y_2 e y_3 são instâncias de teste a serem classificadas. As fronteiras de decisão consideram apenas características físicas dos dados, as quais tornam a classificação de y_2 fácil. Por outro lado, essas técnicas não são capazes de detectar a formação de padrão dos dados, o que torna difícil classificar corretamente y_1 e y_3 . Note-se que os parâmetros das técnicas tradicionais são otimizados. Mesmo assim, elas ainda são incapazes de classificar os itens de teste y_1 e y_3 na classe Azul/◦.	89
Figura 32 – Resultados de classificação da técnica proposta sobre as mesmas bases de dados apresentadas na Fig. 31. As instâncias de teste são classificadas de acordo com seus padrões correspondentes pela técnica baseada em <i>importância</i> . O valor de <i>importância</i> recebido pelos itens de teste em cada classe é mostrado nas subfiguras.	90
Figura 33 – Análise da abordagem proposta sobre as bases de dados mostradas na Fig. 29. Cada subfigura apresenta as decisões geradas pela técnica proposta com um valor específico do parâmetro γ	90
Figura 34 – Análise preditiva da técnica proposta em função dos parâmetros k e γ sobre bases de dados artificiais geradas com diferentes níveis de mistura entre as classes.	91
Figura 35 – Análise de consumo de tempo em função do número de instâncias na base de dados.	92
Figura 36 – Distribuição das classes da base de dados “Cardiac Images”. A figura mostra os primeiros três componentes obtidos a partir do algoritmo de análise de componente principal (PCA). As instâncias rotuladas são representadas por marcadores quadrados (cor verde) e círculos (cor vermelha), os quais denotam as classes de pacientes com coração anormal e normal, respectivamente. As instâncias não rotuladas são representadas por pequenos triângulos (cor azul).	96
Figura 37 – Redes exploradas por (a) PgRkAOG e (b) PgRkNN para conduzir o processo de classificação sobre a base de dados “Cardiac Images”.	98

Figura 38 – Um item de teste da base de dados “Cardiac Images” representado pelo marcador Preto/ $\triangle$. A instância de teste denota um caso de equívoco para a classificação tradicional, uma vez que é erroneamente classificada quando considerando apenas as características físicas. Por exemplo, na figura são mostrados os seis vizinhos mais próximos da nova instância, os quais são analisados pelo algoritmo k-NN para classificar o paciente na classe normal (Vermelha/ $\square$). Por outro lado, o item de teste é corretamente predito pela técnica proposta a qual leva em conta a classe na qual ele possui maior importância (0.065): classe anormal (Azul/ $\circ$).	99
Figura 39 – Exemplo de todas as categorias da base de dados ETH-80. Cada categoria é composta por dez objetos distintos.	100
Figura 40 – Exemplo de dois objetos pertencentes à categoria Maçã da base de dados ETH-80. Cada objeto é representado por meio de 41 imagens em diferentes ângulos e rotações.	100
Figura 41 – <i>Framework</i> geral para otimização estrutural no aprendizado baseado em rede. A ilustração enfatiza as duas etapas da técnica: otimização (também denominada etapa de treinamento) e teste.	104
Figura 42 – Exemplo ilustrativo sobre a conversão do vetor de probabilidade (manipulado por SL-PSO) para rede (na qual as propriedades estruturais são exploradas), e sobre as heurísticas de mapeamento propostas. (a) <i>MapLab</i> considera apenas vizinhos mais próximos de mesmo rótulo para criar a matriz <i>Map</i> ; (b) <i>MapAll</i> leva em conta itens de dados de todas as classes para criar a matriz <i>Map</i> . . .	108
Figura 43 – Bases de dados artificiais geradas com níveis de mistura distintos entre as classes: (a) 10%; (b) 25%; (c) 50%; (d) 75%; (e) 90%.	111
Figura 44 – Análises da convergência na etapa de otimização do <i>framework</i> proposto usando as heurísticas de mapeamento <i>MapLab</i> e <i>MapAll</i> . Cada subfigura exibe a acurácia média obtida por meio da função de qualidade f em função do número de iterações. Os resultados mostram que cem partículas e cem iterações, geralmente, são suficientes para assegurar a convergência de ambas heurísticas. Em termos de desempenho preditivo, a heurística de mapeamento <i>MapAll</i> é um pouco melhor.	115
Figura 45 – Comparação entre redes formadas por diferentes métodos: (a) rede obtida por kNN; (b) rede obtida por BIO considerando uma iteração; (c) rede obtida por BIO ao término do processo de otimização (após cem iterações). As informações extraídas a partir de um componente cujos vértices são os mesmos nas três redes enfatizam uma organização estrutural bastante distinta. Por exemplo, nas redes geradas em kNN, BIO-1gen e BIO-optimized, os nós mais importantes no componente analisado registram pontuação 0.047 (id. 33), 0.014 (id. 39) e 0.066 (id. 30), respectivamente.	120

Figura 46 – Rede otimizada obtida para a base de dados “Sonar”. Itens de dados no formato de círculos e quadrados representam as duas classes de dados existentes na base. Itens de teste representados circulos por pontilhados denotam casos de erro para as três técnicas tradicionais.	121
Figura 47 – Inserção temporária do item de teste y nas classes de dados do exemplo ilustrativo.	124
Figura 48 – Imagens da base de dados “AlphaDigits”.	126
Figura 49 – Erro médio de (a) GFHF e (b) LGC sobre as bases de dados PBbr-dar (esquerda), PBbr-fazer (centro) e PBbr-dizer (direita) com um argumento rotulado por papel semântico.	135
Figura 50 – Erro médio de (a) GFHF e (b) LGC sobre as bases de dados PBbr-dar (esquerda), PBbr-fazer (centro) e PBbr-dizer (direita) com 6, 16 e 10 argumentos rotulados, respectivamente.	136
Figura 51 – Erro médio de (a) GFHF e (b) LGC sobre as bases de dados PBbr-dar (esquerda), PBbr-fazer (centro) e PBbr-dizer (direita) com 9, 24 e 15 argumentos rotulados, respectivamente.	136
Figura 52 – Erro médio de GFHF e LGC nas bases de dados PBbr-dar (esquerda), PBbr-fazer (centro) e PBbr-dizer (direita) com diferentes números de argumentos previamente rotulados: (a) $\eta = 1$, (b) $\eta = 2$ e (c) $\eta = 3$	137

LISTA DE ALGORITMOS

Algoritmo 1 – Construção do grafo k -associados (kac)	46
Algoritmo 2 – Construção do grafo k -associados ótimo	48
Algoritmo 3 – Classificação baseada em <i>importância</i>	85
Algoritmo 4 – Framework para otimização estrutural de redes	106

LISTA DE TABELAS

Tabela 1 – Análise da distância média $\langle d \rangle$ e coeficiente de agrupamento médio C de redes regular ($k = 2$, $p = 0$), pequeno mundo ($p = 0.06$) e aleatória ($p = 0.8$). Para $p > 0$, os resultados são obtidos a partir da média de 20 redes geradas aleatoriamente.	28
Tabela 2 – Argumentos modificadores (ArgM) do PropBank.	57
Tabela 3 – Classes do PropBank-br em função do número de argumentos.	58
Tabela 4 – Resultados obtidos a partir das medidas de redes complexas usadas para classificação de alto nível da base de dados Art-triângulo. As medidas de redes são assortatividade, coeficiente de agrupamento e grau médio.	71
Tabela 5 – Predição final obtida para as bases de dados artificiais quando $\lambda = 0.2$ e $\lambda = 0.6$	72
Tabela 6 – Breve descrição das bases de dados reais: <i>Nome</i> representa o nome da base de dados, <i>#Inst.</i> é o número de instâncias, <i>#Attr.</i> é o número de atributos, <i>#Classes</i> é o número de classes e <i>Classe Maj.</i> denota o percentual de instâncias na classe majoritária.	76
Tabela 7 – Resultados comparativos obtidos por HL- k AOG, CART, k-NN e SVM sobre dez bases de dados reais. Cada célula na tabela é representa por “Acc $\pm$ Std” onde “Acc” e “Std” denotam, respectivamente, a média de acurácia e o desvio padrão em trinta execuções usando validação cruzada estratificada dez vezes. Em HL- k AOG, o resultado de classificação é obtido a partir do melhor valor entre $\lambda_1 = 0.2$ e $\lambda_2 = 0.6$	77
Tabela 8 – Ranque médio com o desempenho das técnicas em cada base de dados. . .	78
Tabela 9 – Breve descrição das bases de dados reais em termos do número de instâncias (<i>#Inst.</i>), número de atributos (<i>#Attr.</i>), e número de classes (<i>#Classes</i>).	92
Tabela 10 – Resultados comparativos obtidos por SVM, RF, k-NN, k-AOG, CART, PgRkNN e PgRkAOG sobre seis bases de dados reais. “Acc” e “Std.” denotam, respectivamente, a média de acurácia preditiva e o desvio padrão sobre trinta execuções usando validação cruzada dez vezes estratificada. Os resultados de SVM, k-NN e PgRkNN mostram os parâmetros ajustados a partir de seleção de modelo.	94
Tabela 11 – Ranque médio com o desempenho das técnicas em cada base de dados. . .	95

Tabela 12 – Resultados preditivos sobre as bases de dados “Cardiac Images” e “Cardiac Diagnosis”. SVM, RF, k-NN, k-AOG e CART denotam máquina de vetores de suporte, floresta aleatória, k-vizinho mais próximo, classificador do grafo k-associado ótimo e árvore de classificação e regressão, respectivamente. Os termos PgRkNN e PgRkAOG referem-se à abordagem proposta usando os métodos de formação de rede <i>k</i> NN e <i>k</i> AOG, respectivamente. Resultados de SVM, k-NN e PgRkNN mostram os parâmetros ajustados por meio de seleção de modelo. Os melhores resultados estão em negrito.	97
Tabela 13 – Resultados preditivos sobre as diferentes categorias que compõe a base de dados ETH-80. RWLP e PgRkNN denotam a técnica de classificação por facilidade de acesso usando caminhada aleatória (<i>random walk limiting probabilities</i>) e a técnica de classificação baseada no conceito de <i>importância</i> em redes complexas, respectivamente. Os resultados de ambas as técnicas mostram os parâmetros ajustados por meio de seleção de modelo.	101
Tabela 14 – Resultados obtidos pelo <i>framework</i> proposto (BIO) sobre as cinco bases de dados mostradas pela Fig. 43, caracterizadas por diferentes níveis de mistura entre as classes. BIO é comparado com métodos de construção da rede bastante usados na literatura, <i>k</i> NN e <i>k</i> AOG.	112
Tabela 15 – Descrição das bases de dados reais em termos do número de instâncias (<i>#Inst.</i>), atributos (<i>#Attr.</i>) e classes (<i>#Classes</i>).	113
Tabela 16 – Análise de convergência usando a heurística de mapeamento <i>MapLab</i> na etapa de otimização. Os resultados representam a acurácia média obtida mediante a função de qualidade <i>f</i> para uma, dez e cem iterações, respectivamente.	114
Tabela 17 – Análise de convergência usando a heurística de mapeamento <i>MapAll</i> na etapa de otimização. Os resultados representam a acurácia média obtida por meio da função de qualidade <i>f</i> para uma, dez e cem iterações, respectivamente. .	114
Tabela 18 – Acurácia preditiva obtida pelo <i>framework</i> proposto quando usando as heurísticas de mapeamento <i>MapLab</i> e <i>MapAll</i> na etapa de teste.	116
Tabela 19 – Acurácia preditiva do <i>framework</i> de otimização (BIO) em comparação métodos de formação da rede bem conhecidos (<i>k</i> NN e <i>k</i> AOG) e outro método estocástico (RM). As últimas linhas denotam os resultados do teste estatístico de Postos Sinalizados de Wilcoxon: (i) os símbolos “>” e “<” indicam uma rejeição da hipótese nula, onde o algoritmo correspondente à coluna é, respectivamente, melhor ou pior que o algoritmo correspondente à linha; (ii) e o símbolo “~” sugere que ambos os algoritmos desempenham igualmente bem, i.e., falha na rejeição da hipótese nula para um nível de confiança de 95%.	117

Tabela 20 – Acurácia preditiva do <i>framework</i> de otimização (BIO) em comparação com métodos de formação de rede bastante utilizados na literatura (<i>SkNN</i> e <i>MkNN</i>). Consulte-se a legenda da Tabela 19 para uma explicação detalhada sobre os resultados dos testes estatísticos.	118
Tabela 21 – Acurácia preditiva do <i>framework</i> de otimização (BIO) em comparação com técnicas de classificação tradicionais bem conhecidas (<i>k</i> -NN, NB e SVM). Uma explicação detalhada sobre os resultados dos testes estatísticos é fornecida na legenda da Tabela 19.	119
Tabela 22 – Resultados de classificação obtidos pelo <i>framework</i> de otimização no exemplo ilustrativo.	124
Tabela 23 – Parâmetros otimizados pelo <i>framework</i> no exemplo ilustrativo.	124
Tabela 24 – Classificação de alto nível por meio de medidas de redes complexas no exemplo ilustrativo. Valores calculados de proximidade, assortatividade e coeficiente de agrupamento para cada classe.	125
Tabela 25 – Breve descrição das bases de dados reais em termos do número de instâncias, número de atributos e número de classes.	125
Tabela 26 – Acurácia preditiva do classificador de regressão logística (<i>LR</i>), do <i>framework</i> com otimização da rede ($\mathcal{M}(f)$) e sem otimização da rede ($\mathcal{M}(kNN)$, com $k = q$)	127
Tabela 27 – Acurácia preditiva do <i>framework</i> de otimização usando diferentes composições da função de qualidade.	127
Tabela 28 – Acurácia preditiva usando o classificador ingênuo de Bayes (<i>NB</i>) e o <i>framework</i> de otimização. A probabilidade dos atributos foi assumida ser Gaussiana nos experimentos.	127
Tabela 29 – Breve descrição das bases de dados extraídas do PropBank-br em termos do número de argumentos e classes.	132
Tabela 30 – Resultados comparativos em termos de erro médio e desvio padrão (em trinta execuções) usando um item rotulado por classe ($\eta = 1$). O melhor resultado para cada base de dados é destacado.	133
Tabela 31 – Resultados comparativos em termos de erro médio e desvio padrão (em trinta execuções) usando $\eta = 2$ (o número de argumentos rotulados é mostrado ao lado do nome da base). O melhor resultado para cada base de dados é destacado.	134
Tabela 32 – Resultados comparativos em termos de erro médio e desvio padrão (em trinta execuções) usando $\eta = 3$ (o número de argumentos rotulados é mostrado ao lado do nome da base). O melhor resultado para cada base de dados é destacado.	134

LISTA DE ABREVIATURAS E SIGLAS

ϵ N	Rede vizinhança de raio ϵ
k -AOG	Classificador ótimo de Bayes usando grafo k associados ótimo
k -NN	Classificador k vizinhos mais próximos
k AOG	Rede k associados ótimo
k NN	Rede k vizinhos mais próximos
APS	Anotação de papéis semânticos
BIO	<i>Framework</i> bioinspirado para otimização estrutural de redes
CART	Classificador árvores de decisão
GFHF	Propagação de rótulos usando campos Gaussianos e funções harmônicas
LGC	Propagação de rótulos usando consistência local e global
LR	Classificador regressão logística
Mk NN	Rede k vizinhos mais próximos mútuos
MapAll	Heurística de mapeamento que considera todos os itens de dados
MapLab	Heurística de mapeamento que considera apenas itens de dados de mesma classe
MST	Rede árvore geradora mínima
NB	Classificador ingênuo de Bayes
PBbr	Recurso PropBank-br
PCA	Algoritmo de análise de componente principal
PSO	Algoritmo de otimização por enxame de partículas
RBF	Função de base radial, normal ou Gaussiana
RF	Classificador florestas aleatórias
Sk NN	Rede k vizinhos mais próximos simétricos
SL-PSO	Algoritmo de otimização por enxame de partículas baseado em teorias sociais
SVM	Classificador máquina de vetores de suporte

LISTA DE SÍMBOLOS

$\mathcal{G}$ — Rede

V — Conjunto de vértices

E — Conjunto de arestas

A — Matriz de adjacências da rede

W — Matriz de pesos da rede

v — Um vértice da rede

n — Número de vértices na rede

e — Uma aresta da rede

m — Número de arestas na rede

ω — Peso de uma aresta

$\langle k \rangle$ — Grau médio

r — Assortatividade

CC — Coeficiente de agrupamento

AC — Proximidade

PR — *PageRank*

$\mathcal{X}$ — Conjunto de dados

$\mathcal{X}_{train}$ — Conjunto de dados de treinamento

$\mathcal{X}_{valid}$ — Conjunto de dados de validação

$\mathcal{X}_{test}$ — Conjunto de dados de teste

$\mathcal{L}$ — Conjunto de classes

x — Item de dado

l — Classe de dado

n — Número de itens de dados de treinamento

z — Número de itens de dados de teste

$\mathcal{X}^{(l)}$ — Conjunto de dados rotulados

$\mathcal{X}^{(u)}$ — Conjunto de dados não rotulados

D — Matriz de distância

S — Matriz de similaridade

L — Laplaciano normalizado

$\mathcal{M}$ — Classificador híbrido

$\mathcal{C}$ — Classificador de baixo nível

$\mathcal{H}$ — Classificador de alto nível

ΔG — Variação de medidas de rede

$\mathcal{E}$ — Eficiência do componente

$\mathcal{I}$ — Classificador por importância

$\mathcal{F}$ — Eficiência diferencial espaço-estrutural

Map — Matriz de mapeamento

SUMÁRIO

1	INTRODUÇÃO	1
1.1	Motivação	3
1.2	Objetivos	9
1.3	Organização do trabalho	11
2	FUNDAMENTAÇÃO TEÓRICA E REVISÃO DA LITERATURA	13
2.1	Redes complexas	14
2.1.1	<i>Conceitos básicos</i>	14
2.1.2	<i>Medidas de redes complexas</i>	18
2.1.2.1	<i>Grau</i>	18
2.1.2.2	<i>Assortatividade</i>	19
2.1.2.3	<i>Coeficiente de agrupamento</i>	20
2.1.2.4	<i>Proximidade</i>	21
2.1.2.5	<i>Intermedialidade</i>	21
2.1.2.6	<i>Modularidade</i>	22
2.1.2.7	<i>PageRank</i>	23
2.1.3	<i>Modelos de redes complexas</i>	25
2.1.3.1	<i>Redes aleatórias</i>	25
2.1.3.2	<i>Redes pequeno mundo</i>	26
2.1.3.3	<i>Redes livres de escala</i>	28
2.1.3.4	<i>Redes modulares</i>	29
2.1.3.5	<i>Otimização estrutural de redes</i>	31
2.2	Aprendizado de máquina	33
2.2.1	<i>Aprendizado supervisionado</i>	35
2.2.2	<i>Aprendizado não supervisionado</i>	38
2.2.3	<i>Aprendizado semissupervisionado</i>	39
2.3	Aprendizado de máquina baseado em redes	40
2.3.1	<i>Métodos de formação de rede</i>	41
2.3.1.1	<i>Rede completamente conectada</i>	41
2.3.1.2	<i>Rede k vizinhos mais próximos</i>	41
2.3.1.3	<i>Rede vizinhança de raio ϵ</i>	42
2.3.1.4	<i>Outros métodos da literatura</i>	42
2.3.2	<i>Classificação supervisionada baseada em grafos</i>	43

2.3.2.1	<i>Grafo k-associados ótimo para classificação de dados</i>	45
2.3.2.2	<i>Classificação de alto nível</i>	49
2.3.2.3	<i>Classificação de dados por facilidade de acesso usando passeio aleatório</i>	50
2.3.2.4	<i>Redução de dimensionalidade para classificação de dados</i>	51
2.3.3	<i>Classificação semissupervisionada baseada em grafos</i>	52
2.4	<i>Anotação automática de papéis semânticos</i>	54
2.4.1	<i>Papéis semânticos</i>	55
2.4.2	<i>PropBank-br e outros recursos</i>	56
2.4.3	<i>Anotação de papéis semânticos</i>	59
2.4.3.1	<i>Revisão dos principais trabalhos da literatura</i>	60
2.5	<i>Considerações finais</i>	62
3	CLASSIFICAÇÃO DE ALTO NÍVEL USANDO GRAFO K-ASSOCIADOS ÓTIMO	65
3.1	<i>Descrição da técnica</i>	65
3.1.1	<i>Classificação de alto nível sem parâmetros</i>	66
3.1.2	<i>Variação de medidas de redes complexas</i>	67
3.1.3	<i>Medida de eficiência do componente</i>	68
3.1.4	<i>Análise de complexidade</i>	69
3.2	<i>Resultados experimentais</i>	70
3.2.1	<i>Bases de dados artificiais</i>	70
3.2.2	<i>Bases de dados reais</i>	76
3.3	<i>Considerações finais</i>	78
4	CLASSIFICAÇÃO ORGANIZACIONAL DE DADOS BASEADA NO CONCEITO DE IMPORTÂNCIA EM REDES COMPLEXAS	81
4.1	<i>Descrição da técnica</i>	82
4.1.1	<i>Classificação baseada em importância</i>	83
4.1.2	<i>Algoritmo e análise de complexidade</i>	85
4.2	<i>Resultados experimentais</i>	86
4.2.1	<i>Bases de dados artificiais</i>	87
4.2.1.1	<i>Análise de parâmetros</i>	88
4.2.1.2	<i>Análise de tempo computacional</i>	89
4.2.2	<i>Bases de dados reais</i>	92
4.2.3	<i>Estudo de caso: detecção de doenças cardíacas</i>	95
4.3	<i>Resultados experimentais com outra medida de similaridade</i>	99
4.3.1	<i>Estudo de caso: reconhecimento de padrões invariantes</i>	99
4.4	<i>Considerações finais</i>	101

5	COMPUTAÇÃO BIOINSPIRADA PARA OTIMIZAÇÃO ESTRUTURAL DE REDES E SUAS APLICAÇÕES	103
5.1	Descrição geral da técnica	103
5.1.1	<i>Algoritmo de otimização por enxame de partículas</i>	104
5.1.2	<i>Framework de otimização estrutural bioinspirado</i>	105
5.1.2.1	<i>Representação geral</i>	105
5.1.2.2	<i>Heurísticas de mapeamento</i>	106
5.1.2.3	<i>Função de qualidade</i>	108
5.1.2.4	<i>Análise de complexidade</i>	109
5.2	<i>Aplicação do framework para a classificação baseada em importância</i>	109
5.2.1	<i>Função de qualidade</i>	110
5.2.2	<i>Resultados em bases de dados artificiais</i>	110
5.2.3	<i>Resultados em bases de dados reais</i>	111
5.2.3.1	<i>Experimentos</i>	112
5.2.3.2	<i>Heurísticas de mapeamento e análises de convergência</i>	113
5.2.3.3	<i>Comparação com métodos de formação de dígrafos</i>	116
5.2.3.4	<i>Comparação com métodos de formação de rede não direcionada</i>	118
5.2.3.5	<i>Comparação com classificadores tradicionais</i>	118
5.2.3.6	<i>Análises das redes geradas</i>	119
5.3	<i>Aplicação do framework para classificação de alto nível</i>	122
5.3.1	<i>Função de qualidade e representação adicional</i>	122
5.3.1.1	<i>Medidas de redes complexas</i>	123
5.3.2	<i>Resultados experimentais</i>	123
5.3.2.1	<i>Exemplo ilustrativo</i>	123
5.3.2.2	<i>Resultados em bases de dados reais</i>	124
5.4	<i>Considerações finais</i>	128
6	DIFUSÃO DE PAPÉIS SEMÂNTICOS BASEADA EM GRAFOS	129
6.1	Descrição do framework	129
6.1.1	<i>Construção da rede</i>	130
6.1.2	<i>Geração da matriz ponderada e do Laplaciano do grafo</i>	130
6.1.3	<i>Propagação de rótulos</i>	131
6.2	Resultados experimentais	131
6.2.1	<i>Bases de dados</i>	132
6.2.2	<i>Parâmetros e baseline</i>	132
6.2.3	<i>Resultados e discussões</i>	133
6.3	<i>Considerações finais</i>	137
7	CONCLUSÕES	139
7.1	Principais contribuições	139

7.2	Oportunidades de trabalhos futuros	141
7.3	Lista de artigos gerados	143
REFERÊNCIAS		145

INTRODUÇÃO

Com o acúmulo de dados cada vez mais facilitado pelas tecnologias desenvolvidas para coleta e armazenamento, a principal preocupação das organizações é a extração de informações relevantes dessas bases de dados. A fim de lidar com as dificuldades iminentes para alcançar tal objetivo, que incluem o grande volume e variedade dos dados, o *Aprendizado de Máquina* visa ao desenvolvimento de métodos computacionais capazes de “aprender” a partir dessas experiências acumuladas (MITCHELL, 1997; BISHOP, 2006; DUDA; HART; STORK, 2007). Dessa forma, as técnicas de aprendizado de máquina atuam na geração automática de modelos capazes de detectar o conhecimento inerente às bases de dados, na tentativa de replicar eficientemente o processo de análise conduzido por um especialista humano.

Devido à extensa lista de aplicações, compreendendo desde diagnóstico de doenças, análises de dados biológicos e reconhecimento de padrões até trajetórias espaciais (AGGARWAL, 2014; LIBBRECHT; NOBLE, 2015; BISHOP, 2006; LI; HUANG; PENG, 2015), o aprendizado de máquina desperta interesse crescente das mais diversas áreas, incluindo matemática, física, biologia e engenharia. Embora existam muitas abordagens diferentes para tarefas de aprendizado, a divisão tradicional considera dois paradigmas (MITCHELL, 1997). No *aprendizado supervisionado*, o objetivo é induzir um modelo preditivo usando dados com informações previamente conhecidas, também denominadas *rótulo*, *classe* ou *saída*. Assim, o processo de aprendizado consiste em construir uma função de mapeamento dos dados de entrada para suas respectivas saídas de modo que dados não rotulados possam ter seus rótulos posteriormente preditos usando essa mesma função. Por outro lado, no *aprendizado não supervisionado*, não há existência de rótulos ou classes e o objetivo normalmente está associado ao agrupamento de dados baseado em algum critério de similaridade.

Adicionalmente, o *aprendizado semissupervisionado* é considerado um paradigma recente e intermediário aos dois anteriores. Ao contrário do aprendizado supervisionado, no qual normalmente uma quantidade razoável de dados rotulados é exigida para um treinamento de

qualidade, em muitas aplicações reais, é possível obter apenas conhecimento parcial sobre os dados. Por exemplo, em uma rede social de grande escala, o que se conhece são algumas relações de amizade; no domínio da biologia, são conhecidas apenas as funções de alguns genes. Dessa forma, o processo de predição, no aprendizado semissupervisionado, é caracterizado pelo uso de informações tanto dos poucos dados rotulados disponíveis quanto dos muitos dados não rotulados existentes (CHAPELLE; SCHOLKOPF; ZIEN, 2006).

Se a existência ou não de rótulos é usada para categorizar os paradigmas de aprendizado de máquina, os valores que tais rótulos podem assumir também determinam o nome dado à tarefa de predição. Se os valores assumidos pelos rótulos são discretos, a tarefa é denominada *classificação* e a respectiva função de mapeamento é chamada classificador, caso sejam contínuos, ela é chamada *regressão* e a função é denominada regressor. As investigações apresentadas neste documento contemplam a tarefa de classificação de dados (BISHOP, 2006). Embora haja muitas técnicas de classificação conhecidas na literatura, o interesse deste trabalho é a investigação e o desenvolvimento de métodos e heurísticas baseadas em redes complexas para a tarefa de classificação.

O uso de redes¹ tem se constituído em uma ferramenta robusta para a modelagem de sistemas reais. Frequentemente, as redes obtidas por tais sistemas são de grande escala e exibem padrões de conexão não triviais (ALBERT; BARABÁSI, 2002; NEWMAN, 2003; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008), i.e., nem completamente regular nem completamente aleatório, sendo, portanto, denominadas *Redes Complexas*. Seja pela natureza ubíqua ou por oferecer um conjunto de ferramentas unificado, que reúne conceitos de estatística, sistemas dinâmicos e teoria dos grafos, redes complexas têm despontado como tópico de pesquisa promissor para várias áreas da ciência (BARABASI, 2003; BORNHOLDT; SCHUSTER, 2003; COHEN; HAVLIN, 2010; NEWMAN, 2010), incluindo aprendizado de máquina e mineração de dados (SILVA; ZHAO, 2016).

Exemplos de aplicações naturais de redes complexas incluem redes de interação social entre indivíduos (SCOTT, 2012) e entre companhias (MIZRUCHI, 1982), redes neurais biológicas (SPORNS, 2002; BULLMORE; SPORNS, 2009), redes de parceiros sexuais e românticos (LILJEROS *et al.*, 2001; BEARMAN; MOODY; STOVEL, 2004), Internet (FALOUTSOS; FALOUTSOS, 1999), *World Wide Web* (ALBERT; JEONG; BARABÁSI, 1999), redes de colaboração entre cientistas (NEWMAN, 2001; BARABÁSI *et al.*, 2002), cadeias de interações metabólicas (JEONG *et al.*, 2000), redes de distribuição de energia elétrica (ALBERT; ALBERT; NAKARADO, 2004), entre outras (NEWMAN, 2010).

No contexto do aprendizado de máquina, apesar de a maioria das bases de dados não serem, naturalmente, representadas através de redes, é bastante usual transformá-las em tal representação. Como uma rede compreende, basicamente, um conjunto de vértices e conexões entre eles, em bases de dados do tipo vetor de atributos, cada item de dado pode ser representado por

¹ Os termos *rede* e *grafo* são tratados como sinônimos neste documento.

um vértice enquanto as conexões são dadas a partir da afinidade entre esses dados, por exemplo, conectando cada item de dado aos seus k vizinhos mais próximos. Aliás, tal transformação é vista como uma forma de redução de dimensionalidade não linear (ROWEIS; SAUL, 2000; TENENBAUM; SILVA; LANGFORD, 2000; ZHU; LAFFERTY; ROSENFELD, 2005). Outra vantagem importante da representação em rede é a possibilidade de analisar as características estruturais, dinâmicas e funcionais dos dados. As características estruturais são descritas mediante as interações entre os vértices, as dinâmicas através da evolução de tais interações e as funcionais por meio da análise dos efeitos da estrutura e da dinâmica no funcionamento geral da rede. Nesse sentido, vários estudos relacionam as propriedades das redes com o funcionamento dos sistemas modelados por elas (ALBERT; BARABÁSI, 2002; NEWMAN, 2003; BOCCALETTI *et al.*, 2006; COHEN; HAVLIN, 2010; FORTUNATO, 2010; NEWMAN, 2010; COSTA *et al.*, 2011).

De maneira geral, a habilidade para capturar as relações espaciais, topológicas e funcionais dos dados é uma das principais motivações para a pesquisa em redes complexas. Muitos estudos têm sido bem sucedidos explorando diferentes propriedades da rede em diferentes aplicações. Um exemplo bastante conhecido é o PageRank, método baseado em grafos, que opera por trás da *engine* universal de busca do *Google*, a fim de caracterizar a importância de páginas da web (BRIN; PAGE, 1998; PAGE *et al.*, 1999; LANGVILLE; MEYER, 2011; TABRIZI *et al.*, 2013). Outro exemplo conhecido no domínio de aprendizado de máquina e mineração de dados, é a detecção de comunidades, em que a estrutura topológica da rede é usada para encontrar comunidades, i.e., grupos caracterizados por conjuntos de vértices densamente conectados entre si, porém esparsamente conectados com vértices de outros grupos (NEWMAN; GIRVAN, 2004; CLAUSET; NEWMAN; MOORE, 2004; BOCCALETTI *et al.*, 2007; FORTUNATO, 2010; GOLDENBERG *et al.*, 2010; YANG; LESKOVEC, 2015).

1.1 Motivação

As motivações para a pesquisa de redes complexas, no contexto da classificação de dados, estão relacionadas às diferentes heurísticas de classificação que o aprendizado baseado em redes pode oferecer. Originalmente, a classificação é uma habilidade natural dos seres humanos. Todos os dias em suas vidas, classificam pessoas, lugares, comidas, ameaças, etc. Apesar da classificação ser um processo natural, o mecanismo implícito pelo qual o cérebro humano pode realizar esse tipo de tarefa de diferentes maneiras está longe de ser esclarecido. Embora não tenhamos o entendimento de todo o processo, algumas evidências relacionadas são conhecidas. Por exemplo, alguns estudos mostram que nosso cérebro pode reconhecer movimentos biológicos a partir de conjuntos de neurônios os quais detectam propriedades específicas distintas relacionadas à seletividade, visualização e robustez (GIESE; POGGIO, 2003). Outros estudos apontam fortes indícios de que decisões perceptuais são construídas por integrar evidências entre populações de neurônios distribuídas em áreas de processamento sensorial (HEEKEREN *et al.*, 2004; SIEGEL; ENGEL; DONNER, 2011).

Dada a sua importância prática, a literatura é composta por muitas técnicas para a classificação de dados, tais como redes neurais artificiais (HAYKIN, 2009), algoritmos Bayesianos (NEAPOLITAN, 2003; NIELSEN; JENSEN, 2009), árvores de decisão (BREIMAN, 1984; QUINLAN, 1986; BREIMAN, 2001), aprendizado baseado em instância (KIDWELL, 2013), e máquina de vetores de suporte (CORTES; VAPNIK, 1995; VAPNIK, 2013). Em comum, todas elas realizam a tarefa de classificação considerando apenas os atributos físicos dos dados de entrada (similaridade, distância ou distribuição). Em outras palavras, técnicas de classificação tradicional definem fronteiras de decisão no espaço, e uma nova instância é classificada por verificar sua posição relativa às fronteiras. Logo, as relações semânticas ou topológicas existentes entre os itens de dados são ignoradas. Por outro lado, o cérebro humano (animal) utiliza a habilidade de inferência de vários conjuntos de neurônios para detectar propriedades distintas e combiná-las dentro de um padrão completo, correspondente a uma classe de dados. Dessa maneira, o cérebro humano possui habilidade para identificar a estrutura organizacional dos dados não importa quão fisicamente diferentes os itens de dados de mesma classe são. Considere-se um conjunto de dados de treinamento onde haja algumas instâncias representando chassis (B), portas (D) e pneus (T). Considere-se, também, uma instância de teste y - uma porta - a qual precisa ser classificada. A Fig. 1a mostra que técnicas de classificação computacionais classificarão y na classe de portas (D), uma vez que y é fisicamente similar às portas d_1 e d_3 . Pelo contrário, o cérebro humano pode não apenas gerar a mesma classificação como as técnicas tradicionais fazem (rotulando y como da classe porta), mas pode ainda organizar diferentes tipos de objetos dentro de categorias organizacionais, i.e., combinando um chassi e um pneu, a classificação realizada pelo cérebro pode associar y a uma classe carro C_2 , como ilustrado pela Fig. 1b. Considere-se outro exemplo, dado um conjunto de estudantes, professores e prédios, a classificação computacional é mais propensa a gerar uma classe estudante, uma classe professor e uma classe prédio em vez de combinar todos aqueles componentes para formar classes escola, como a classificação cerebral faz. Aqui, a questão é: como realizar tal classificação organizacional por computador? Uma resposta parcial é apresentada neste documento.

Apesar da classificação por computador ainda estar distante da humana, algum esforço precisa ser feito no sentido de agregar as propriedades distintas dos dados no processo de classificação. Por exemplo, considere-se a Fig. 2a, a qual mostra uma base de dados simples com duas classes denotadas pelos itens de dados na forma de círculos e quadrados, i.e., $\mathcal{L} = \{\circ, \square\}$. Na figura, os itens na forma de triângulo ($\triangle$) representam novos itens de dados que precisam ser classificados. Uma vez que classificadores tradicionais, como k vizinhos mais próximos (k -NN) e máquina de vetores de suporte (SVM), realizam a classificação baseados apenas nos atributos físicos dos dados, eles, normalmente, possuem dificuldade para detectar esses novos itens de dados como pertencentes à classe de círculos. Pelo contrário, tais técnicas classificam-nos como itens de dados da classe de quadrados, como exibido nas Figs. 2b e 2c. Analisando o processo de decisão por trás desses resultados, é possível observar que as técnicas tradicionais falham em identificar os padrões semânticos formados pelos dados. Por exemplo, k -NN classifica um item

Figura 1 – Ilustração da diferença entre classificação computacional e cerebral. (a) resultado de classificação considerando apenas os atributos físicos dos dados de entrada. Nesse caso, a instância de teste y - uma porta - é associada à classe porta (D), porque é similar a d_1 e d_3 , porém é fisicamente diferente de outros objetos. (b) a classificação cerebral considera a relação organizacional entre os itens de dados, o que, por vez, organiza as instâncias de dados em um sistema repleto de sentido: y pode ser associado a uma classe carro C_2 .

de teste por verificar os rótulos de seus k vizinhos mais próximos (circulados na Fig. 2b), e SVM leva em conta os vetores de suporte (circulados na Fig. 2c) para aproximar cada classe com um fecho convexo e, então, encontrar o melhor hiperplano de separação entre as classes.

Em contrapartida, ainda tomando como exemplo a Fig. 2, redes complexas podem prover diferentes conceitos e heurísticas para a tarefa de classificação, por considerar a estrutura semântica dos dados. Embora seja um tema bastante investigado no contexto do aprendizado não supervisionado e semissupervisionado, especialmente, em tarefas de agrupamento de dados (também denominado detecção de comunidades), propagação de rótulos e redução de dimensionalidade (PALLA *et al.*, 2005; CHAPPELLE; SCHOLKOPF; ZIEN, 2006; YAN *et al.*, 2007; ZHU, 2008; QUILES *et al.*, 2008; FORTUNATO, 2010; SILVA; ZHAO, 2012b; NEWMAN, 2012; SILVA; ZHAO, 2013), existem ainda poucas investigações relacionadas ao uso de redes para as tarefas do aprendizado supervisionado (BERTINI *et al.*, 2011; SILVA; ZHAO, 2012a). A classificação híbrida proposta em Silva e Zhao (2012a), a qual combina técnicas de classificação tradicionais, também denominadas de baixo nível, e técnicas baseadas em medidas de redes complexas, denominadas de alto nível, representou um avanço nesse sentido.

A classificação de alto nível é um conceito recente de classificação baseado em redes. Nessa técnica, a classificação acontece por verificar em qual dos componentes da rede (classe) a inserção de um item de teste causa menor variação nas medidas de redes complexas, i.e., o novo item recebe o rótulo da classe na qual está em maior conformidade com o padrão. Se, por um lado, a classificação híbrida combina propriedades distintas na classificação de dados, por outro, o número de parâmetros da técnica é demasiadamente elevado. São parâmetros na etapa de construção da rede, parâmetros no cálculo das variações das medidas de redes, parâmetros da técnica de baixo nível, além do parâmetro de combinação linear entre as associações de baixo e alto nível. Como todos estes parâmetros são orientados ao problema, a seleção de seus valores toma tempo considerável e possui uma forte influência na qualidade da predição final. Para lidar

Figura 2 – Análise do processo de classificação de técnicas tradicionais. (a) Base de dados artificial composta por duas classes denotadas por itens de dados no formato de círculos e quadrados, enquanto itens de teste são representados na forma de triângulo; (b) k vizinhos mais próximos classifica os itens de teste verificando o rótulo dos seus k vizinhos mais próximos (circulados na figura); (c) Máquina de vetores de suporte levam em conta os vetores de suporte (circulados na figura) para aproximar cada classe com um fecho convexo e encontrar o melhor hiperplano de separação entre as classes, o qual é usado como fronteira de decisão para classificar os itens de teste. Técnicas de classificação tradicionais são incapazes de considerar a estrutura semântica dos dados e, conseqüentemente, os itens de teste são rotulados como pertencentes à classe de quadrados.

com tal complexidade de parâmetros, uma técnica de classificação simples que usa uma única rede, para prover tanto as informações de baixo quanto de alto nível, é proposta nesta tese.

Em se tratando de combinar propriedades ou visões distintas sobre os dados, algumas técnicas de aprendizado de máquina precisam ser mencionadas, embora tais visões não estejam associadas ao conceito de classificação, mas ao desempenho (acurácia) de classificação. Por exemplo, máquinas de comitê usam um conjunto de classificadores em que cada classificador toma uma decisão por si próprio e todas essas decisões são combinadas em uma única resposta por um esquema de votação (HAYKIN, 2009). Outro exemplo é cotreinamento, o qual requer duas visões dos dados e dois classificadores. Essa técnica assume que cada instância é descrita recorrendo a dois diferentes conjuntos de atributos que fornecem informações complementares sobre as instâncias (BLUM; MITCHELL, 1998; GOLDMAN; ZHOU, 2000). Contudo, uma vez que os classificadores usados nessas técnicas se preocupam apenas com as características físicas dos dados de entrada, cada classificador provê uma decisão com um certo nível de dependência estatística dos outros. Em outras palavras, cada classificador não provê realmente uma visão independente no nível semântico ou em uma forma particular de organização de cada classe de dados.

Embora a análise da conformidade de padrão dos dados seja o conceito de classificação explorado na classificação de alto nível, ele não é o único possível de ser derivado a partir de redes complexas. Muitos outros conceitos, visões e heurísticas de classificação podem ser obtidos a partir, por exemplo, das propriedades estruturais e dinâmicas da rede. Aliás, tais visões são capazes de fornecer uma camada adicional de informação às bases de dados. Como exemplo, considere-se a suposição comum na classificação de dados que afirma que todos os itens de dados possuem a mesma relevância. Além de tal suposição não ser compatível com a classificação natural realizada pelo cérebro humano, essa negligência referente à importância individual de cada item de dado pode mudar o entendimento sobre toda a base de dados. No exemplo da Fig. 1, cada objeto (chassis, portas e pneus) tem sua própria relevância e um não pode ser trocado por outro. Suponha-se que os chassis são removidos ou trocados por pneus, os carros não existem mais. A fim de tomar em conta a relevância de cada item de dado, outra pesquisa apresentada neste documento contempla o desenvolvimento de uma técnica capaz de detectar a formação de padrão dos dados de maneira global, por extrair da rede construída propriedades espaciais e estruturais via medidas de redes complexas, de modo que um item de teste recebe o rótulo da classe na qual ele possui maior *importância*.

O exemplo mostrado pela Fig. 2a consiste em uma base de dados simples cujas classes apresentam padrões muito claros. Como mostrado, k -NN e SVM, por exemplo, não são capazes de levar em conta a formação de padrão dos dados de entrada. Por outro lado, por considerar a estrutura topológica dos dados por meio da rede apresentada na Fig. 3a, técnicas baseadas em redes podem capturar informações semânticas sobre os dados. Na Fig. 3b, é possível observar que a classificação baseada em *importância* atribui a classe círculo para os novos itens de dados.

Figura 3 – Análise da classificação realizada pela técnica baseada em *importância*. (a) rede formada a partir da base de dados mostrada pela Fig. 2a; (b) explorando propriedade topológicas, funcionais e espaciais representadas nas conexões da rede, a técnica baseada em *importância* é capaz de detectar relações semânticas entre os dados e, consequentemente, classifica os itens de teste como pertencentes à classe de círculos.

Apesar da contribuição em diferentes conceitos de classificação providos através de redes complexas, seja na classificação de alto nível ou na classificação baseada em importância, um denominador comum entre as técnicas de aprendizado baseado em grafos é que uma rede, na maioria das vezes, precisa ser gerada a partir das bases de dados do tipo vetor de atributos. Embora alguns estudos apontem a etapa de formação da rede como crucial para o bom desempenho preditivo dos métodos de aprendizado baseado em redes (ZHU; LAFFERTY; ROSENFELD, 2005; MAIER; LUXBURG; HEIN, 2008; SOUSA; REZENDE; BATISTA, 2013), estes mesmos estudos também destacam a pouca atenção que este tópico tem recebido dentro da área de aprendizado baseado em grafos. Enquanto o estudo e desenvolvimento de medidas e propriedades é bastante comum, especialmente, para as etapas de agrupamento e propagação de rótulos, investigações que visam ao desenvolvimento de métodos para a construção da rede são escassas. Prova disso é que a grande maioria dos trabalhos na literatura empregam métodos heurísticos como k vizinhos mais próximos para obtenção da rede (OZAKI *et al.*, 2011). Aliás, os métodos de construção da rede para propósitos gerais, i.e., não específicos para um conjunto limitado de tarefa ou aplicações, são, geralmente, caracterizados pela localidade e uniformidade como lidam com os dados.

O tópico de geração de redes não é inédito na área de redes complexas. Pelo contrário, há um conjunto de estudos e ferramentas relacionados às propriedades e reproduções de diversos fenômenos existentes em redes reais. Contudo, enquanto a maioria dos modelos de redes complexas são descritos em função de algum mecanismo de crescimento e muitos outros trabalhos estão concentrados em explorar medidas de redes, a otimização estrutural de redes é um tópico raramente abordado (NEWMAN, 2010). A otimização das conexões da rede de acordo com uma meta ou certas características são desejáveis para alguns problemas, como, por

exemplo, o sistema *hub-and-spoke* das redes de linhas aéreas (CANCHO; SOLÉ, 2003; COHEN; HAVLIN, 2010). Como meio de preencher essa lacuna, especialmente, no que diz respeito ao aprendizado baseado em grafos, esta tese propõe um *framework* geral para a construção da rede por meio da otimização estrutural. A técnica desenvolvida constrói o grafo enquanto otimiza uma função de qualidade relacionada a uma tarefa de aprendizado, a qual pode ser, por exemplo, a classificação de dados, a redução de dimensionalidade, a detecção de anomalia, etc.

Nesta tese, o potencial do aprendizado baseado em redes também é avaliado na classificação de papéis semânticos em sentenças de textos jornalísticos. Em processamento de línguas naturais (PLN), a anotação automática de papéis semânticos (APS) consiste em identificar e classificar os argumentos de um predicador (verbo) com papéis que indiquem relações semânticas entre eles. Por conseguinte, APS pode representar as relações lógicas que existem entre um evento e seus participantes, como “quem fez o quê?”, “para quem?”, “onde?”, “quando?” e “como?”. Motivado pela vasta possibilidade de aplicações de APS (PALMER; GILDEA; XUE, 2010) e pela pequena quantidade de pesquisas explorando o tópico para a língua portuguesa (ALVA-MANCHEGO; ROSA, 2012; FONSECA; ROSA, 2013; HARTMANN, 2015), técnicas baseadas em rede são investigadas na tarefa de difusão de papéis semânticos, a qual nada mais é do que a tarefa de APS no contexto de aprendizado semissupervisionado. Especificamente, o foco é a classificação de argumentos usando o PropBank-br (DURAN; ALUÍSIO, 2012). Além de ser um corpus de língua portuguesa, outra motivação importante para a escolha do PropBank-br é a pouca quantidade de dados anotados, cujo tamanho é cerca de um sétimo da versão inglesa. Tal característica representa um cenário difícil para o aprendizado de máquina onde técnicas baseadas em grafos podem ser avaliadas em bases de dados com distribuições muito arbitrárias e desbalanceadas.

Em síntese, as investigações apresentadas neste documento combinam ferramentas e conceitos de classificação, sistemas de aprendizado, estatística e redes complexas.

1.2 Objetivos

O presente trabalho tem como objetivo principal o desenvolvimento e a aplicação de métodos e heurísticas baseadas em teorias de redes complexas para a classificação de dados. Nesse sentido, a hipótese investigada é se as características estruturais e dinâmicas advindas da representação gráfica dos dados e capturadas a partir de medidas e propriedades de redes complexas podem prover métodos computacionais eficientes para a tarefa de classificação. O termo eficiência denota aqui um balanço entre desempenho preditivo e complexidade computacional de modo que as novas características e vantagens providas pelos métodos propostos possam ser verificadas em várias aplicações reais. Especificamente, os seguintes objetivos são perseguidos:

1. Projetar uma técnica capaz de considerar a conformidade de padrão dos dados no processo de classificação, e caracterizada pela simplicidade (especificamente, o baixo número de

- parâmetros) e pelo bom desempenho preditivo em relação aos classificadores tradicionais. Dado o elevado número de parâmetros da técnica de classificação híbrida (SILVA; ZHAO, 2012a) e sobre a hipótese de que tal característica dificulta a sua reprodução e aplicação prática, propõe-se um novo método computacional no qual a mesma rede, construída a partir dos dados de entrada e sem uso de parâmetros, é usada para fornecer as associações de baixo e alto nível.
2. Desenvolver um conceito de classificação que, a partir da abstração gráfica, caracterize a importância individual dos itens de dados, e usufrua dessa informação no processo de classificação de maneira eficiente. Ao contrário das técnicas tradicionais, que realizam a classificação considerando apenas os atributos físicos dos dados, e diferente da conformidade de padrão, que realiza a classificação de acordo com a variação nas medidas de rede, propõe-se um método definido a partir do PageRank. Tal método calcula a importância dos itens de dados, de modo que um novo item receba o rótulo da classe na qual ele recebe maior importância. Ainda, a técnica explora as características físicas e estruturais dos dados usando uma medida de eficiência na rede e é capaz de capturar relações locais e globais entre os dados, a partir da medida de *eficiência diferencial espaço-estrutural* proposta.
 3. Provar que redes geradas a partir de um processo de otimização estrutural podem apresentar melhor desempenho que outras geradas heurísticamente em aplicações do aprendizado baseado em grafos. Embora o aprendizado de máquina baseado em grafos receba atenção crescente por parte da comunidade científica, muito pouco tem sido feito no que diz respeito à etapa de construção da rede, i.e., a transformação das bases de dados do tipo vetor de atributos em grafos. A investigação sobre otimização estrutural da rede vem preencher parte dessa lacuna, por investigar se o processo de otimização das conexões da rede, em relação a uma função objetivo, pode prover resultados melhores do que métodos definidos heurísticamente. Um *framework* geral baseado em enxame de partículas, para otimização estrutural de redes, é proposto, a fim de permitir estudar e analisar empiricamente o processo de otimização de redes em diversas aplicações relacionadas ao uso de grafos para aprendizado, por exemplo, a classificação de dados, a redução de dimensionalidade, a detecção de anomalia, entre outras. Especificamente neste documento, o *framework* proposto é analisado na classificação baseada em importância e na classificação híbrida.
 4. Analisar e comparar o desempenho preditivo de técnicas do aprendizado baseado em grafos na tarefa de APS para a língua portuguesa. Na técnica proposta, as sentenças extraídas do corpus PropBank-br são mapeadas para uma rede de modo que cada argumento representa um vértice e as arestas entre eles são definidas mediante algum critério de similaridade. Diferentes métodos de formação de rede e estratégias de propagação de rótulos são avaliados na classificação de papéis semânticos a partir de poucos argumentos previamente rotulados.

1.3 Organização do trabalho

Os demais capítulos da tese são organizados da seguinte forma. No Capítulo 2, são revisados conceitos fundamentais para as investigações apresentadas neste trabalho. Especificamente, são revisadas definições básicas das áreas de redes complexas, aprendizado de máquina e aprendizado baseado em redes, bem como os principais trabalhos correlatos existentes na literatura, além da tarefa de anotação automática de papéis semânticos. Os desenvolvimentos e resultados gerados pelas investigações são apresentados a partir do Capítulo 3, onde uma técnica que combina associações de baixo e alto nível é capaz de considerar a estrutura topológica dos dados, mesmo com um número muito reduzido de parâmetros. No Capítulo 4, um novo conceito de classificação de dados baseado em redes, capaz de considerar a importância dos itens de dados, é descrito em detalhes e suas características salientes são destacadas mediante vários experimentos em bases de dados artificiais, bem como em aplicações reais. O *framework* inspirado em enxame de partículas para otimização estrutural da rede é apresentado no Capítulo 5, novamente, seguido por uma considerável quantidade de experimentos e comparações com métodos amplamente utilizados na literatura. O Capítulo 6 contempla a investigação da tarefa de difusão de papéis semânticos a partir de diferentes métodos de construção da rede e estratégias de propagação de rótulos. As conclusões sobre as investigações realizadas no documento estão no Capítulo 7, o qual também apresenta uma lista contendo algumas oportunidades de trabalhos futuros, bem como os artigos produzidos durante o período do doutoramento.

FUNDAMENTAÇÃO TEÓRICA E REVISÃO DA LITERATURA

Os principais conceitos e trabalhos relacionados às investigações contempladas pela tese são apresentados neste capítulo, a saber:

- A Seção 2.1 introduz conceitos fundamentais da área de redes complexas. Grafos são uma ferramenta robusta para modelagem de sistemas reais e são aplicados em uma variedade de problemas. Devido ao padrão de conexão não trivial destas redes, i.e. nem completamente regular nem totalmente aleatórias, elas são denominadas redes complexas. Conceitos básicos da área, a caracterização de medidas e modelos de redes e a otimização estrutural de redes são tópicos abordados na seção;
- O aprendizado de máquina, área relacionada ao desenvolvimento de modelos capazes de extrair conhecimento útil a partir de bases de dados, é apresentado na Seção 2.2. Especificamente, a seção oferece uma descrição geral de tarefas e técnicas associadas aos três paradigmas de aprendizado: não supervisionado, em que não há qualquer informação de rótulo ou classe; semissupervisionado, caracterizado pela existência de pouquíssimos itens de dados rotulados; e supervisionado, quando há informação disponível sobre os rótulos de todos os itens de dados usados no treinamento;
- Os principais métodos e heurísticas usados na literatura para a construção de grafos são discutidos na Seção 2.3, bem como as principais técnicas baseadas em grafos relacionadas aos métodos desenvolvidos na tese;
- A Seção 2.4 introduz os principais conceitos relacionados à tarefa de anotação automática de papéis semânticos, além do corpus considerado nas investigações e os principais trabalhos arrolados da literatura.

2.1 Redes complexas

Redes complexas lidam com conceitos de estatística, sistemas dinâmicos e teoria dos grafos, e despertam o interesse de muitas áreas do conhecimento, tais como física, matemática, biologia e sociologia, devido a suas aplicações sobre uma grande variedade de problemas, os quais incluem redes sociais, redes biológicas, Internet e *World Wide Web* e redes de energia elétrica (STROGATZ, 2001; ALBERT; BARABÁSI, 2002; DOROGOVTSSEV; MENDES, 2002; NEWMAN, 2003; BOCCALETTI *et al.*, 2006; BOCCALETTI *et al.*, 2007; FORTUNATO, 2010; NEWMAN, 2010). Como as características topológicas dessas redes não são triviais, i.e., nem completamente regulares nem completamente aleatórias, elas são denominadas redes complexas. A literatura possui muitas medidas e modelos para caracterizar a estrutura dessas redes. As medidas são, frequentemente, usadas para analisar as propriedades estatísticas que descrevem a estrutura e o comportamento de sistemas em rede, enquanto a criação de modelos de rede está normalmente relacionada ao entendimento do significado dessas propriedades.

Exemplos de medidas de rede bastante conhecidas incluem a *Modularidade* e o *PageRank* (NEWMAN; GIRVAN, 2004; BRIN; PAGE, 1998). A primeira é muito explorada para a tarefa de detecção de comunidades, uma vez que ela calcula a força da divisão de uma rede em grupos, e a segunda é a famosa medida de ranqueamento que opera por trás da *engine* de busca universal do *Google*, a fim de caracterizar a importância de páginas da web.

Em relação aos modelos de redes complexas, dentre os mais conhecidos estão as *Redes Pequeno Mundo* e as *Redes Livres de Escalas*, os quais apresentam características estruturais específicas (WATTS; STROGATZ, 1998; BARABÁSI; ALBERT; JEONG, 2000; PASTOR-SATORRAS; VESPIGNANI, 2001). Enquanto o primeiro é caracterizado por elevado agrupamento e proximidade entre os nós, o segundo é descrito por uma lei de potência.

Ao contrário da maioria dos modelos de redes, os quais são descritos em função de mecanismos de crescimento, outro tópico relevante em redes complexas considera que alguns problemas reais podem ser mais bem explicados pela otimização de uma meta ou certas características, o que é denominado otimização estrutural de redes. As redes de linhas aéreas são exemplos naturais dessa categoria (NEWMAN, 2010).

O conteúdo apresentado nessa seção é organizado da seguinte maneira: a Seção 2.1.1 apresenta conceitos básicos da teoria dos grafos; a Seção 2.1.2 traz medidas de redes muito estudadas na literatura de redes complexas e que são relevantes para os desenvolvimentos discutidos na tese; e a Seção 2.1.3 descreve alguns dos mais importantes modelos de redes, incluindo uma breve discussão sobre otimização estrutural de redes.

2.1.1 Conceitos básicos

Em redes complexas, o termo rede denota o conceito informal que descreve um objeto composto de elementos e as conexões entre estes elementos. Tomando como exemplo uma rede

social, os elementos estão, normalmente, associados às pessoas, enquanto as conexões podem estar associadas às relações de amizade entre elas. Matematicamente, a forma natural de modelar estas redes é a partir dos conceitos de grafos. Dessa forma, os termos rede e grafo são tratados como sinônimos neste trabalho. A seguir, são apresentadas definições e conceitos importantes da teoria de grafos (BOLLOBÁS, 1998; DIESTEL, 2006; GROSS; YELLEN, 2006).

Um *grafo* ou *rede* $\mathcal{G} = (V, E)$ é uma estrutura matemática composta por dois conjuntos finitos V e E , onde V é o conjunto de n vértices (ou nós) e E o conjunto de m arestas (ou conexões) do grafo. Cada vértice do grafo é normalmente identificado por um valor inteiro ordenado $i = 1, 2, \dots, n$, enquanto a conexão entre dois nós i e j é representada por (i, j) , ou seja $E \subseteq \{(i, j) \mid i, j \in V\}$.

De acordo com o tipo de aresta admitido, os grafos podem ser classificados em direcionados, não direcionados ou mistos.

- *Grafos não direcionados*. Um grafo é denominado não direcionado se suas arestas apresentam relação simétrica, isto é, o fato de haver uma conexão (i, j) implica a existência de (j, i) .
- *Grafos direcionados* ou *dígrafos*. Um grafo é denominado direcionado, se suas arestas apresentam origem e destino. Desta forma, a conexão (i, j) não necessariamente implica em (j, i) .
- *Grafos mistos*. Um grafo é denominado misto, se ele admite tanto arestas direcionadas quanto não direcionadas.

Os grafos também podem ser classificados em relação à existência de peso nas arestas:

- *Grafos ponderados*. Um grafo é dito ponderado, se valores numéricos são associados às suas arestas. Dessa forma, o peso das arestas pode ser representado como uma função $\omega : E \rightarrow \mathbb{R}$, que atribui a cada aresta $(i, j) \in E$ um peso $\omega(i, j)$.

Do ponto de vista da representação, um dígrafo não ponderado $\mathcal{G}$ pode ser definido em termos de uma matriz de adjacências $\mathbf{A}$, onde a existência de conexão do nó i para j é representada por $a_{ij} = 1$. Se $\mathcal{G}$ é não direcionado, então a matriz $\mathbf{A}$ é simétrica, i.e., $\mathbf{A} = \mathbf{A}^T$. Por outro lado, se $\mathcal{G}$ é um grafo ponderado ele é, normalmente, definido por meio de uma matriz ponderada $\mathbf{W}$, em que cada entrada $w_{ij} \in \mathbf{W}$ representa o peso associado à conexão de i para j , $\omega(i, j)$.

Ainda, os grafos são classificados por admitir ou não arestas de um vértice para si próprio, e por admitir ou não arestas idênticas:

- *Grafos com autoligação.* Um grafo é dito com autoligação, se ele admite arestas de um vértice para si próprio.
- *Grafos com arestas múltiplas.* Um grafo é denominado de arestas múltiplas, se ele admite uma coleção de duas ou mais arestas idênticas.

A partir dos conceitos apresentados, têm-se os conceitos de grafos simples e multigrafos:

- *Grafos simples e Multigrafos.* Um grafo é dito simples, se ele não admite autoligações nem arestas múltiplas. Caso admita, ele é denominado multigrafo.

As seguintes definições também são bastante difundidas em trabalhos da literatura:

- *Vértices adjacentes ou vizinhos.* Seja $\mathcal{G}$ uma rede não direcionada, dois vértices i e j conectados mediante uma aresta, i.e., $a_{ij} = 1$, são denominados adjacentes ou vizinhos. Seja $\mathcal{G}$ um dígrafo, j é dito vizinho de i se $a_{ij} = 1$.
- *Grau.* Se $\mathcal{G}$ é não direcionado, o grau de um vértice i , denotado por k_i , é o número total de vértices adjacentes à i :

$$k_i = \sum_{j=1}^n a_{ij} . \quad (2.1)$$

- *Grau de entrada e grau de saída.* Se $\mathcal{G}$ é uma rede direcionada, o grau de um vértice i é dado pela soma dos graus de entrada k_i^{in} e saída k_i^{out} do vértice:

$$k_i = k_i^{in} + k_i^{out} ; \quad k_i^{in} = \sum_{j=1}^n a_{ji} ; \quad k_i^{out} = \sum_{j=1}^n a_{ij} . \quad (2.2)$$

- *Grau médio.* O grau médio de uma rede é a média aritmética do grau de todos os seus vértices, i.e.:

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^n k_i \quad (2.3)$$

- *Grafo regular.* Um grafo é dito regular, se todos os vértices apresentam o mesmo grau.
- *Grafo completo.* Um grafo $\mathcal{G}$ é denominado completo se todos os seus nós são adjacentes uns aos outros, i.e., se $\mathcal{G}$ é não direcionado então $m = \frac{n(n-1)}{2}$, caso contrário, $m = n(n-1)$.
- *Grafo conectado.* Um grafo é chamado conectado, se existe um caminho entre qualquer par de vértices distintos.

- *Passeio*. Suponha-se um grafo $\mathcal{G}$ o qual possui dois nós não adjacentes v_1 e v_l . Embora não sejam vizinhos, os nós podem estar conectados através de uma sequência de arestas a , tal como $v_1, a_1, v_2, a_2, \dots, a_l, v_l$. Uma sequência de vértices em que cada dois vértices consecutivos são conectados por uma aresta é denominada passeio. O comprimento da sequência está associado ao seu número de arestas (l no exemplo).
- *Caminho*. Se todos os nós e arestas de um passeio são distintos, ele é denominado caminho.
- *Ciclo*. Um passeio que termina no vértice em que ele começa e onde apenas o primeiro vértice aparece mais de uma vez é denominado ciclo.
- *Circuito*. Um passeio que começa e termina no mesmo vértice e passa uma única vez em cada aresta é denominado circuito. Note-se que, diferente do ciclo, vértices podem ser repetidos em um circuito.
- *Ciclo Hamiltoniano*. Um ciclo que passa por todos os n vértices do grafo é chamado ciclo Hamiltoniano.
- *Circuito Euleriano*. Denomina-se circuito Euleriano aquele que passa por todas as m arestas de um grafo.

Algumas medidas de distância no grafo também são bastante empregadas na literatura:

- *Distância geodésica*. A distância entre dois vértices i e j é o número de arestas do menor caminho entre eles.
- *Excentricidade*. A maior distância de um vértice i para qualquer outro vértice no grafo é denominada excentricidade.
- *Diâmetro*. O diâmetro de um grafo é a excentricidade máxima de qualquer vértice do grafo.
- *Raio*. Ao contrário do diâmetro, o raio de um grafo representa a excentricidade mínima de qualquer vértice do grafo.

Sejam dois grafos, $\mathcal{G}_1 = (V_1, E_1)$ e $\mathcal{G}_2 = (V_2, E_2)$, tem-se:

- *Subgrafo ou subrede*. $\mathcal{G}_2$ é denominado subgrafo ou subrede de $\mathcal{G}_1$ se $V_2 \subseteq V_1$ e $E_2 \subseteq E_1$.
- *Componente*. Um subgrafo $\mathcal{G}_2$ onde quaisquer vértices $i \in V_2$ são alcançáveis e não podem se conectar com outros vértices $j \in V' = V_1 - V_2$ é denominado um componente conexo de $\mathcal{G}_1$.
- *Subgrafo gerador e subgrafo induzido*. $\mathcal{G}_2$ é chamado um subgrafo gerador de $\mathcal{G}_1$ se $V_2 = V_1$ e $E_2 \subseteq E_1$. Nesse caso, diz-se, também, que $\mathcal{G}_1$ é um subgrafo induzido de $\mathcal{G}_2$.

- *Árvore Geradora*. Se $\mathcal{G}_2$ é um subgrafo gerador de $\mathcal{G}_1$, conexo e não possui ciclos, ele é denominado uma árvore geradora de $\mathcal{G}_1$.
- *Árvore Geradora Mínima*. Se $\mathcal{G}_2$ é uma árvore geradora de $\mathcal{G}_1$, ponderado e minimiza a seguinte equação, ele é nomeado uma árvore geradora mínima de $\mathcal{G}_1$.

$$\sum_{(i,j) \in E_2} \omega(i, j) \quad (2.4)$$

2.1.2 Medidas de redes complexas

A literatura sobre redes complexas engloba um vasto conjunto de medidas de rede. Essas medidas possuem aplicações diversas em diferentes sistemas reais e são oriundas de estudos do campo da estatística, sistemas complexos, matemática, sistemas não lineares, entre outros (NEWMAN, 2003; COSTA *et al.*, 2007). A seguir, são apresentadas algumas medidas bastante utilizadas na literatura. Para fim de exemplificação das medidas, considere-se a rede “Caratê”, exposta na Fig. 4, a qual foi compilada em Zachary (1977).

Figura 4 – Rede “Caratê” compilada em Zachary (1977). As arestas caracterizam as relações de amizade de pessoas envolvidas no episódio do rompimento das colaborações entre dois instrutores de caratê de uma academia, representados pelos nós 1 e 34 da rede. Por conta da mudança de um dos instrutores para outra academia, vários alunos também se desligaram da academia para acompanhá-lo.

2.1.2.1 Grau

O grau médio $\langle k \rangle$ é uma das medidas de rede mais simples. Tal medida quantifica estatisticamente o grau médio dos vértices em uma rede. Como apresentado na Subseção 2.1.1, o grau médio da rede é dado pela Eq. (2.3) e o grau de um vértice pela Eq. (2.1).

Na Fig. 5 é apresentado o grau de cada um dos vértices da rede “Caratê” ilustrada na Fig. 4. Note-se que os dois instrutores registram os maiores valores de grau. O valor de grau médio da rede é 4.59.

Figura 5 – Grau dos nós da rede “Caratê”.

2.1.2.2 Assortatividade

A medida de assortatividade quantifica a tendência de conexões entre vértices (NEWMAN, 2002). Assim, esta medida analisa se uma ligação ocorre, preferencialmente, entre vértices com grau similar ou não. Originalmente, o coeficiente de assortatividade r representa o coeficiente de correlação de Pearson de grau entre pares de nós conectados. Tal coeficiente pode assumir valores entre $[-1, 1]$, de modo que valores positivos indicam que pares de vértices diretamente conectados estão mais suscetíveis a se comportar da mesma forma, enquanto valores negativos indicam maior probabilidade dos vértices conectados terem comportamento distintos. No primeiro caso ($r \rightarrow 1$), o padrão de mistura assortativa é dito perfeito e no segundo caso ($r \rightarrow -1$), é denominado disassortativo.

Seja $L = |E|$ o número de conexões na rede e i_u, k_u os graus dos vértices i e k , que compõe uma aresta $u \in E$, a assortatividade da rede pode ser calculada por:

$$r = \frac{L^{-1} \sum_u i_u k_u - [L^{-1} \sum_u \frac{1}{2} (i_u + k_u)]^2}{L^{-1} \sum_u \frac{1}{2} (i_u^2 + k_u^2) - [L^{-1} \sum_u \frac{1}{2} (i_u + k_u)]^2} \quad (2.5)$$

O valor de assortatividade obtido para o grafo exibido na Fig. 4 é -0.48, o que denota que os vértices adjacentes, normalmente, apresentam diferentes valores de grau, tal como pode ser visto na Fig. 5.

2.1.2.3 Coeficiente de agrupamento

A medida de coeficiente de agrupamento quantifica o grau para o qual vértices locais em uma rede tendem a se agrupar. Várias evidências sugerem que, em muitas redes reais, tais como as redes sociais, os vértices tendem a criar grupos coesos caracterizados por uma densidade relativamente alta de laços (WATTS; STROGATZ, 1998).

Há muitas maneiras para se calcular o coeficiente de agrupamento, todavia esta medida foi originalmente proposta por Watts e Strogatz (1998). Inicialmente, suponha-se que um vértice i possua k vizinhos, o que possibilita um número máximo de $k(k-1)$ arestas entre eles, se qualquer vizinho u de i é conectado a qualquer outro vizinho s de i . Assim, o coeficiente de agrupamento local de um vértice i quantifica quão próximos seus vizinhos estão de um grafo completo (clique) e é dado por:

$$CC_i = \frac{|e_{us}|}{k_i(k_i - 1)}, \quad (2.6)$$

onde $CC_i \in [0, 1]$, $|e_{us}|$ denota o número de conexões compartilhadas pelos vizinhos diretos do vértice i e k_i é o grau do vértice i . Consequentemente, o coeficiente de agrupamento médio da rede é obtido por:

$$CC = \frac{1}{|V|} \sum_{i=1}^{|V|} CC_i. \quad (2.7)$$

Na Fig. 6, é apresentado o coeficiente de agrupamento de cada um dos vértices da rede mostrada na Fig. 4. O coeficiente de agrupamento médio da rede é 0.57.

Figura 6 – Coeficiente de agrupamento dos elementos da rede “Caratê”.

2.1.2.4 Proximidade

A proximidade mede o inverso do caminho mínimo (distância geodésica) médio entre um nó e todos os outros nós da rede (FREEMAN, 1979) e é dada por:

$$AC_i = \frac{n-1}{\sum_{j=1}^n d(i,j)}, \quad (2.8)$$

onde $AC_i \in [0, 1]$, $d(i, j)$ é o caminho mínimo entre os vértices i e j , e n denota o número de vértices na rede. A proximidade média é dada por:

$$AC = \frac{1}{n} \sum_{i=1}^n AC_i. \quad (2.9)$$

A Fig. 7 mostra a proximidade normalizada de cada um dos vértices da rede apresentada na Fig. 4. A proximidade média da rede é 0.43.

Figura 7 – Proximidade normalizada dos elementos da rede “Caratê”.

2.1.2.5 Intermedialidade

A intermedialidade ou *betweenness* é uma medida que indica a centralidade de vértices na rede e foi proposta, originalmente, em (FREEMAN, 1977; FREEMAN, 1979). Considere-se uma rede $\mathcal{G}$ composta por um conjunto de vértices V , os quais trocam mensagens entre eles utilizando o conjunto de arestas E , de modo que estas mensagens sempre tomam o menor caminho da rede (caminho geodésico). Dessa forma, o grau de intermediação b_i de um vértice i equivale ao número de caminhos geodésicos (η), que possuem i no percurso, ou seja, avalia-se a presença do

vértice i no menor caminho existente de um vértice u para outro vértice v , tal que $u, v \in V - i$. Formalmente, o grau de intermediação de um vértice i é definido por:

$$b_i = \sum_{u,v \in V - \{i\}} \eta_{uv}^i. \quad (2.10)$$

Uma vez que pode haver mais de um menor caminho entre os vértices u e v , denomina-se ng_{uv} o número de menores caminhos entre os vértices. Considerando-se, ainda, que o vértice i também pode participar de muitos caminhos existentes em ng_{uv} , denomina-se ng_{uv}^i o número de caminhos em ng_{uv} que passam por i . Dessa forma, η_{uv}^i é dado por:

$$\eta_{uv}^i = \frac{ng_{uv}^i}{ng_{uv}}. \quad (2.11)$$

A Fig. 8 exhibe a intermedialidade de cada um dos vértices da rede apresentada na Fig. 4.

Figura 8 – Intermedialidade dos nós da rede “Caratê”.

2.1.2.6 Modularidade

Esta medida possui relação com o modelo de rede modular. Conceitualmente, a modularidade Q quantifica quão boa é uma particular divisão da rede. Ela varia de 0, representando total aleatoriedade da rede, para 1, o qual indica comunidades claramente visíveis na rede. Originalmente, essa medida foi proposta em (CLAUSET; NEWMAN; MOORE, 2004; NEWMAN, 2006), onde foi utilizada para detecção de comunidades no contexto de aprendizado não supervisionado. Tal medida se tornou rapidamente um elemento essencial em muitos algoritmos de agrupamento. De fato, ela é a mais utilizada e mais bem conhecida função de qualidade, pois representa uma das primeiras tentativas de alcançar um primeiro princípio de entendimento do

problema de agrupamento, e ela incorpora à sua forma compacta todos os ingredientes essenciais e questões, a partir da definição de comunidade, para a escolha de um modelo nulo, para a expressão da “força” de comunidades e partições (FORTUNATO, 2010).

Formalmente, a modularidade é denotada por:

$$Q = \frac{1}{2E} \sum_{u,v} \left[e_{uv} - \frac{k_u k_v}{2E} \right] \delta(c_u, c_v) \quad (2.12)$$

onde E é o número total de arestas na rede, e_{uv} indica a fração de arestas que conectam vértices de u para v na comunidade, k_u denota o grau do vértice u e $\delta(c_u, c_v)$ representa o delta de Kronecher, o qual produz 1 se $c_u = c_v$ e 0, caso contrário. Basicamente, a modularidade calcula a fração de conexões que acontecem dentro dos grupos, subtraída pelo número esperado de conexões caso elas fossem distribuídas aleatoriamente (note-se que conceitos de redes aleatórias também são empregados na medida). Assim, para uma dada divisão dos vértices da rede dentro de alguns módulos, a medida de modularidade reflete a concentração de vértices dentro dos módulos, comparada com uma distribuição aleatória de ligações entre todos vértices, independentemente dos módulos (NEWMAN, 2006).

O valor de modularidade obtido para os grupos formados na rede apresentada na Fig. 4 é 0.40, o que é um forte indício da existência de comunidades no grafo de acordo com a densidade das conexões intra e inter grupos.

2.1.2.7 PageRank

PageRank é uma medida de centralidade baseada em passeios aleatórios, a qual foi introduzida em Brin e Page (1998). Intuitivamente, a medida pode ser interpretada como um modelo de caminhada aleatório, o qual assume que um agente está passeando sobre o grafo. Em cada nó, o agente, aleatoriamente, seleciona uma das conexões de saída e vai para aquele nó vizinho. Além disso, o agente pode saltar, aleatoriamente, para qualquer outro nó do grafo, independente de estar ou não conectado ao nó atual, com uma certa probabilidade. Este último mecanismo assegura que o processo de caminhada não ficará preso em um subgrafo sem conexões de saída, evitando os tão conhecidos problemas de convergência (*spider-trap* e *dead-end*). A probabilidade de o agente estar em cada nó é o seu valor de PageRank. Em poucas palavras, quanto maior o número de conexões de entrada de um nó, mais importante ele é, e se os nós que realizam conexões para um dado nó possuem grande relevância, este nó também possuirá. Matematicamente, PageRank pode ser descrito pela seguinte formulação:

$$PR_j^{(t+1)} = \sum_{i \rightarrow j} \beta \cdot \frac{PR_i^{(t)}}{k_i^{out}} + (1 - \beta) \frac{1}{n}, \quad (2.13)$$

onde $PR_i^{(t)}$ é o PageRank do vértice i , k_i^{out} é o grau de saída do vértice i , n é o número de nós na rede, β é a probabilidade de saltos aleatórios, e $i \rightarrow j$ denota a conexão de um vértice i para o vér-

tice j . Inicialmente, os valores de PR podem ser definidos aleatoriamente respeitando a seguinte condição $\sum_j PR_j = 1$, uma vez que a equação é calculada iterativamente até a convergência.

É possível interpretar o PageRank como um tipo de caminhada aleatória, onde em um dado instante de tempo t um caminhante está no nó i , no tempo $t + 1$ em um dos nós conectados por i , e assim por diante, tal como:

$$p^{(t+1)} = M \cdot p^{(t)}, \quad (2.14)$$

onde p é uma distribuição de probabilidade e M é uma matriz de adjacência estocástica definida por:

$$M_{ji} = \begin{cases} \frac{1}{k_i^{out}}, & \text{se } i \rightarrow j, \\ 0, & \text{caso contrário,} \end{cases} \quad (2.15)$$

na qual a soma de quaisquer de suas colunas é igual a 1. Supondo que a caminhada aleatória alcança um estado estacionário, i.e., $p(t + 1) = p(t)$, p representa uma distribuição estacionária para a caminhada aleatória. Segundo a teoria de caminhadas aleatórias, para grafos que satisfazem certas condições, a distribuição estacionária é única e eventualmente será alcançada não importa qual a distribuição de probabilidade adotada no tempo $t = 0$.

A fim de garantir as condições necessárias para alcançar o estado estacionário, é preciso lidar com dois problemas bastante conhecidos: *spider-trap* e *dead-end*. O primeiro acontece quando todas as conexões de saída estão dentro do mesmo grupo, enquanto no segundo não há conexões de saída a partir de um dado nó. A Fig. 9 ilustra os dois problemas de convergência mencionados.

Figura 9 – Exemplo ilustrativo dos problemas de convergência *spider-trap* e *dead-end*.

A solução adotada pelo PageRank para lidar com os problemas de convergência são os saltos aleatórios (*random teleport*), onde a partir de uma probabilidade β , o caminhante segue uma das conexões de saída do nó atual, e com uma probabilidade $1 - \beta$, o caminhante vai para qualquer outro nó do grafo. Logo, a formulação do PageRank também pode ser dada por:

$$p^{(t+1)} = \beta M \cdot p^{(t)} + \left[\frac{1 - \beta}{n} \right]_n, \quad (2.16)$$

a qual é equivalente à formulação definida na Eq. 2.13.

A Fig. 10 apresenta os valores de PageRank para os elementos da rede Karatê, ilustrada na Fig. 4. Uma vez que a formulação do PageRank considera graus de saída dos vértices, as redes adotadas são normalmente dígrafos. Dessa forma, para exemplificar, adequadamente, a medida, cada aresta original da rede “Karatê” foi transformada em duas arestas, uma para cada direção.

Figura 10 – Valores de PageRank obtidos para os vértices da rede “Karatê” adaptada.

2.1.3 Modelos de redes complexas

O desejo de modelar propriedades e dinâmicas de rede, a fim de entender e reproduzir o comportamento de vários sistemas reais motiva diretamente o estudo de redes complexas. Muitas investigações teóricas e empíricas sobre redes reais, tais como Watts e Strogatz (1998), Barabási e Albert (1999), Newman (2010) culminaram no desenvolvimento de modelos de rede que refletem de forma comportada características próprias desses sistemas e que se constituem em uma “caixa de ferramentas” para a análise de vários outros sistemas complexos relacionados. Diante de tantas contribuições, cinco classes principais de modelagem de redes são apresentadas a seguir: *Redes Aleatórias*, *Redes Pequeno Mundo*, *Redes Livres de Escala*, *Redes Modulares* e *Otimização Estrutural de Redes*.

2.1.3.1 Redes aleatórias

Redes aleatórias são modelos de rede onde as conexões entre os vértices são realizadas de maneira aleatória a partir de um conjunto específico de parâmetros definidos. Normalmente, redes com topologias complexas ou princípios de organização desconhecidos parecem aleatórias. Dessa forma, a teoria de grafos aleatórios é bastante empregada no estudo de sistemas reais.

É importante ressaltar que propriedades de redes aleatórias não são investigadas em termos de uma única rede, mas em termos de um conjunto delas uma vez que o interesse dos estudos, geralmente, é a descoberta e a análise de propriedades típicas dessas redes. Nesse sentido, o objetivo é obter uma distribuição de probabilidade mais precisa por considerar as redes possíveis de ser geradas a partir dos parâmetros definidos. Duas abordagens são, normalmente, empregadas na geração de redes aleatórias: $G(n, m)$ e $G(n, p)$. A primeira abordagem define um número de vértices n e um número de arestas m . Na sequência m pares de vértices são definidos aleatoriamente a partir de todos os pares possíveis e são conectados por uma aresta. A segunda abordagem, também denominada *modelo Erdős-Rényi*, em homenagem às contribuições desses pesquisadores nesse tema, define um número de vértices n e realiza as ligações entre eles a partir de uma probabilidade $p \in [0, 1]$ de conectar cada par de vértices. Assim, para $p = 0$ não há arestas na rede, enquanto para $p = 1$ todos os vértices são ligados, resultando em uma rede totalmente conectada. A Figura 11 mostra exemplos de redes aleatórias geradas a partir do modelo *Erdős-Rényi* (ERDÖS; RÉNYI, 1959), considerando $n = 10$ e p igual a 0, 0.2, 0.5 e 0.8, respectivamente. Para uma probabilidade p qualquer, o grau médio da rede é dado por $c = p(n - 1)$. Redes geradas por esse modelo possuem distribuição de grau binomial, i.e.:

$$p_k = \binom{n}{k} p^k (1 - p)^{n-k} . \quad (2.17)$$

Figura 11 – Redes aleatórias geradas a partir do modelo *Erdős-Rényi* com número de vértices $n = 10$ e diferentes probabilidades.

A distribuição de grau médio obtido em cinco redes aleatórias geradas a partir de $n = 5000$ e $p = 0.1$ é apresentada na Figura 12. Pela figura, é possível notar que a distribuição de grau médio se aproxima da distribuição de Poisson para redes de grande tamanho ($n \rightarrow \infty$):

$$p_k = e^{-c} \frac{c^k}{k!} \quad (2.18)$$

2.1.3.2 Redes pequeno mundo

As propriedades das redes pequeno mundo podem ser encontradas em muitas redes reais, especialmente, naquelas onde a maioria dos vértices pode ser alcançada por outros com um

Figura 12 – Distribuição de grau médio de redes aleatórias geradas a partir do modelo *Erdős-Rényi*. Os parâmetros usados foram $n = 5000$ e $p = 0.1$.

pequeno número de conexões, tal como ocorre nas redes sociais. Aliás, um estudo importante que enfatiza características e propriedades dessas redes foi realizado pelo psicólogo Stanley Milgram, que conduziu seu famoso experimento “pequeno mundo” na década de 1960 (MILGRAM, 1967). O experimento consistiu em enviar 96 pacotes para pessoas escolhidas aleatoriamente, a partir de uma lista telefônica na cidade de Omaha, USA. Em cada pacote, havia um passaporte que deveria ser entregue a um destinatário, bem como o nome, endereço e profissão do destinatário. Além disso, havia um pedido para não enviar o passaporte diretamente para o destinatário, mas para algum conhecido, o qual julgasse ter mais chances de conhecer alguém ligado ao destinatário ou o próprio. Desse modo, o passaporte deveria passar de conhecido em conhecido até chegar ao destinatário. Como os participantes deveriam registrar cada passo do passaporte, ao término do experimento, Milgram sabia que o tamanho médio dos caminhos completados até o destinatário consistia em 5.9 passos, o que resultou na origem da famosa ideia de “seis graus de separação”, a qual afirma que apenas seis passos separam quaisquer duas pessoas no mundo.

O modelo de rede pequeno mundo é caracterizado por registrar alto grau de agrupamento C e baixa distância média entre os vértices da rede $\langle d \rangle$. Watts e Strogatz (1998) propuseram o primeiro algoritmo bem sucedido para gerar redes baseadas no modelo pequeno mundo. O algoritmo proposto pode ser definido em dois passos:

Rede regular Inicialmente, a rede compreende n vértices, cada um deles conectado aos seus k vizinhos mais próximos;

Probabilidade p A partir de uma probabilidade $p \in [0, 1]$, cada aresta é reconectada de modo que *self-loops* e arestas duplicadas são excluídos.

A probabilidade p é responsável por interpolar as características da rede entre as características das redes regular e aleatória. Dessa forma, a rede é regular quando $p = 0$. Ao incrementar o valor de $p : 0 \leq p \leq 0.1$, a rede é capaz de apresentar as propriedades do modelo pequeno mundo. Por outro lado, quando $p \rightarrow 1$, tem-se uma rede aleatória. A Tabela 1 mostra a baixa distância média e o elevado coeficiente de agrupamento médio das redes pequeno mundo a partir de uma rede regular ($k = 2$) com $n = 20$ vértices. Na tabela, os valores para $p > 0$ são obtidos por meio da média de 20 redes geradas aleatoriamente.

Tabela 1 – Análise da distância média $\langle d \rangle$ e coeficiente de agrupamento médio C de redes regular ($k = 2$, $p = 0$), pequeno mundo ($p = 0.06$) e aleatória ($p = 0.8$). Para $p > 0$, os resultados são obtidos a partir da média de 20 redes geradas aleatoriamente.

p	0	0.06	0.8
$\langle d \rangle$	2.75	2.34	2.08
C	0.5	0.38	0.22

A Figura 13 ilustra a formação de uma rede pequeno mundo através de uma rede regular obtida a partir de $k = 2$. Na Fig. 13a, a probabilidade p de reconectar uma aresta é igual a 0, logo, uma rede com elevado coeficiente médio de agrupamento e distância média é obtida. Por exemplo, a distância entre os dois nós destacados na figura é igual a 5. Uma rede pequeno mundo obtida com $p = 0.06$ é exibida na Fig. 13b. Nela, é possível perceber que a distância entre os nós destacados diminuiu. A Fig. 13c apresenta uma rede quase aleatória ($p = 0.8$), em que a distância entre os nós é ainda menor, porém o coeficiente de agrupamento também diminuiu bastante.

Figura 13 – Simulação do modelo de Watts e Strogatz (1998) para geração de redes pequeno mundo com $n = 20$ vértices: (a) $p = 0$; (b) $p = 0.06$; (c) $p = 0.8$.

2.1.3.3 Redes livres de escala

Os modelos de rede anteriores assumem que a probabilidade de dois vértices serem conectados é independente do grau dos vértices de modo que novas arestas são estabelecidas aleatoriamente. Por outro lado, muitas redes reais apresentam uma propriedade denominada

conexão preferencial, cujas arestas de um novo vértice adicionado à rede se conectam, preferencialmente, aos vértices que possuem maior grau.

Enquanto os modelos de rede aleatória e pequeno mundo estão diretamente relacionados às características topológicas da rede, consideram um número fixo de vértices e estabelecem as arestas entre vértices de maneira aleatória, o modelo de rede livre de escala visa capturar as dinâmicas da rede, considerando o crescimento da rede mediante a adição contínua de novos vértices e utilizando a propriedade de conexão preferencial, de modo que as conexões dos novos vértices são estabelecidas com os vértices da rede que possuem maior grau. De fato, este modelo pode ser percebido em muitas redes reais, tais como a própria *World Wide Web* e redes de publicação de artigos. Formalmente, uma rede é denominada livre de escala quando seu grau de distribuição segue uma lei de potência do tipo:

$$p_k \sim k^{-\gamma}, \quad (2.19)$$

onde γ é denominado expoente de livre escala. A partir dos conceitos de crescimento e conexão preferencial, [Barabási e Albert \(1999\)](#) propuseram o primeiro algoritmo, denominado modelo *Barabási-Albert*, para a construção de redes livre de escala:

Crescimento A partir de um número m_0 de vértices, a cada etapa, um novo vértice é conectado com $m < m_0$ vértices $\in m_0$;

Conexão preferencial A probabilidade p de um novo vértice u ser conectado a um vértice preexistente i depende do grau k_i , tal que:

$$p(i \rightarrow u) = \frac{k_i}{\sum_j k_j}. \quad (2.20)$$

A Figura 14 mostra um exemplo de rede livre de escala com $n = 60$ vértices. Na figura, é possível notar a propriedade de conexão preferencial, a qual colabora para a formação de *hubs* - vértices que apresentam muito mais conexões que os demais - na rede. Tais vértices são destacados na figura pela cor escura e tamanho maior.

A distribuição de grau médio obtido através de cinco redes aleatórias geradas a partir do modelo *Barabási-Albert* com $n = 5000$ e $m_0 = 5$ é apresentada na Figura 15. Ao contrário das redes aleatórias, as quais são bem definidas em termos de distribuição de grau médio, a figura ilustra os padrões de conectividade não uniformes relacionados às redes livres de escala.

2.1.3.4 Redes modulares

Uma característica importante de várias redes naturais e construídas pelo homem é a presença de estruturas locais denominadas módulos ou comunidades. Uma comunidade é definida como um grupo de vértices densamente conectados, cujas conexões entre grupos distintos são

Figura 14 – Exemplo de rede livre de escala obtida a partir do algoritmo proposto em [Barabási e Albert \(1999\)](#) com $n = 60$ e $m_0 = 5$.

Figura 15 – Distribuição de grau médio de redes livres de escala geradas a partir do modelo *Barabasi-Albert*. Os parâmetros usados foram $n = 5000$ e $m_0 = 5$.

esparsas. De modo geral, as comunidades representam padrões de interação entre vértices da rede e sua identificação é crucial para a análise e entendimento dos mecanismos de crescimento, formação e processos que ocorrem através da rede ([CLAUSET; NEWMAN; MOORE, 2004](#)).

Um modelo para gerar redes modulares foi originalmente proposto em [Girvan e Newman \(2002\)](#), com o objetivo de avaliar métodos para detecção de comunidades em redes. Os passos do algoritmo são apresentados a seguir:

- Um conjunto de n vértices é agrupado em $\mathcal{C}$ comunidades distintas sem sobreposição;

- Dois vértices são seleccionados e conectados com probabilidade p_{in} , se eles pertencem à mesma comunidade, ou $p_{out} < p_{in}$, se eles são de comunidades distintas.

Os valores de p_{in} e p_{out} são escolhidos de acordo com o nível de evidência da comunidade. Assim, altos valores de p_{in} resultam em uma rede com módulos bem definidos, enquanto baixos valores os tornam muito mal definidos. Empiricamente, $p_{out} \leq p_{in}$ permitem comunidades bem definidas na rede. A Figura 16 mostra uma rede gerada a partir do modelo de Girvan e Newman (2002). Na figura, é possível identificar duas comunidades por meio da análise das conexões inter e intra-grupos.

Figura 16 – Rede modular com $n = 60$ vértices gerada a partir do modelo de Girvan e Newman (2002). Para gerar a rede, os vértices foram divididos em dois módulos de igual tamanho e todas as possíveis arestas do grafo simples foram consideradas com probabilidade $p_{in} = 0.3$ para nós do mesmo módulo e $p_{out} = 0.03$ caso contrário.

A investigação de propriedades e dinâmicas para detecção de comunidades é extremamente relevante para muitos sistemas reais. Prova disso é o grande interesse que este tópico desperta na comunidade científica. A Fig. 17 apresenta um exemplo prático de detecção de comunidades sobre a rede “Caratê”, apresentada na Fig. 4. Pela figura, é possível observar a clara divisão dos alunos em função do rompimento das colaborações entre os instrutores da academia (nós 1 e 34 da rede). As comunidades exibidas na figura são obtidas após executar o algoritmo de detecção de comunidades proposto em Newman e Girvan (2004), denominado *Edge-Betweenness*.

2.1.3.5 Otimização estrutural de redes

Enquanto a maior parte dos modelos de redes complexas são descritos em função de algum mecanismo de crescimento e muitos outros trabalhos são fortemente concentrados em

Figura 17 – Detecção de comunidades na rede “Caratê” (Fig. 4) utilizando o algoritmo proposto em [Newman e Girvan \(2004\)](#).

explorar medidas de rede, a otimização de uma meta ou certas características pode ser desejável para alguns problemas. Um exemplo muito intuitivo descrito por [Newman \(2010\)](#) é o projeto *hub-and-spoke* das redes de linhas aéreas, o qual otimiza a eficiência por assegurar que voos de/para menores destinos vão para ou vêm de grandes centros, garantindo aviões mais cheios e oferecendo aos passageiros uma viagem curta.

Na literatura, o principal estudo relacionado à otimização estrutural de redes é apresentado em [Cancho e Solé \(2003\)](#), no qual uma técnica baseada em um modelo simples de busca local (similar ao algoritmo de recozimento simulado) é investigada. Os autores mostraram que quatro tipos de redes bastante conhecidos, tal como apresentado na Fig. 18, podem ser encontrados a partir da técnica, a qual é dirigida por uma função de energia que minimiza o número de arestas p e o menor caminho médio d , de acordo com a seguinte equação:

$$E(\lambda) = \lambda d + (1 - \lambda)p, \quad (2.21)$$

onde $0 \leq \lambda, d, p \leq 1$ e λ é um parâmetro que controla a combinação linear entre d e p , d é a distância normalizada entre vértices denotada por:

$$d = \frac{D}{D^{max}}; D = \frac{1}{\binom{n}{2}} \sum_{i < j} D_{ij}; D^{max} = \frac{(n+1)}{3}, \quad (2.22)$$

onde D_{ij} é a distância mínima entre os vértices i e j , e D^{max} é o valor máximo de D que pode ser alcançado por uma rede conectada. O número de arestas normalizado, p , é dado por:

$$p = \frac{1}{\binom{n}{2}} \sum_{i < j} a_{ij}. \quad (2.23)$$

Outro exemplo de otimização estrutural é apresentado em [Cohen e Havlin \(2010\)](#), que discutem a resistência de redes reais à ataques aleatórios e direcionados, como, por exemplo, o

Figura 18 – Redes encontradas em [Cancho e Solé \(2003\)](#) mediante otimização estrutural: exponencial, estrela, livre de escala e densidade elevada, respectivamente. O eixo y é denotado pela entropia do grau, dada por $H = -\sum_{k=1}^{n-1} p_k \log p_k$, onde p_k é a frequência de vértices com grau k . Figura retirada de [Cancho e Solé \(2003\)](#).

ataque de *hackers* sobre redes de computadores. Nesse sentido, o autor sugere que a formação da rede precisa ser otimamente robusta para suportar tais ataques. Em simulações usando múltiplas ondas de ataques simultâneas, os resultados mostram que o processo de otimização da robustez da rede leva a uma distribuição de grau bi-modal.

No Capítulo 5 da tese, a otimização estrutural é investigada no contexto do aprendizado baseado em redes.

2.2 Aprendizado de máquina

Como maneira de introduzir o aprendizado de máquina, considerem-se os dois problemas a seguir. O primeiro problema é o de ordenação, onde a entrada é um conjunto de números e a saída deve ser a apresentação desses números de maneira ordenada. O segundo problema é o de detecção de *spam*, onde a entrada é um documento de *email* e a saída deve ser uma resposta indicando se a mensagem é um *spam* ou não. Analisando os dois problemas, é possível notar que o problema de ordenação exibe um conceito imparcial que, independentemente da entrada, permite conhecer a saída, enquanto que o conceito de *spam* é bastante subjetivo, uma vez que existe uma definição parcial associada à resolução da tarefa.

Uma solução para resolver o problema de detecção de *spam* é reunir um conjunto de mensagens em um dado domínio, algumas das quais caracterizadas como *spam* por um especialista, e “aprender” a classificar *spam* através delas, de modo que a máquina extraia automaticamente as instruções para esta tarefa. Diferentemente do problema de ordenação, em que não é necessário aprender a ordenar os números, existem muitas aplicações nas quais não há uma sequência de instruções prontas, mas um conjunto de dados a ser aprendido, tais como: análise de crédito bancário, detecção de intrusão em redes, categorização de textos, entre outros. Assim, quando se faz uso do termo “aprender a classificar *spam*”, pode-se, imediatamente, criar uma relação com a habilidade de que o ser humano dispõe para o aprendizado. De fato, tal habilidade torna as pessoas capazes de assimilar conhecimento mediante experiências obtidas nas mais diferentes situações.

Essencialmente baseado nisto, o aprendizado de máquina é uma área de pesquisa relacionada ao desenvolvimento de técnicas computacionais que visam aprender determinados padrões ou comportamentos a partir de observações ou amostras, também denominadas experiências (MITCHELL, 1997; BISHOP, 2006). Logo, o aprendizado de máquina visa reduzir a dependência em relação aos especialistas, tendo cada vez mais importância para o processo de extração do conhecimento, em razão do grande volume e da complexidade dos dados gerados pelas organizações.

Tradicionalmente, o aprendizado de máquina é dividido em dois paradigmas: aprendizado supervisionado e não supervisionado. No aprendizado supervisionado, o objetivo é aprender um mapeamento a partir dos vetores de entrada dos dados para os vetores de saídas desejadas correspondentes. A característica principal é a presença de um supervisor, o qual provê os valores de rótulos corretos para cada entrada. Dentre os problemas mais pesquisados no aprendizado supervisionado, está a classificação de dados, quando os valores assumidos pelos rótulos são discretos, e a regressão, quando os valores são contínuos (BISHOP, 2006; ALPAYDIN, 2009). O aprendizado não supervisionado, também denominado estimação de densidade ou agrupamento, é caracterizado pela ausência de supervisor, ou seja, apenas os vetores de entrada dos dados estão disponíveis e não há qualquer informação de rótulo ou classe. Dessa forma, o objetivo é encontrar regularidades nestes dados, normalmente, a partir da identificação de estruturas onde certos padrões ocorrem mais frequentemente que outros. Entre as principais tarefas, estão o agrupamento de dados e a redução de dimensionalidade (JAIN; MURTY; FLYNN, 1999; SCHAEFFER, 2007; YAN *et al.*, 2007).

Recentemente, um paradigma intermediário entre o supervisionado e não supervisionado tem sido explorado. O aprendizado semissupervisionado usa, simultaneamente, conjuntos de dados rotulados e não rotulados para o aprendizado. A propagação de rótulos é uma das principais tarefas investigadas neste paradigma, especialmente, o aprendizado transdutivo (CHAPELLE; SCHOLKOPF; ZIEN, 2006; ZHU, 2008). Quando há informação sobre rótulos nas bases de dados, o aprendizado também pode ser categorizado em indutivo ou transdutivo. No aprendizado

indutivo, a meta é obter uma função que generalize o conhecimento aprendido para além dos dados de treinamento, compreendendo, assim, todo o espaço dos dados. No caso do aprendizado transdutivo, o objetivo é mais simples, uma vez que o objetivo é obter uma função que realize a predição apenas dos itens de dados não rotulados considerados durante o treinamento.

A Figura 19 exibe uma ilustração básica das principais diferenças entre as tarefas de classificação supervisionada indutiva (a), semissupervisionada transdutiva (b) e de agrupamento de dados (c). Note-se que, na tarefa de agrupamento de dados, os itens de dados são representados por meio de cores apenas para enfatizar a separação deles entre grupos. As subseções seguintes apresentam os conceitos básicos, as principais tarefas, bem como as técnicas mais conhecidas dos paradigmas três do aprendizado de máquina.

Figura 19 – Ilustração da diferença básica entre os principais paradigmas do aprendizado de máquina: (a) classificação supervisionada; (b) classificação semissupervisionada; e (c) agrupamento de dados.

2.2.1 Aprendizado supervisionado

Formalmente, uma tarefa do aprendizado supervisionado pode ser definida da seguinte maneira. Considere-se um conjunto de dados $\mathcal{X}_{train} = \{(x_1, l_1), \dots, (x_n, l_n)\}$, onde o par (x_i, l_i) denota a i -ésima instância do conjunto, e $x_i = (a_1, \dots, a_d)$ representa o vetor de entrada que contém os atributos do domínio d -dimensional dos dados, e $l_i \in \mathcal{L} = \{1, \dots, c\}$ define o vetor

de saída, o qual informa o rótulo (ou classe) associado ao item de dado i dentre o conjunto de c possíveis rótulos.

Há duas etapas principais no processo de aprendizado supervisionado. A primeira delas é o *treinamento*, cujo principal objetivo é aprender um mapa ou classificador $x \rightarrow l$ utilizando o conjunto de dados $\mathcal{X}_{train}$. A segunda etapa é denominada *classificação* ou *predição* e consiste em avaliar este classificador construído por predizer um conjunto de instâncias de teste $\mathcal{X}_{test} = \{(x_{n+1}, ?), \dots, (x_{n+z}, ?)\}$, o qual contém instâncias não vistas no treinamento e desprovidas de rótulos.

Existem algumas suposições em relação aos dados de treinamento e de teste que precisam ser endereçadas, a fim de obter um classificador com bom desempenho de generalização. Entre as hipóteses mais importantes, destacam-se a representatividade dos dados de treinamento, na qual assume-se que os dados de treinamento representam uma amostra representativa da distribuição ou da população por meio da qual foi gerado, e dados de teste imparciais, em que se assegura que os dados de teste não são usados junto aos dados de treinamento. Ambas as hipóteses garantem que um classificador treinado com dados de treinamento representativos deve ser capaz de predizer dados de teste da mesma distribuição. Embora seja comum algum nível de violação dessas suposições na prática, quando fortemente violadas, os resultados de classificação se traduzem em taxas de erro elevadas (MITCHELL, 1997). Dentre as várias técnicas de classificação tradicionais existentes na literatura, algumas das mais usadas são apresentadas a seguir.

- **Árvores de decisão:** uma árvore de decisão é uma técnica hierárquica baseada na estrutura de árvores, a qual consiste de vértices de decisão internos e folhas, os quais tentam inferir um conjunto de regras de decisão para classificar um conjunto de amostras. Em geral, dada uma entrada, em cada vértice de decisão, um teste ou função é aplicado e uma das subárvores é tomada dependendo da saída. O processo inicia do vértice raiz e é repetido, recursivamente, até que se atinge um vértice folha, o qual realiza uma decisão sobre a atribuição da classe. Uma árvore de decisão é ainda um modelo não paramétrico no sentido de que não é assumida qualquer forma paramétrica para as densidades da classe, e a estrutura de árvore não é fixada a priori, mas a árvore cresce, uma vez que subárvores e folhas são adicionadas durante o processo de aprendizado. Cada função define uma discriminante no espaço de entrada de dimensão d , dividindo-o em regiões menores, que são mais subdivididas, à medida que se toma um caminho a partir da descida da raiz (BREIMAN, 1984; QUINLAN, 1986).
- **Aprendizado baseado em instâncias:** estes algoritmos formulam hipóteses diretamente a partir das instâncias de treinamento, uma vez que eles não realizam generalização explícita dos dados, mas comparam as instâncias de teste com instâncias empregadas no treinamento, as quais são armazenadas em memória. De fato, o k -NN é o algoritmo baseado

em instâncias mais conhecido. Ele emprega uma medida de similaridade (normalmente, a distância euclidiana) e classifica uma nova instância por predizer que ela possui a classe majoritária de seus k vizinhos mais próximos (KIDWELL, 2013).

- **Redes bayesianas:** também denominada modelo bayesiano ou rede probabilística, redes bayesianas proveem um *framework* probabilístico para raciocínio sobre incerteza. Informalmente, redes bayesianas são grafos acíclicos direcionados, onde os vértices representam variáveis aleatórias de um domínio e as arestas denotam as dependências condicionais ou informativas existentes entre estas variáveis. Após construída a rede, várias técnicas exatas e aproximadas podem ser empregadas para realizar inferência estatística sobre ela. Entre as técnicas de classificação mais conhecidas no contexto do aprendizado bayesiano, estão o classificador ótimo de Bayes e o classificador ingênuo de Bayes (NEAPOLITAN, 2003).
- **Redes Neurais Artificiais:** Redes neurais artificiais (RNA) têm sido muito empregadas para aprendizado de máquina e reconhecimento de padrões. Basicamente, RNA são sistemas adaptativos compostos por camadas de “neurônios” interconectados que processam informações através desta rede de conexões. Em termos práticos, a rede neural, inicialmente, sem conhecimento, passa por um processo de aprendizado que a torna capaz de representar as dependências entre as variáveis descritivas e de modelar o comportamento a ser explicado. Para construir esse modelo, a rede neural mapeia as relações entre os vetores de entrada e saída através das conexões sinápticas existentes entre os neurônios (HAYKIN, 2009).
- **Máquina de Vetores de Suporte:** baseada no princípio de minimização do risco estrutural, as máquinas de vetores de suporte (SVM) trabalham sobre o conceito de margem e realizam a classificação de dados por construir vários hiperplanos. O termo margem se refere à distância mínima a partir do hiperplano de separação até as instâncias de dados mais próximas. A técnica maximiza a margem, ou seja, cria a maior distância possível entre os hiperplanos de separação e as instâncias próximas a eles. O fato de considerar apenas instâncias próximas às margens é uma característica particular da técnica. Aliás, estas instâncias são denominadas vetores de suporte. Todavia a maioria dos problemas reais envolvem dados não linearmente separáveis, nos quais não existe hiperplano que possa separar instâncias positivas das negativas. Neste caso, a solução é mapear os dados sobre um espaço de maior dimensão (denominado espaço de característica transformado), a fim de definir esses hiperplanos de separação. Outro ponto importante das SVMs é que a implementação desta transformação e a construção do hiperplano de separação no espaço de característica de alta dimensão podem ser substituídas por utilizar uma representação de *kernel*, onde a solução é escrita como uma soma ponderada dos valores de uma certa função *kernel* avaliada nos vetores de suporte (VAPNIK, 2013; HSU; LIN, 2002).

2.2.2 Aprendizado não supervisionado

A principal diferença do aprendizado não supervisionado em relação aos demais paradigmas é a completa ausência de informação de classe ou rótulo (MITCHELL, 1997). Formalmente, seja $\mathcal{X} = \{x_i, i = 1, \dots, n\}$ um conjunto de dados não rotulado onde cada item de dado é descrito por d atributos $x_i = \{x_{i1}, \dots, x_{id}\}$, o objetivo de um algoritmo não supervisionado é inferir conhecimento unicamente a partir da informação inerente ao conjunto de dados (MITCHELL, 1997; BISHOP, 2006).

Confundida muitas vezes com o próprio paradigma de aprendizado não supervisionado, por ser o seu exemplo mais conhecido, a tarefa de agrupamento de dados está diretamente associada à descoberta de conhecimento em bases de dados. O agrupamento consiste em encontrar subconjuntos de dados que compartilham alguma relação ou característica a partir de alguma suposição feita sobre os dados, tal como distribuição, por exemplo (JAIN; MURTY; FLYNN, 1999; BISHOP, 2006). A seguir, são apresentadas as duas principais categorias de algoritmos de agrupamento:

Agrupamento hierárquico. Técnicas pertencentes a esta categoria são caracterizadas por encontrar sucessivos grupos usando grupos estabelecidos previamente. Dois tipos de abordagens são adotadas, a *aglomerativa*, que inicia com cada item de dado representando um único grupo, a partir dos quais grupos maiores serão formados sucessivamente, e a *divisiva*, que inicia com um único grupo contendo todos os itens de dados a partir do qual grupos menores são formados sucessivamente (ALPAYDIN, 2009).

Agrupamento particional. Ao contrário do conjunto de grupos aninhados obtido pelos algoritmos hierárquicos, algoritmos particionais são caracterizados por particionar os itens de dados sem sobreposição de grupos. O exemplo mais conhecido dessa categoria é o algoritmo k -Médias (ALPAYDIN, 2009; JAIN, 2010), onde os itens de dados são particionados em k grupos por alternar os seguintes passos até a convergência do algoritmo:

1. atribua cada item de dado x para o grupo cuja média produza a menor soma dos quadrados intragrupos, i.e.:

$$\min \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2, \quad (2.24)$$

onde μ_i é a média dos pontos no grupo S_i , também denominado centroide de S_i ;

2. calcule-se o centroide dos novos grupos.

Outra tarefa bastante conhecida no aprendizado não supervisionado é a redução de dimensionalidade. Quando os dados encontram-se em um espaço com muitas dimensões, vários problemas podem surgir ao computar a similaridade entre eles, o que é denominado *maldição da*

dimensionalidade (DUDA; HART; STORK, 2007). A técnica mais representativa para essa tarefa é a *análise de componente principal* (PCA), que, basicamente, é uma transformação ortogonal que representa os dados por meio de componentes principais, i.e., as maiores variâncias na projeção dos dados (JOLLIFFE, 2002). Normalmente, um pequeno número de componentes principais é suficiente para manter boa parte da estrutura dos dados.

2.2.3 Aprendizado semissupervisionado

Formalmente, o aprendizado semissupervisionado (SSL) pode ser definido da seguinte maneira. Seja $\mathcal{X}^{(l)} = \{x_i, i = 1, \dots, n\}$ um conjunto de dados onde, para cada item de dado, é atribuída uma classe $l_i = \{l_i \in \mathcal{L} \mid 1 \leq i \leq c\}$. E seja $\mathcal{X}^{(u)} = \{x_i, i = 1, \dots, m\}$ um conjunto de dados onde as instâncias não possuem rótulo. Visto que cada item de dado, independente de pertencer à $\mathcal{X}^{(l)}$ ou $\mathcal{X}^{(u)}$, é descrito por d atributos $x_i = \{x_{i1}, \dots, x_{id}\}$, o objetivo está, normalmente, relacionado a prever as classes dos itens de dados em $\mathcal{X}^{(u)}$, usando os itens de dados rotulados disponíveis. Diferentemente do aprendizado supervisionado, aqui, é assumido que $n \ll m$, i.e., a grande maioria das instâncias não possui rótulo.

De acordo com Chapelle, Scholkopf e Zien (2006), no aprendizado semissupervisionado, todos os dados $\mathcal{X} = \mathcal{X}^{(l)} \cup \mathcal{X}^{(u)}$ são usados ou na etapa de treinamento em caso de técnicas indutivas ou na etapa de classificação em caso de técnicas transdutivas¹. Contudo, devido às características particulares do SSL, um conjunto de suposições, normalmente, precisa ser realizado em relação à distribuição dos dados. A seguir são, apresentadas as três suposições mais frequentes na literatura.

- Suposição de *suavidade*: tal suposição afirma que variações de rótulos de dados em regiões de alta densidade devem ser suaves, i.e., se itens de dados são próximos em uma região de alta densidade, então, seus rótulos devem ser similares também.
- Suposição de *grupos*: tal suposição afirma que, se dois itens de dados estão no mesmo grupo, então, eles devem possuir o mesmo rótulo. Enquanto a suposição de grupos garante consistência global, a suposição de suavidade garante consistência local entre rótulos e grupos. Dessa forma, as duas suposições são complementares, ao considerar grupos bem definidos ou em regiões de alta densidade.
- Suposição de *variedade*: tal suposição afirma que dados, em um espaço com muitas dimensões, podem ser representados em um espaço com menos dimensões, cujas propriedades assegurem maior acerto na estimação de densidade e medidas de similaridades mais apropriadas. Desse modo, dados pertencentes à mesma variedade devem possuir um mesmo rótulo.

¹ O aprendizado transdutivo, ao contrário do indutivo, não busca inferir uma regra de decisão geral, mas apenas prever os rótulos dos itens de dados pertencentes à $\mathcal{X}^{(u)}$.

De acordo com [Chapelle, Scholkopf e Zien \(2006\)](#), as técnicas de SSL podem ser classificadas em técnicas geradoras, técnicas de separação por baixa densidade e técnicas baseadas em grafos sendo uma das principais diferenças entre elas a maneira pela qual cada uma lida com as suposições mencionadas acima. Enquanto, em modelos geradores, a inferência envolve a estimação de uma densidade condicional, como, por exemplo, a técnica de maximização de expectativas (EM) ([NIGAM *et al.*, 2000](#)), nos modelos de separação por baixa densidade, como, por exemplo, a técnica SVM transdutiva (TSVM) ([CORTES; VAPNIK, 1995; VAPNIK; VAPNIK, 1998](#)), a inferência é baseada sobre a definição de fronteiras de decisão preferencialmente afastadas de regiões de alta densidade. Nas técnicas baseadas em grafo, como, por exemplo, a propagação de rótulos ([ZHU; GHAHRAMANI; LAFFERTY, 2003](#)), nós rotulados e não rotulados são associados aos itens de dados e as arestas refletem relações de similaridade entre eles. Por consequência, assume-se que os rótulos devem ser propagados pela rede.

Adicionalmente, a ideia de autotreinamento e cotreinamento também são bastante conhecidas dentro desse paradigma ([ZHU; GOLDBERG, 2009](#)). O autotreinamento considera a classificação de dados não rotulados iterativamente a partir dos dados rotulados disponíveis. Um classificador é retreinado a cada iteração e classifica os itens de dados não rotulados com maior certeza, os quais se juntam aos demais dados rotulados para treinar o classificador na iteração seguinte. Por sua vez, o cotreinamento requer dois classificadores distintos, usando duas visões distintas sobre os dados. Alternadamente, cada classificador é treinado com uma das visões sobre os dados e classifica os itens de dados não rotulados com maior confiabilidade. Então, esses itens de dados são adicionados aos dados rotulados e os classificadores ensinam-se mutuamente por prover exemplos rotulados para o outro ([BLUM; MITCHELL, 1998](#)).

2.3 Aprendizado de máquina baseado em redes

Nesta seção, técnicas relevantes para o contexto do trabalho, i.e., relacionadas ao aprendizado baseado grafo são apresentadas. As aplicações das técnicas contemplam conceitos importantes para os desenvolvimentos da tese. Logo, trabalhos que investigam formação do grafo, medidas de rede, passeio aleatório e propagação de rótulos serviram como inspiração para as contribuições apresentadas nos capítulos seguintes. A seção é organizada da seguinte maneira:

- A Subseção [2.3.1](#) apresenta uma breve revisão da literatura sobre métodos para construção da rede. Apesar de poucos trabalhos se dedicarem especificamente à investigação da construção da rede, o processo de inferência por trás do aprendizado baseado em grafos deriva dela. Consequentemente, a rede tem papel crucial na motivação, projeto e desempenho de técnicas baseadas em grafo. Além de apresentar métodos de formação de redes usados no desenvolvimento desta tese, esta subseção também serve como motivação para a investigação apresentada no Capítulo [5](#), a qual está relacionada à otimização estrutural de redes.

- As principais técnicas baseadas em redes usadas para classificação de dados no contexto supervisionado são apresentadas na Subseção 2.3.2. Enquanto o aprendizado baseado em grafos tem sido bastante explorado em tarefas do aprendizado não supervisionado e semisupervisionado, existem poucas investigações relacionadas ao aprendizado supervisionado. O objetivo dessa subseção é apresentar características importantes dos principais trabalhos existentes na literatura, as quais serviram de motivação para as investigações apresentadas nos Capítulos 3 e 4.
- As técnicas de propagação de rótulos utilizadas para investigar a tarefa de anotação automática de papéis semânticos no contexto semissupervisionado são apresentadas na Subseção 2.3.3. O *framework* desenvolvido para tal investigação, bem como os resultados obtidos são exibidos no Capítulo 6.

2.3.1 Métodos de formação de rede

Em técnicas de classificação baseada em redes, os vértices e arestas representam, respectivamente, os itens de dados e as relações existentes entre eles. Diferentemente dos algoritmos tradicionais, a etapa do treinamento envolve, principalmente, a formação da rede a partir da conexão entre os vértices. A construção do grafo é uma etapa crucial para qualquer técnica baseada em rede, pois, a partir da rede, são extraídas informações estruturais, dinâmicas e funcionais dos dados. Apesar da grande relevância, na literatura, há poucos algoritmos relacionados à construção da rede (ZHU; LAFFERTY; ROSENFELD, 2005). Os métodos de formação mais empregados são k vizinhos mais próximos (k NN) e vizinhança de raio ϵ (ϵ N) (BELKIN; NIYOGI, 2003; ZHU, 2008). Nas próximas subseções, são apresentados conceitos relacionados a essas e outras técnicas de formação da rede.

2.3.1.1 Rede completamente conectada

Considere-se a matriz de similaridade S obtida a partir da base de dados original, uma abordagem direta para construir uma rede seria estabelecer ligações entre pares de nós com pesos definidos de acordo com S_{ij} . Contudo tal abordagem, normalmente, resulta em redes completas, o que é muito similar à representação da matriz de similaridade original. Como a rede obtida é densa, há que ser mencionado ainda o alto custo computacional para lidar com tais redes, o qual pode ser um fator limitante para a sua aplicação em muitos problemas reais. Ademais, alguns trabalhos apontam empiricamente que redes completamente conectadas apresentam pior desempenho que redes esparsas (ZHU; LAFFERTY; ROSENFELD, 2005).

2.3.1.2 Rede k vizinhos mais próximos

As abordagens mais estudadas para construção do grafo a partir de bases de dados do tipo vetor de atributos são baseadas em relações de vizinhança (BERTINI *et al.*, 2011), onde a rede k

vizinhos mais próximos (k NN) tem sido, quase exclusivamente, usada na literatura (OZAKI *et al.*, 2011). k NN é um método de formação da rede de propósito geral, onde cada item de dado é representado por um vértice e há uma conexão de um vértice v_i para outro v_j se e somente se x_j pertence ao conjunto dos k elementos mais similares à x_i , definido como $k\text{NN}^{x_i}$, i.e.:

$$e_{ij} = \begin{cases} 1, & \text{se } x_j \in k\text{NN}^{x_i}, \\ 0, & \text{caso contrário.} \end{cases} \quad (2.25)$$

Pela definição, k NN produz um grafo direcionado esparsos para pequenos valores de k . Há ainda algumas variações desse método, as quais são bastante usadas, especialmente, por tornar o grafo resultante não direcionado. O k NN Simétrico (Sk NN), o qual gera uma matriz de adjacência simétrica A' a partir da matriz de adjacência original A , obtida pelo k NN da seguinte forma:

$$A'_{ij} = \max(A_{ij}, A_{ij}^T). \quad (2.26)$$

Outro método derivado é a rede k NN Mútuo (Mk NN), onde a matriz de adjacência simétrica A' é obtida por:

$$A'_{ij} = \min(A_{ij}, A_{ij}^T). \quad (2.27)$$

Se, por um lado, os métodos baseados em k vizinhos, normalmente, produzem grafos esparsos os quais oferecem vantagens sobre redes completamente conectadas, por outro lado, há o problema de saber o melhor valor para o parâmetro k .

2.3.1.3 Rede vizinhança de raio ε

O método de vizinhança de raio ε (ε N) constrói uma rede não direcionada cujas arestas consistem de pares de vértices (v_i, v_j) , que seguem a seguinte fórmula:

$$e_{ij} = \begin{cases} 1, & S_{ij} \geq \varepsilon, \\ 0, & \text{caso contrário.} \end{cases} \quad (2.28)$$

De fato, a rede de vizinhança de raio ε permite uma grande variação do grau de cada vértice da rede de acordo com a distribuição dos dados originais. Contudo tal método possui o problema em determinar um valor adequado para ε .

2.3.1.4 Outros métodos da literatura

Os métodos de construção de rede k vizinhos mais próximos e vizinhança de raio ε tratam os itens de dados de maneira uniforme em regiões densas e esparsas. Recentemente, alguns métodos de formação da rede que empregam informações adaptativas têm sido propostos. Alguns exemplos são:

- a combinação de k NN e ϵ N para construção da rede, de modo que o k NN seja usado para definir conexões em regiões esparsas, enquanto ϵ N para regiões densas (SILVA; ZHAO, 2012a; SILVA; ZHAO, 2015);
- o método *b-matching*, que, ao contrário do k NN, assegura que cada nó na rede possui o mesmo número de ligações b e, portanto, produz uma rede balanceada ou regular (JEBARA; WANG; CHANG, 2009);
- o método de vizinhança linear, cuja ideia é aproximar a rede inteira por uma série de caminhos de vizinhança linear sobrepostos, sendo que os pesos das arestas em cada caminho são determinados por um procedimento de programação quadrática padrão (WANG; ZHANG, 2008);
- o método de formação da rede baseado em heurísticas de agrupamento emprega o método *single linkage*, heurística de agrupamento capaz de construir redes conectadas e esparsas, enquanto mantém a estrutura de agrupamento original dos dados (CUPERTINO; HUERTAS; ZHAO, 2013);
- o método AdaRadius é baseado na combinação de árvore geradora mínima e ϵ N. Inicialmente, MST é usado para fornecer uma espécie de esqueleto da rede a partir da matriz de similaridade dos dados originais. Então, um valor ϵ_i para cada vértice v_i é definido como o tamanho da sua aresta mais distante na MST. Finalmente, a rede é construída usando o método ϵ N considerando o valor de ϵ definido para cada nó (ARAÚJO; ZHAO, 2016);
- enquanto a maioria dos métodos de formação da rede orientados à tarefa são desenvolvidos para o aprendizado não supervisionado ou semissupervisionado, k AOG é projetado para o aprendizado supervisionado, especificamente, para a classificação de dados. Como k NN, k AOG cria uma ligação direcionada de um vértice v_i para v_j combinando k NN e uma medida de *pureza*, a qual expressa o nível de mistura de um componente em relação a outros componentes de classes distintas usando a informação de rótulo dos dados de treinamento. Basicamente, o método k AOG combina as redes k NN subsequentes por aumentar o valor de k , enquanto aumentando ou mantendo a *pureza* da rede encontrada até então, até que a rede ótima medida pela *pureza* seja alcançada (BERTINI *et al.*, 2011).

2.3.2 Classificação supervisionada baseada em grafos

O aprendizado baseado em redes é amplamente investigado no contexto do aprendizado não supervisionado e semissupervisionado (CHAPELLE; SCHOLKOPF; ZIEN, 2006; FORTUNATO, 2010; JAIN, 2010), especialmente em tarefas de agrupamento de dados, propagação de rótulos e redução de dimensionalidade (BELKIN; NIYOGLI, 2003; SCHAEFFER, 2007; YAN *et al.*, 2007; ZHU, 2008; SILVA; ZHAO, 2012c; SILVA; ZHAO, 2012b). Por outro lado, há

muito espaço para a investigação e desenvolvimento de novas técnicas de aprendizado supervisionado baseado em grafos. De fato, existem poucas investigações relacionadas ao uso de redes para problemas supervisionados, especialmente, aqueles em que os dados são representados na forma de vetor de atributos (BERTINI *et al.*, 2011; SILVA; ZHAO, 2012a). Se, por um lado, há várias técnicas de classificação semissupervisionadas, que, em princípio, poderiam ser aplicadas para problemas supervisionados (SINDHWANI; NIYOGI; BELKIN, 2005; BELKIN; NIYOGI; SINDHWANI, 2006), por outro lado, tais técnicas foram projetadas para tomar vantagens de itens de dados rotulados e não rotulados durante a etapa de treinamento, de modo que o grafo é normalmente usado para modelar os dados em uma representação que permita a propagação de rótulos. Por isso, em problemas nos quais uma quantidade razoável de dados rotulados está disponível, uma abordagem propriamente supervisionada é preferível (BERTINI *et al.*, 2011).

Tradicionalmente, algoritmos de classificação recebem como entrada um vetor de atributos, o qual representa um item de dado em termos de características numéricas ou categóricas. Por outro lado, no aprendizado relacional estatístico, a representação de um item de dado normalmente está associada às suas relações com outras instâncias, estando os dados, naturalmente, representados na forma de um grafo (NICKEL *et al.*, 2015). Dessa forma, a classificação relacional viola a suposição de independência dos dados de modo que a classe de uma instância pode não depender apenas de seus próprios atributos, mas também da classe de seus vizinhos (MACSKASSY; PROVOST, 2003). As aplicações do aprendizado relacional incluem reconhecimento de cadeias moleculares em expressões gênicas (SEGAL; WANG; KOLLER, 2003), classificação de artigos científicos relacionados (LU; GETOOR, 2003) e a predição de ligações (HASAN *et al.*, 2006; LIBEN-NOWELL; KLEINBERG, 2007).

Ainda no contexto do aprendizado relacional, um tópico denominado inferência coletiva consiste em propagar rótulos de vértices pré-rotulados para vértices não rotulados, considerando a interligação entre eles de maneira conjunta (MCDOWELL; GUPTA; AHA, 2009). Diz-se coletiva porque informações dos rótulos e dos atributos dos vértices podem ser consideradas isoladamente, em conjunto com seus vizinhos e também considerando vértices não rotulados (SEN *et al.*, 2008). Entre as técnicas mais conhecidas dessa categoria, estão classificadores locais como ICA (*iterative classification algorithm*) (LU; GETOOR, 2003) e Gibbs (JENSEN; NEVILLE; GALLAGHER, 2004), e classificadores globais, onde o treino e a inferência são realizados pelo mesmo algoritmo, como LBP (*loopy belief propagation*) (SEN *et al.*, 2008). Outro trabalho relacionado ao assunto é o *framework* geral de classificação em redes, proposto em Macskassy e Provost (2007), onde um classificador local indutivo, um classificador relacional e um componente de inferência coletiva são empregados em conjunto.

Diferentemente do aprendizado relacional, as principais contribuições desta tese não se referem a técnicas de propagação de rótulos na rede. Pelo contrário, o objetivo é explorar propriedades e medidas de redes complexas para caracterizar a formação de padrão da rede durante a classificação de cada item de teste. Apesar deste ser um tópico bastante recente, os resultados

obtidos mostram que as técnicas desenvolvidas até então evidenciam vantagens em alguns aspectos sobre os algoritmos tradicionais, tais como a ausência de parâmetros e as habilidades para detectar classes de diferentes formas e capturar relações espaciais, topológicas e funcionais dos dados. A seguir, três abordagens de classificação baseada em rede são apresentadas. A primeira delas é uma técnica de classificação sem parâmetros, denominada grafo k -associados ótimo (BERTINI *et al.*, 2011). A segunda, denominada classificação de alto nível, usa medidas de redes complexas, a fim de detectar a formação de padrão dos dados (SILVA; ZHAO, 2012a). A terceira técnica é baseada em passeio aleatório e utiliza probabilidades limitantes para classificar novos itens de dados (CUPERTINO; ZHAO; CARNEIRO, 2015). Além das técnicas de classificação propriamente ditas, um *framework* baseado em grafos para redução de dimensionalidade no contexto supervisionado (CARNEIRO; CUPERTINO; ZHAO, 2014) também é apresentado nesta subseção.

2.3.2.1 Grafo k -associados ótimo para classificação de dados

A técnica de aprendizado supervisionado baseada no grafo k -associados ótimo pode ser caracterizada por meio de dois conceitos principais: (i) uma medida de *pureza* que usa a abstração gráfica da rede para medir o nível de mistura entre os componentes de classes distintas; e (ii) o grafo k -associados ótimo (k AOG), o qual pode ser considerado uma versão adaptativa e sem parâmetros do grafo k vizinhos mais próximos. Aliás, a ausência de seleção de parâmetro é uma das principais vantagens da técnica apresentada a seguir (BERTINI *et al.*, 2011).

Dado um conjunto de dados rotulados, a etapa de treinamento consiste na construção da rede k AOG, a qual representa os itens de dados em um grafo no qual os componentes armazenam informação local (*pureza*) sobre a distribuição dos dados. Na etapa de classificação, o mesmo grafo é usado por um classificador para estimar a probabilidade de um item de dado não rotulado pertencer a um dado componente (representando uma classe de dados).

Treinamento: A fase de treinamento, ou construção da rede k AOG, pode ser explicada em três passos.

1. No primeiro, os dados de treinamento na forma de vetor de atributos são representados como um dígrafo, denominado grafo k -associados, o qual pode ser associado a um grafo k vizinhos mais próximos cujas conexões apenas se dão entre os vizinhos que possuem a mesma classe. O Alg. 1 apresenta os passos para construção do grafo k -associados, o qual é usado para obtenção do grafo k -associados ótimo no terceiro passo. Ele recebe como entrada um conjunto de dados e um parâmetro k , relacionado ao número de vizinhos a serem considerados. Nas linhas um e dois, inicializam-se os componentes e o grafo k -associados. Entre as linhas três e seis, tem-se um laço que determina as conexões para cada vértice v_i , a princípio verificando os k vizinhos mais próximos de v_i ($\Lambda_{v_i,k}$) e a igualdade entre as classes. Na linha sete há uma função para encontrar cada componente

do grafo. O restante do algoritmo adiciona cada um dos componentes encontrados ao grafo k -associados e calcula uma medida de pureza sobre eles, retornando tais informações como saída do algoritmo.

Algoritmo 1: Construção do grafo k -associados (kac)

Require: k e conjunto de dados X

```

1:  $C \leftarrow \emptyset$ 
2:  $\mathcal{G}^{(k)} \leftarrow \emptyset$ 
3: for all  $v_i \in V$  do
4:    $\Delta_{v_i,k} \leftarrow \{v_j | v_j \in \Lambda_{v_i,k} \text{ e } l_j = l_i\}$ 
5:    $E \leftarrow E \cup \{e_{i,j} | v_j \in \Delta_{v_i,k}\}$ 
6: end for
7:  $C \leftarrow \text{encontraComponentes}(V, E)$ 
8: for all  $C_\alpha \in C$  do
9:    $\Phi_\alpha \leftarrow \text{calculaPureza}(C_\alpha)$ 
10:   $\mathcal{G}^{(k)} \leftarrow \mathcal{G}^{(k)} \cup \{(C_\alpha(V', E'); \Phi_\alpha)\}$ 
11: end for
12: return Grafo  $k$ -associados  $\mathcal{G}^{(k)}$ 

```

2. O segundo passo consiste no cálculo da pureza de cada um dos componentes do grafo k -associados. Para determinar a medida de pureza de um componente, é preciso considerar o percentual de ligações que ocorrem dentro do componente em relação ao número total de conexões possíveis (soma das conexões que ocorrem e que não ocorrem). Assim, a fórmula para cálculo da pureza de um componente é dada por:

$$\Phi_\alpha = \frac{D_\alpha}{2k}, \quad (2.29)$$

onde D_α representa o grau médio do componente α . Assim, valores de pureza próximos a 1 indicam que o componente apresenta elevada compacidade, enquanto valores baixos de pureza evidenciam elevada mistura entre itens de dados de classes distintas. A Fig. 20 apresenta as medidas de purezas calculadas para três componentes com diferentes níveis de mistura entre as classes.

3. No terceiro passo, a rede k -associados ótimo ($kAOG$) é formada mediante a estimação do valor de k para cada componente. Intuitivamente, um único valor de k limita a estrutura e a pureza dos componentes, uma vez que raramente produzirá a melhor configuração de itens de dados dentro dos componentes para um dado conjunto de dados. Por outro lado, seria interessante uma abordagem que permitisse diferentes valores de k de modo a ajustar melhor a pureza e o tamanho dos componentes às características das bases de dados. Desse modo, o grafo k -associados ótimo visa obter a melhor organização de dados dentro dos componentes, por permitir que os componentes da rede sejam formados a partir de diferentes valores de k . Em outras palavras, a ideia por trás da rede k -associados é incrementar o valor de k enquanto maximiza a pureza dos componentes da rede. Desse

Figura 20 – Medida de *pureza* calculada sobre componentes com diferentes níveis de mistura de classes. Figura retirada de Bertini *et al.* (2011).

modo, para cada valor de k , a pureza dos componentes obtidos é calculada e usada na comparação com outros grafo k -associados gerados a partir de diferentes valores de k , de modo que o componente com maior pureza é mantido, enquanto os outros são descartados. O Alg. 2 apresenta a técnica para construção do grafo k -associados ótimo. O algoritmo não requer nenhum parâmetro e utiliza como entrada apenas o conjunto de dados de treino para construção do classificador. Nas primeiras linhas, k é inicializado com valor 1 e o grafo 1-associado é tomado como o grafo ótimo ($\mathcal{G}^{ot}$). A partir daí, inicia-se um laço cuja condição de término é uma heurística que visa garantir a continuação do algoritmo até que ele alcance o grafo com grau máximo de pureza (BERTINI *et al.*, 2011). Enquanto tal condição de parada não é satisfeita k é incrementado. Entre as linhas sete e doze, o algoritmo verifica, para cada um dos componentes deste grafo k -associados ($C_\beta^{(k)}$), se a condição dada pela linha oito do algoritmo é satisfeita. Caso seja, na linha nove, tem-se uma operação para remover do grafo ótimo cada um dos componentes que compõe $C_\beta^{(k)}$, ou seja, $C_\alpha^{(k)} \subseteq C_\beta^{(k+z)}$. Na linha seguinte, o novo componente é acrescentado ao $\mathcal{G}^{ot}$. Por fim, o algoritmo retorna cada componente obtido, com seu respectivo valor de k e pureza.

Classificação: Após a formação da rede k AOG usando os dados de treinamento, a etapa de classificação de novos itens de dado pode ser realizada. Nesta fase, a rede k AOG é utilizada por um classificador ótimo de Bayes, para fazer a classificação de novas instâncias. Supondo uma nova instância $v_y \in \mathcal{X}_{test}$, o classificador ótimo de Bayes calcula a probabilidade de y pertencer a cada um dos componentes. A partir da teoria de Bayes, a probabilidade a posteriori de v_y pertencer ao componente C_α dado todos os vizinhos observados deste novo caso (Λ_{v_y}), é representada por:

$$P(v_y \in C_\alpha | \Lambda_{v_y}) = \frac{P(\Lambda_{v_y} | v_y \in C_\alpha) P(v_y \in C_\alpha)}{P(\Lambda_{v_y})}, \quad (2.30)$$

onde a probabilidade a priori é calculada usando a *pureza* normalizada dos componentes que possuem vértices vizinhos de v_y . A seguir, cada termo que compõe essa formulação é descrito em detalhes.

Algoritmo 2: Construção do grafo k -associados ótimo

Require: Conjunto de dados X_{train}

- 1: $k \leftarrow 1$
- 2: $\mathcal{G}^{(Or)} \leftarrow kac(k, X)$
- 3: **repeat**
- 4: $ultimoGrauMedio \leftarrow D^{(k)}$
- 5: $k \leftarrow k + 1$
- 6: $\mathcal{G}^{(k)} \leftarrow kac(k, X)$
- 7: **for all** $C_\beta^{(k)} \subset \mathcal{G}^{(k)}$ **do**
- 8: **if** $\Phi_\beta^{(k)} \geq \Phi_\alpha^{(Or)}$ para todo $C_\alpha^{(Or)} \subseteq C_\beta^{(k)}$ **then**
- 9: $\mathcal{G}^{(Or)} \leftarrow \mathcal{G}^{(Or)} - \bigcup_{C_\alpha^{(Or)} \subseteq C_\beta^{(k)}} C_\alpha^{(Or)}$
- 10: $\mathcal{G}^{(Or)} \leftarrow \mathcal{G}^{(Or)} \cup \{C_\beta^{(k)}\}$
- 11: **end if**
- 12: **end for**
- 13: **until** $D^{(k)} - ultimoGrauMedio < D^{(k)}/k$
- 14: **return** Grafo k -associados ótimo $\mathcal{G}^{(Or)}$

Uma vez que cada componente no grafo k -associados ótimo possui o seu próprio valor de k , para estimar a probabilidade da vizinhança Λ_{v_y} dado que $v_y \in C_\alpha$ é necessário calcular o número de conexões de v_y para C_α considerando os valores de k que cada componente assume (k_α). Assim, assumindo-se que Λ_{v_y, k_α} representa o conjunto dos k_α vizinhos mais próximos de v_y e $|\cdot|$ aponta para a cardinalidade do conjunto, a probabilidade da vizinhança Λ_{v_y} condicionada ao fato de que $v_y \in C_\alpha$ é dada por:

$$P(\Lambda_{v_y} | v_y \in C_\alpha) = \frac{|\Lambda_{v_y, k_\alpha} \cap C_\alpha|}{k_\alpha}. \quad (2.31)$$

O termo de normalização $P(\Lambda_{v_y})$, apresentado na Eq. 2.30, pode ser determinado considerando todos os componentes C_β (tal que $C_\beta \leq k_\alpha$) aos quais o novo vértice v_y se conecta ($N_{v_y, C_\beta} \neq 0$, onde N_{v_y, C_β} é o número de conexões de v_y para C_β), i.e.:

$$P(\Lambda_{v_y}) = \sum_{N_{v_y, C_\beta} \neq 0} P(\Lambda_{v_y} | v_y \in C_\alpha) P(v_y \in C_\alpha). \quad (2.32)$$

De fato, a medida de pureza, quando normalizada, pode ser vista como a probabilidade a priori de conexões dentro de um componente (BERTINI *et al.*, 2011). Logo, o processo de normalização da pureza na classificação de um novo vértice v_y considera apenas os componentes aos quais v_y se conecta, tal como:

$$P(v_y \in C_\alpha) = \frac{\Phi_\alpha}{\sum_{N_{v_y, C_\beta} \neq 0} \Phi_\beta}. \quad (2.33)$$

Sabe-se que no grafo k -associados ótimo o número de componentes é maior ou igual ao número de classes. Assim, para definir a probabilidade de um novo vértice pertencer a uma dada

classe é preciso somar as probabilidades a posteriori de todos os componentes da mesma classe, i.e.:

$$P(v_y|\omega_i) = \sum_{l(C_\alpha)=\omega_i} P(v_y \in C_\alpha|\Lambda_{v_y}), \quad (2.34)$$

onde $l(C_\alpha)$ está relacionado à classe dos vértices no componente C_α .

Finalmente, para determinar a classe em que o novo vértice v_y será classificado, escolhe-se a classe com maior probabilidade a posteriori, de acordo com:

$$\varphi(v_y) = \operatorname{argmax}\{P(v_y|\omega_1), \dots, P(v_y|\omega_M)\}. \quad (2.35)$$

2.3.2.2 Classificação de alto nível

A técnica de classificação de alto nível proposta em [Silva e Zhao \(2012a\)](#) consiste em um *framework* híbrido para combinação de classificadores de baixo e alto nível. Diferentemente das técnicas tradicionais de classificação de dados, também denominadas de baixo nível, a classificação de alto nível considera não apenas os atributos físicos dos dados (distância, similaridade ou distribuição), mas também o padrão de formação deles, o qual é capturado a partir de medidas de redes complexas. Logo, as medidas são empregadas a fim de melhorar o desempenho preditivo dos classificadores tradicionais.

A técnica de alto nível pode ser dividida em dois passos: treinamento e classificação. No treinamento, os dados rotulados são utilizados para a formação da rede. Na classificação, medidas de redes complexas são calculadas antes e após a inserção dos itens de teste considerando as redes construídas para as classes de dados. Havendo uma grande variação das medidas de rede para uma dada classe, diz-se que o item de teste não está em conformidade com a formação de padrão expressa pelos dados daquela classe. Caso a variação das medidas seja pequena, diz-se que há conformidade de padrão entre a nova instância e os dados da classe.

Treinamento: Como qualquer outra técnica de aprendizado baseado em rede, o primeiro passo da classificação de alto nível é a construção da rede a partir dos dados de entrada na forma de vetores de atributos. Nesse contexto, [Silva e Zhao \(2012a\)](#) propuseram a combinação de k NN e ε N para a construção da rede, de modo que o critério k NN seja usado para definir conexões em regiões esparsas enquanto o critério ε N para regiões densas (veja Subseção 2.3.1), como apresentado a seguir:

$$\mathcal{E}(x_i) = \begin{cases} \varepsilon N(x_i), & \text{se } |\varepsilon N(x_i)| > k, \\ kNN(x_i), & \text{caso contrário.} \end{cases} \quad (2.36)$$

O algoritmo conecta um dado vértice i utilizando o critério ε N se o número de conexões alcançadas é maior que o valor de k , caso contrário, as conexões são determinadas pelo critério k NN. Como as conexões são realizadas apenas considerando os vértices de mesma classe, uma rede $\mathcal{G}_c$ é formada para cada classe separadamente, resultando em um total de $|\mathcal{L}|$ redes.

Classificação: O segundo passo da técnica de alto nível compreende o uso das redes formadas para classificar novos itens de dados. Na classificação de alto nível, isso é feito por calcular medidas de rede para cada rede antes e após a inserção de um novo item de dado. Dado um item de teste $y \in \mathcal{X}_{test}$, a classificação de alto nível examina o impacto de inserir y para cada classe de maneira isolada. Basicamente, a variação dos resultados obtidos pelas medidas de rede indicam a classe à qual y pertence. Se a inserção de y em $\mathcal{G}_c$ causa uma pequena variação nas medidas de rede, a classificação de alto nível produzirá um grande valor de associação de y para a classe c , indicando que y está em conformidade com este padrão. Do contrário, se a inserção de y em $\mathcal{G}_c$ causa grande variação das medidas, a técnica retorna um pequeno valor de associação para a classe c , indicando que o item de teste, provavelmente, não pertence àquela classe.

Formalmente, a classificação de um dado item de teste y pelo *framework* híbrido proposto em [Silva e Zhao \(2012a\)](#) pode ser descrita por:

$$\mathcal{M}_y^{(c)} = (1 - \lambda) \mathcal{C}_y^{(c)} + \lambda \mathcal{H}_y^{(c)}, \quad (2.37)$$

com $\mathcal{M}_y^{(c)}$ denotando a associação produzida pelos algoritmos de baixo e alto nível, quando y é avaliado para a classe c . Ainda em relação à equação, $\mathcal{C}_y^{(c)} \in [0, 1]$ estabelece a associação produzida por um classificador tradicional; $\mathcal{H}_y^{(J)} \in [0, 1]$ a associação produzida pelo classificador de alto nível; e $\lambda \in [0, 1]$ é o parâmetro de combinação convexa entre os dois classificadores.

A classificação de alto nível de um item de teste y para uma dada classe c é dada por:

$$H_y^{(c)} = \frac{\sum_{u=1}^Z \delta(u) [1 - f_y^{(c)}(u)]}{\sum_{g \in L} \sum_{u=1}^Z \delta(u) [1 - f_y^{(g)}(u)]}, \quad (2.38)$$

onde $\mathcal{H}_y^{(c)} \in [0, 1]$, u denota as medidas de redes empregadas pelo algoritmo de alto nível, $\delta(u) \in [0, 1]$ e $\sum_{u=1}^Z \delta(u) = 1$, $\forall u \in \{1 \dots, Z\}$ são parâmetros que indicam a influência de cada medida de rede no processo de classificação, e $f_y^{(c)}(u)$ provê uma resposta, se o item de teste y apresenta conformidade com os padrões da classe c ou não, considerando a u -ésima medida de rede aplicada. O denominador na equação tem propósito apenas de normalização.

Sobre a função $f_y^{(c)}(u)$, ela é dada por:

$$f_y^{(c)}(u) = \Delta G_y^{(c)}(u) p^{(c)}, \quad (2.39)$$

no qual $\Delta G_y^{(c)}(u) \in [0, 1]$ representa a variação que ocorre em uma medida de rede sempre que um item de teste é inserido em $\mathcal{G}_c$, e $p^{(c)} \in [0, 1]$ é a proporção de instâncias de treinamento que pertencem à classe c . Exemplos de medidas de rede usadas para capturar relações semânticas entre os dados são: assortatividade, coeficiente de aglomeração, grau médio, entre outras.

2.3.2.3 Classificação de dados por facilidade de acesso usando passeio aleatório

A técnica proposta em [\(CUPERTINO; ZHAO; CARNEIRO, 2015\)](#) usa o processo de passeio aleatório para classificar dados por facilidade de acesso. Tal processo pode ser

entendido em termos de cadeias de Markov por meio de três definições básicas, espaço de estados, probabilidade de transição e probabilidades limitantes: i) Uma cadeia de Markov denota uma sequência de estados visitados por um caminhante aleatório em que a probabilidade de ir de um estado para outro é independente dos estados anteriores, visto que o estado atual é conhecido (GALLAGER, 2012); (ii) A probabilidade de transição é a probabilidade de um caminhante aleatório ir de um estado para outro; e (iii) sob certas condições e após um número infinito de transições, o processo de passeio aleatório atinge um estado estacionário, denominado probabilidades limitantes. A partir do conceito de espaço de estados, probabilidade de transição e probabilidades limitantes, a técnica pode ser descrita em três passos.

1. O primeiro passo é a construção da rede, denominada rede de treinamento. Basicamente, esse passo consiste em mapear os itens de dados rotulados como conjunto de espaço de estados de um caminhante aleatório, cujas similaridades entre eles representam as probabilidades de transição entre os estados. Dessa maneira, os vértices da rede representam estados e as arestas probabilidades de transição entre eles;
2. Na etapa de classificação, dado um item de teste $y \in \mathcal{X}_{test}$, sua similaridade é calculada em relação a todos os outros nós da rede e adicionada como viés a todas as probabilidades de transição associadas a cada nó. Dessa forma, a rede obtida no passo anterior é modificada de acordo com o viés de informação de cada item de dado não rotulado. A nova rede é denominada rede de classificação;
3. Finalmente, as probabilidades limitantes são calculadas sobre a rede de classificação e os vértices rotulados mais facilmente alcançados (com maior probabilidade) representam o rótulo de y .

2.3.2.4 Redução de dimensionalidade para classificação de dados

As investigações conduzidas em Cupertino, Carneiro e Zhao (2013), Carneiro, Cupertino e Zhao (2014) adaptam diferentes métodos de formação da rede para a tarefa de redução de dimensionalidade, usando o *framework* proposto em Yan *et al.* (2007). Neste *framework*, dois grafos carregam informações distintas sobre os dados. O primeiro grafo, denominado grafo intrínseco W , carrega a informação sobre a compacidade intraclasse dos componentes, enquanto o segundo grafo, denominado grafo de penalização W^P , carrega conexões de penalização as quais refletem a separabilidade interclasses. A Fig. 21 apresenta uma ilustração geral sobre os grafos. Após a formação das redes, o Laplaciano de ambos os grafos, denotadas por L e L^P , são obtidos e inseridos no *framework* o qual é otimizado em termos de um vetor de projeção $\mathbf{w}$, tal como:

Dado um conjunto de dados $\mathcal{X}$, o *framework* é basicamente composto pelos seguintes passos:

Figura 21 – Grafo intrínseco e grafo penalidade empregados pelo *framework* de redução de dimensionalidade. Retirado e adaptado de Yan *et al.* (2007).

1. Construção dos grafos intrínseco e penalidade a partir de $\mathcal{X}$;
2. Cálculo do Laplaciano combinatório de ambos os grafos, denotados por L e L^P ;
3. Otimização do *framework* em termos de um vetor de projeção $\mathbf{w}$, tal como:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{\mathbf{w}^T X L X^T \mathbf{w}}{\mathbf{w}^T X L^P X^T \mathbf{w}}. \quad (2.40)$$

o qual pode ser resolvido por $X L X^T \mathbf{w} = \lambda X L^P X^T \mathbf{w}$.

Especificamente, nos trabalhos apresentados em Cupertino, Carneiro e Zhao (2013), Carneiro, Cupertino e Zhao (2014), os métodos de formação da rede k NN, ϵ N e k AOG foram adaptados para prover tanto o grafo intrínseco quanto o grafo penalidade. Após o processo de redução das dimensões, um classificador 1NN é usado para avaliar se além da redução do custo computacional, o desempenho preditivo também é melhorado. Vários experimentos comparando o *framework* com técnicas de redução de dimensionalidade bastante conhecidas, como PCA e *Linear Discriminant Analysis* (LDA), usando bases de dados de imagens e de outros domínios, evidenciam que a abordagem é promissora e que a investigação de novas heurísticas para formação dos grafos pode trazer resultados ainda melhores.

2.3.3 Classificação semissupervisionada baseada em grafos

Nesta subseção são apresentadas duas estratégias de propagação de rótulos relevantes ao contexto do trabalho. Embora tais técnicas sejam, normalmente, propostas usando um método de rede específico, elas podem ser combinadas com quaisquer outros métodos de formação de rede de propósito geral. Formalmente, dado um conjunto de itens de dados $\mathcal{X} = \{x_i, i = 1 \dots, l, \dots, n\}$ e um conjunto de rótulos $\mathcal{L} = \{1, \dots, c\}$, apenas os l primeiros itens possuem rótulos conhecidos $\{y_1, \dots, y_l\} \in \mathcal{L}$. Assim, o objetivo é prever a classe dos itens de dados sem rótulo. A seguir são definidos termos importantes para a posterior explicação dos algoritmos de propagação de rótulos:

- F' denota o conjunto de matrizes $n \times c$ com entradas não negativas, onde a matriz $F = [F_1^T, \dots, F_n^T]^T \in F'$, correspondente ao rótulo associado a um item de dado $x_i \in \mathcal{X}$, é dada por:

$$y_i = \arg \max_{j \in \mathcal{L}} F_{ij} . \quad (2.41)$$

- $\mathbf{Y} \in F'$ representa uma matriz $n \times c$, onde $Y_{ij} = 1$ se x_i apresenta rótulo $y_i = j$, ou $Y_{ij} = 0$ caso contrário.
- $\mathbf{W}$ denota uma matriz de afinidade gerada a partir de uma rede completamente conectada $\mathcal{G}$, tal como:

$$W_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), & \text{se } i \neq j, \\ 0, & \text{caso contrário,} \end{cases} \quad (2.42)$$

onde σ é o parâmetro da largura de banda do *kernel*.

O trabalho apresentado em [Zhu, Ghahramani e Lafferty \(2003\)](#) utiliza uma função harmônica f de maneira a permitir que o rótulo atribuído a um item de dado não rotulado consista na média ponderada de seus vizinhos. Tal função permite que o processo de classificação aconteça de maneira suave, i.e., $f(x) \in [0, 1]$. Os principais passos dessa técnica, denominada campos Gaussianos e funções harmônicas (GFHF), são apresentados a seguir:

1. obtenha a matriz de afinidade $\mathbf{W}$ usando [2.42](#);
2. calcule o Laplaciano combinatório Δ a partir de $\mathbf{W}$:

$$\Delta = \mathbf{D} - \mathbf{W}, D_{ij} = \begin{cases} \sum_j W_{ij}, & \text{se } i = j, \\ 0, & \text{caso contrário;} \end{cases} \quad (2.43)$$

3. retorne a matriz de saída $\mathbf{F}$ resolvendo a seguinte formulação:

$$\mathbf{F}_{\mathcal{U}} = -\Delta_{\mathcal{U}\mathcal{U}}^{-1} \Delta_{\mathcal{U}\mathcal{L}} \mathbf{Y}_{\mathcal{L}} . \quad (2.44)$$

Assumindo-se que os vértices estão ordenados de modo que os rotulados são listados primeiro, tal formulação é fruto do particionamento do Laplaciano combinatório Δ nas quatro sub-matrizes a seguir:

$$\begin{bmatrix} \Delta_{\mathcal{L}\mathcal{L}} & \Delta_{\mathcal{L}\mathcal{U}} \\ \Delta_{\mathcal{U}\mathcal{L}} & \Delta_{\mathcal{U}\mathcal{U}} \end{bmatrix} . \quad (2.45)$$

4. Usando $\mathbf{F}$, associe o rótulo correspondente para cada vértice não rotulado $x_i, i = l+1, \dots, n$ através da Eq. [2.41](#).

Ao contrário do algoritmo GFHF, a técnica de propagação de rótulos proposta em [Zhou et al. \(2004\)](#), denominada consistência local e global (LGC), permite a alteração de rótulos

previamente fornecidos, por considerar também as informações disponíveis a partir dos vizinhos durante o processo de propagação. Tal característica é, particularmente, interessante quando tais dados podem ter sido incorretamente rotulados ou sejam inconsistentes em relação aos seus vizinhos. Desse modo, um parâmetro α é responsável por definir a quantidade relativa de informação a ser obtida tanto a partir dos vizinhos quanto dos rótulos inicialmente fornecidos. A técnica pode ser descrita por meio das seguintes etapas:

1. Obtenha-se a matriz de afinidade $\mathbf{W}$ usando 2.42;
2. Calcule-se o Laplaciano normalizado $\mathbf{L}$ a partir de $\mathbf{W}$:

$$\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}} ; \quad (2.46)$$

3. Itere-se a seguinte formulação até sua convergência, onde $\alpha \in [0, 1]$:

$$\mathbf{F}(t+1) = \alpha \mathbf{L} \mathbf{F}(t) + (1 - \alpha) \mathbf{Y} , \quad (2.47)$$

4. Seja $\mathbf{F}^*$ o limite da sequência $\{\mathbf{F}(t)\}$, cada vértice $x_i \in \mathcal{X}$ é associado a um rótulo usando Eq. 2.41.

Os autores em [Zhou et al. \(2004\)](#) também encontraram que o problema de propagação de rótulos pode ser resolvido mediante um *framework* de regularização, cuja solução para o problema é dada por:

$$F^* = \arg \min_{F \in F'} Q(F) , \quad (2.48)$$

com Q sendo uma função de energia minimizada por F^* , cuja definição é feita de forma a satisfazer dois critérios. O primeiro critério, denominado critério de suavidade, o qual denota que uma função de classificação adequada não deve variar muito os rótulos entre itens de dados próximos. O segundo critério, denominado critério do erro ou ajuste, afirma que uma função de classificação adequada deve aproximar os rótulos propagados daqueles previamente conhecidos. Logo, $Q(F)$ é dada por:

$$Q(F) = \frac{1}{2} \left(\sum_{i,j=1}^n w_{ij} \left\| \frac{1}{\sqrt{D_{ii}}} F_i - \frac{1}{\sqrt{D_{jj}}} F_j \right\|^2 + \mu \sum_{i=1}^n \|F_i - Y_i\|^2 \right) , \quad (2.49)$$

onde $\mu > 0$ é o parâmetro de regularização.

2.4 Anotação automática de papéis semânticos

Esta seção introduz a anotação automática de papéis semânticos. A tarefa consiste em detectar as relações lógicas existentes entre um evento e seus participantes, ou de maneira formal, entre um predicador e seus argumentos. Com contribuições significativas em muitas

tarefas do processamento de línguas naturais (PLN), a anotação de papéis semânticos (APS) permite a análise do significado das sentenças, característica extremamente relevante para que muitos sistemas computacionais possam manipular, efetivamente, a informação codificada no texto. Alguns exemplos de aplicações de APS incluem o uso de modelos de domínio específico para extrair informações sobre notícias financeiras (FILIPPOVA *et al.*, 2009), a construção de sistemas de pergunta-resposta relacionados a atividades médicas (NIELSEN *et al.*, 2010), o desenvolvimento de sistemas para detecção de plágio (OSMAN *et al.*, 2012) e também o emprego de APS em sistemas de tradução automática de máquina (WU; FUNG, 2009).

As subseções a seguir visam abordar o assunto e são organizadas da seguinte maneira: a Subseção 2.4.1 apresenta uma visão geral sobre APS, descrevendo conceitos e exemplos; a Subseção 2.4.2 aborda os principais recursos linguísticos para APS, especificamente, o *corpus* de língua Portuguesa PropBank-br, o qual é usado em alguns experimentos computacionais apresentados na tese; e a Subseção 2.4.3 destaca a tarefa de classificação de papéis semânticos, bem como os principais trabalhos de APS existentes na literatura, especificamente, aqueles voltados para a língua Portuguesa.

2.4.1 Papéis semânticos

No contexto da teoria linguística, Fillmore (1968) tentou estabelecer propriedades universais de todas as línguas por meio de um conjunto de conceitos universais, presumidamente naturais, que poderiam identificar certos tipos de julgamentos que os seres humanos são capazes de fazer sobre os eventos que acontecem a sua volta. Para Fillmore (1968), há muitas relações sintáticas relevantes semanticamente envolvendo os substantivos, porém estão, em grande parte, escondidas. O modelo de casos proposto em Fillmore (1968) corresponde à grande parte da gramática de casos contemporânea (DOWTY, 1991) e se baseia na relação particular existente entre um argumento e o predador, por meio da qual cada associação ocorre somente uma vez em uma sentença simples.

Em sua estrutura básica, as sentenças consistem de um predador (especificamente, verbal neste trabalho) e de argumentos (sujeito, objetos) (PUNYAKANOK; ROTH; YIH, 2005). Assim, papéis semânticos representam as relações lógicas entre um grupo de predadores e um conjunto de argumentos (DOWTY, 1991). Além disso, essas relações costumam ser definidas em *frames*, uma vez que isso evita alguns problemas, por exemplo, papéis temáticos abstratos (FILLMORE, 1968; JACKENDOFF, 1972). A definição de *frames*, proposta em Fillmore (1976), consiste em criar um cenário contendo papéis que possam ser facilmente identificados em um evento. Pensando em um domínio semântico ligado a eventos comerciais, o *frame* para o predador *comprar* deverá conter argumentos relacionados a comprador, dinheiro, mercadoria, vendedor, entre outros, em que cada elemento do *frame* desempenha um papel denominado papel temático (FRAWLEY, 1992), caso semântico (FILLMORE, 1968) ou papel semântico (DILLON, 1977).

O principal objetivo da anotação de papéis semânticos é indicar qual relação semântica existe entre um predicador e os participantes e propriedades associados a ele, com estas relações advindas de uma lista pré-determinada de possíveis papéis semânticos para o predicador em questão (MàRQUEZ *et al.*, 2008). Assim, os constituintes que recebem papéis em uma sentença precisam ser identificados (identificação) e seus rótulos de papel semântico, corretamente atribuídos (classificação). Alguns rótulos bastante utilizados em APS são, por exemplo, AGENTE e PACIENTE para entidades que participam de um evento, e TEMPO e MANEIRA para a caracterização de aspectos das relações de eventos ou participantes.

Como exemplo, na Sentença 1, é possível perceber dois argumentos para o predicador *derrubar*: *chuva* e *árvore*. Em relação aos papéis semânticos, a sentença possui *chuva* como CAUSA, pois inicia a ação expressa pelo predicador sem ter o controle dela e *árvore* como PACIENTE. Uma última análise de papéis semânticos é realizada para o predicador comprar na Sentença 2, que tem *Roberto* como AGENTE, *brinquedo* como TEMA, pois é afetado indiretamente pela ação do predicador (sem mudança de estado), *loja* como ORIGEM, pois refere-se à entidade a partir da qual alguma coisa é movida e *R\$10,00* como VALOR², pois representa um valor utilizado para realização do predicador.

1. A chuva derrubou a árvore.
2. Roberto comprou um brinquedo da loja por R\$10,00.

2.4.2 PropBank-br e outros recursos

O interesse de pesquisadores por papéis semânticos é tão grande que muitos deles concentraram seus esforços durante muitos anos na construção manual de grandes recursos de anotações semânticas, tais como a VerbNet (KIPPER; DANG; PALMER, 2000), a FrameNet (FILLMORE; JOHNSON; PETRUCK, 2003) e o PropBank (PALMER; GILDEA; KINGSBURY, 2005). Como consequência, uma variedade de modelos tem sido proposta para APS mediante o uso destes recursos, que também possuem corpus anotados. Especificamente, o PropBank foi o primeiro recurso cujas decisões no estilo de anotação objetivaram facilitar a aplicação de técnicas de aprendizado de máquina em APS (PALMER; GILDEA; XUE, 2010). Dessa forma, muitos corpus desenvolvidos para outras línguas, como o PropBank Chinês (XUE, 2008), o PropBank Espanhol (TAULÉ; RECASENS, 2008) e o PropBank Português Brasileiro (DURAN; ALUÍSIO, 2012), têm sido fortemente baseados neste modelo.

O PropBank é um recurso linguístico-computacional que adicionou uma camada de rótulos de papéis semânticos às estruturas sintáticas de um subcorpus relacionado a notícias financeiras do *Penn Treebank*³. Os argumentos que pertencem às proposições são anotados pelo PropBank com rótulos de papéis numerados (Arg0 à Arg5), e modificadores são anotados

² Não há um consenso na literatura sobre o nome do papel semântico para dinheiro. Neste projeto, assim como em Rosa e Adan-Coello (2010), foi utilizado VALOR.

³ <http://www.cis.upenn.edu/~treebank/>

com rótulos de papéis específicos ArgM (Modificadores de Argumento). Além disso, um léxico de verbos e seus sentidos está disponível para auxiliar na identificação dos significados dos papéis semânticos. Por isso, para cada verbo no corpus, há um *frame* associado que determina o conjunto de papéis semânticos para aquele verbo. Assim, é possível “traduzir” os argumentos numerados em uma dada sentença de acordo com as descrições de papéis para aquele verbo específico. Por exemplo, quando um argumento do verbo “comprar” é rotulado como Arg0, ele é descrito como “comprador” (AGENTE), tal como *Roberto* na Sentença 2. Em relação aos argumentos modificadores, eles possuem características mais gerais e podem ser aplicados para quaisquer verbos. A Tabela 2 apresenta os principais argumentos modificadores do PropBank, bem como o significado de cada um deles.

Tabela 2 – Argumentos modificadores (ArgM) do PropBank.

Classe	Significado
ADV	Adverbial
CAU	Causa
DIR	Direção
DIS	Marcador discursivo
EXT	Extensão
LOC	Lugar
MNR	Maneira
NEG	Negação
PNC	Propósito
PRD	Predicação secundária
REC	Recíproco
TMP	Temporal

O PropBank-br é um recurso lexical para o português brasileiro, baseado no estilo de anotação do PropBank. Ele foi criado baseado nas anotações da seção da língua portuguesa do Brasil no CETENFolha do corpus Bosque da Floresta Sintá(c)tica⁴, que é um corpus automaticamente anotado pelo analisador sintático Palavras (BICK, 2000) e manualmente corrigido por linguistas. O corpus é composto de 4213 sentenças e 1023 verbos. O total de argumentos por classe é dado pela Tabela 3, o qual evidencia um cenário de distribuições bastante desbalanceadas.

A Figura 22 apresenta a árvore sintática de uma sentença pertencente ao PropBank-br. Na figura, para cada palavra da sentença (expressa na coluna “Forma”) é associado um número identificador (“ID”), o seu lema (“Lema”), a sua classificação morfosintática (“POS-tag”) e também sintática (“Árvore sintática”), além dos respectivos papéis semânticos atribuídos aos participantes da sentença na coluna “Papéis semânticos” e a identificação dos verbos expressa na coluna “Predicador”. Ainda na Figura 22, as siglas “NP”, “VP” e “PP” exibidas na árvore referem-se, respectivamente, a sintagmas nominais, verbais e preposicionais⁵.

⁴ <http://www.linguateca.pt/Floresta/principal.html>

⁵ <http://www.linguateca.pt/Floresta/BibliaFlorestal/>

Tabela 3 – Classes do PropBank-br em função do número de argumentos.

Classes	Argumentos
A0	3,058
A1	5,148
A2	1,101
A3	113
A4	75
A5	1
AM-ADV	369
AM-CAU	156
AM-DIR	15
AM-DIS	294
AM-EXT	81
AM-LOC	778
AM-MNR	410
AM-NEG	335
AM-PNC	171
AM-PRD	192
AM-REC	65
AM-TMP	1,142
	13,504

ID	Forma	Lema	POS-tag	Árvore sintática	Predicador	Papéis Semânticos
1	Ele	ele	PRON-PERS	(FCL (NP*))	-	(A0*)
2	receberá	receber	V-FIN	(VP*)	receber	(V*)
3	o	o	ART	(NP*)	-	(A1*
4	valor	valor	N	*	-	*
5	a	a	PRP	(PP*	-	*
6	a	o	ART	(NP*	-	*
7	vista	vista	N	*)))	-	*)
8	após	após	PRP	(PP*	-	(AM-TMP*
9	30	30	NUM	(NP*	-	*
10	dias	dia	N	*))	-	*)
11	.	-	PU	*)	-	*

Figura 22 – Representação e árvore sintática de uma sentença do PropBank-br. Retirado e adaptado de [Alva-Manchego \(2013\)](#).

A atribuição dos papéis semânticos para a Fig. 22 é exibida na Sentença 3, onde A0 determina o papel semântico de AGENTE, A1 TEMA e AM-TMP, um argumento modificador de TEMPO. Uma simples consulta ao sentido do verbo “receber” no repositório do PropBank⁶ é capaz de traduzir o significado semântico relacionado à A0 e A1 na sentença, os quais denotam o recebedor e a coisa recebida, respectivamente.

3. [*Ele*_{A0}] receberá [*o valor à vista*_{A1}] [*após 30 dias*_{AM-TMP}] .

Nos experimentos apresentados nesta tese, o pré-processamento realizado sobre o PropBank-br, bem como os atributos extraídos das sentenças seguiram o estudo desenvolvido em Alva-Manchego (2013). No pré-processamento, 3308 das sentenças anotadas no corpus são utilizadas e as demais são descartadas devido a erros na árvore sintática, ocorrências do verbo ser, argumentos embutidos, entre outras razões detalhadas em Alva-Manchego (2013).

2.4.3 Anotação de papéis semânticos

De acordo com Palmer, Gildea e Xue (2010), a anotação de papéis semânticos pode ser tratada no *framework* geral de uma tarefa de classificação. Assim, para um dado predicador, e dada a análise sintática de cada constituinte, a tarefa é selecionar, a partir de um conjunto pré-definido, o papel semântico dos constituintes em relação ao predicador.

Apesar do grande número de constituintes em uma árvore sintática, apenas uma pequena fração deles são argumentos para algum predicador na árvore. Dessa forma, é comum a tarefa de APS ser dividida nas duas etapas seguintes:

- **Identificação de argumentos:** onde um processo de decisão binário aponta se um dado constituinte recebe ou não papel semântico de um dado predicador. Em caso afirmativo, o constituinte é marcado como argumento. Em caso negativo, o constituinte é desconsiderado para a etapa seguinte. Considerando o exemplo da Fig. 22, a etapa de identificação de argumentos retornaria os seguintes argumentos:
4. [*Ele*_{Arg}] receberá [*o valor à vista*_{Arg}] [*após 30 dias*_{Arg}] .
- **Classificação de argumentos:** em que se atribuem os papéis semânticos para os argumentos identificados na etapa anterior, conforme apresentado na Sentença 3.

Ainda em relação à identificação de argumentos, uma abordagem comum na área é o uso de conjuntos de heurísticas para realizar a **poda de constituintes** que não sejam argumentos como uma pré-etapa do processo de identificação. De acordo com Palmer, Gildea e Xue (2010), o uso de bons algoritmos heurísticos, como o proposto em Xue e Palmer (2004), leva a uma melhoria geral na precisão da anotação de papéis semânticos. Além disso, tanto as etapas de

⁶ <https://verbs.colorado.edu/propbank/framesets-english/receive-v.html>

identificação quanto de classificação são caracterizadas pela extração de diferentes conjuntos de atributos de modo a obter o melhor desempenho em cada etapa. As medidas de desempenho usadas na literatura normalmente são precisão, cobertura, F1 e acurácia.

Um exemplo prático do grande potencial de APS é descrito em [Osman *et al.* \(2012\)](#) em que é apresentado um sistema baseado em papéis semânticos para a detecção de plágio. Dada uma sentença original e outra suspeita, a ideia proposta pelos autores é realizar a anotação de papéis semânticos das duas sentenças de modo que apenas os argumentos delas sejam usados no processo de detecção. Assim, cada argumento é designado para sua classe de papel semântico correspondente e a detecção de plágio ocorre por calcular a similaridade entre as sentenças de acordo com os argumentos de mesma classe. Uma característica importante provida pelos papéis semânticos nessa aplicação é que a captura de tais papéis é independente da mudança de posição dos argumentos na sentença. Além disso, eles fornecem também um mecanismo robusto para lidar com a mudança de voz passiva-ativa nas sentenças e a substituição de palavras por sinônimos.

2.4.3.1 Revisão dos principais trabalhos da literatura

As pesquisas em anotação automática de papéis semânticos mais desenvolvidas são para a língua inglesa. Várias razões podem ser enumeradas para justificar isto. Uma das principais diz respeito à grande infraestrutura em termos de recursos disponíveis para o inglês, através de grandes projetos tais como VerbNet, FrameNet e PropBank. Outra razão poderia ser o fato de que os recursos para a língua inglesa já existem há mais de uma década, enquanto que, para outras línguas, as primeiras versões dos recursos foram liberadas recentemente.

Pela forma de lidar com a APS, a abordagem apresentada em [Gildea e Jurafsky \(2002\)](#) é tomada como referência na área ([JOHANSSON, 2008](#)). A ideia geral de [Gildea e Jurafsky \(2002\)](#) foi modelar o problema como uma tarefa de classificação no qual um modelo de classificação estatístico é criado por meio de dados anotados manualmente em cópys, e, posteriormente, usado para realizar a anotação do papel semântico de predicados, em uma nova sentença sem anotação. Vários trabalhos têm sido desenvolvidos baseado nestes conceitos ([HACIOGLU, 2004](#); [JOHANSSON, 2008](#)) e alguns resultados em aplicações são considerados promissores ([BETHARD *et al.*, 2008](#); [SANTOS; MILIDIÚ, 2012](#); [OSMAN *et al.*, 2012](#)).

Outro trabalho importante na área é apresentado em [Pradhan *et al.* \(2005\)](#) e, posteriormente, refinado em [Pradhan, Ward e Martin \(2008\)](#). Basicamente, esta abordagem apresenta duas modificações em relação ao método desenvolvido em [Gildea e Jurafsky \(2002\)](#): troca do algoritmo de classificação estatístico por um algoritmo baseado em máquina de vetores de suporte (SVM) e adição de novos atributos à abordagem. Como resultado, tem-se que o desempenho da abordagem proposta é superior a de outros modelos da literatura.

Em relação à língua portuguesa, a anotação automática de papéis semânticos ainda é pouco explorada. Os trabalhos existentes costumavam operar sobre domínios muito específicos

devido a limitações metodológicas e de recursos disponíveis (BICK, 2007; AMANCIO; DURAN; ALUÍSIO, 2011; SEQUEIRA; GONÇALVES; QUARESMA, 2012). Entretanto a recente construção do PropBank-br mudou este cenário e novas propostas de APS têm surgido.

A primeira técnica de anotação de papéis semânticos utilizando o PropBank-br foi proposta em Alva-Manchego e Rosa (2012). O trabalho é baseado em técnicas estado-da-arte na anotação de papéis semânticos para a língua inglesa, tais como os algoritmos de regressão logística e SVM (PRADHAN; WARD; MARTIN, 2008; BAZRAFESHAN; GILDEA, 2013). Na técnica, o autor emprega o algoritmo de regressão logística devido ao menor custo computacional e desempenho semelhante quando comparado à SVM. Segundo o autor, os resultados obtidos para identificação de argumentos estão próximos aos de sistemas estado-da-arte para o inglês, contudo os resultados obtidos para a classificação de argumentos ainda estão distantes. Outra contribuição importante do trabalho é o estudo e aplicação de um conjunto de atributos do estado-da-arte inglesa para a língua portuguesa. Nos experimentos de APS apresentados nesta tese, os atributos selecionados são derivados de tal investigação.

Outro sistema de anotação automática de papéis semânticos para a língua portuguesa, utilizando o PropBank-br, é proposto em Fonseca e Rosa (2013), Fonseca (2013). O trabalho é baseado em Collobert *et al.* (2011) (língua inglesa) e propõe uma rede neural artificial (RNA) para realizar as seguintes tarefas: etiquetagem morfossintática, análise sintática superficial, reconhecimento de entidades nomeadas e anotação de papéis semânticos, sendo este último o objetivo principal da proposta. Como resultado do trabalho, o autor enfatiza a independência do modelo em relação às informações produzidas por ferramentas externas, isto é, independência de análise sintática, classificação morfossintática, etc. Contudo o autor afirma que o desempenho alcançado pelo sistema ainda está muito abaixo do estado da arte para o inglês e também para o português, quando comparado ao trabalho de Alva-Manchego (2013).

Mais recente, o trabalho apresentado em Hartmann (2015) analisa a tarefa de APS para o gênero de textos de opinião sobre produtos da Web usando os dois sistemas de APS existentes, além de avaliar o reuso do aprendizado intergêneros, i.e., usar o classificador treinado sobre um gênero para prever argumentos de outro gênero. Como resultados, os autores encontraram que o sistema usando regressão logística obtém melhor desempenho que a RNA mesmo em bases de dados maiores e mais balanceadas que o PropBank-br. O trabalho também mostrou que o uso de árvores sintáticas não revisadas para treinamento do sistema APS permite um melhor desempenho do que quando o sistema é treinado sobre árvores sintáticas revisadas.

Em relação ao desempenho das técnicas de APS avaliadas sobre o PropBank-br, o classificador de regressão logística (LR) usado em Alva-Manchego (2013) apresenta os melhores resultados, sendo inclusive utilizado em trabalhos seguintes (HARTMANN, 2015). Contudo, ao comparar o desempenho obtido sobre o PropBank-br em relação àqueles obtidos sobre o PropBank, é possível concluir que APS, para língua portuguesa, está bastante distante dos principais sistemas para a língua inglesa. Por exemplo, Alva-Manchego (2013) alcança acurácia

de 81.70% sobre o PropBank-br enquanto [Pradhan, Ward e Martin \(2008\)](#) obtêm acurácia de 93.00% sobre o PropBank. Apesar de distante, o resultado obtido para a língua portuguesa pode ser considerado bom ao analisar o tamanho reduzido do PropBank-br ante o PropBank. Por outro lado, esse número reduzido de instâncias no PropBank-br e o desbalanceamento do córpus abrem espaço para que novas técnicas sejam investigadas, especialmente, no que tange ao aprendizado semissupervisionado, uma vez que rotular os papéis semânticos dos argumentos é caro, demorado e exige muito esforço por parte de anotadores humanos especializados.

Mesmo com a dificuldade em prover sentenças rotuladas para APS, a maior parte das técnicas desenvolvidas contemplam o aprendizado supervisionado, em que as abordagens mais comuns estão relacionadas ao uso de classificadores estatísticos ([GILDEA; JURAFSKY, 2002](#); [SURDEANU *et al.*, 2003](#); [PRADHAN; WARD; MARTIN, 2008](#); [KALJAH, 2010](#); [ROSA; ADAN-COELLO, 2010](#); [COLLOBERT *et al.*, 2011](#)). Inclusive no PropBank-br, no qual os dados anotados são escassos, o contexto supervisionado é o mais explorado ([ALVA-MANCHEGO; ROSA, 2012](#); [FONSECA; ROSA, 2013](#); [CARNEIRO *et al.*, 2016](#)). Por outro lado, a literatura de APS contém poucos trabalhos que contemplam os aprendizados não supervisionados e semissupervisionados. Trabalhos não supervisionados como os apresentados em [Lang e Lapata \(2011a\)](#), [Lang e Lapata \(2011b\)](#) referem-se à tarefa como indução de papéis semânticos e tem por objetivo agrupar os argumentos de cada verbo de acordo com suas funções semânticas. Por sua vez, os trabalhos semissupervisionados incluem, por exemplo, a investigação de métodos usando autotreinamento e cotreinamento ([HE; GILDEA, 2006](#)), e o emprego de um pequeno número de argumentos rotulados para construir uma distribuição a priori a ser usada por um método não supervisionado ([TITOV; KLEMENTIEV, 2012](#)). Além destes, outras abordagens semissupervisionadas visam aumentar os dados anotados com dados não anotados, cujos papéis semânticos são inferidos através de projeção ([FÜRSTENAU; LAPATA, 2012](#)), bem como reduzir a esparsidade dos atributos usando representação de palavras ([DESCHACHT; MOENS, 2009](#)).

Dada a relevância da tarefa de APS no contexto semissupervisionado, o pequeno número de trabalhos sobre o assunto e a escassez de dados anotados no PropBank-br, o Capítulo 6 da tese contempla a investigação técnicas de aprendizado baseado em grafos para a difusão de papéis semânticos em sentenças extraídas do PropBank-br.

2.5 Considerações finais

Este trabalho tem como principal objetivo o desenvolvimento de técnicas e heurísticas baseadas em redes complexas para a classificação de dados. Nesse sentido, a Seção 2.1 forneceu uma revisão importante sobre conceitos básicos de grafos, além de discutir medidas e modelos de redes complexas. Por exemplo, as medidas de rede apresentadas na Subseção 2.1.2 são empregadas na classificação por conformidade de padrão e por importância apresentadas, respectivamente, nos Capítulos 3 e 4. Dentre os modelos de redes discutidos na Subseção 2.1.3,

a otimização estrutural de redes é o principal tópico investigado no Capítulo 5.

Na Seção 2.2, o aprendizado de máquina e seus principais paradigmas são descritos para, na seção seguinte, a Seção 2.3, serem discutidos os principais trabalhos baseados em redes relacionados ao contexto do trabalho, tais como métodos de formação da rede, técnicas de classificação supervisionada e semissupervisionadas. Os conceitos e técnicas apresentados nestas seções podem ser vistos ao longo dos capítulos de desenvolvimento da tese.

Por fim, a Seção 2.4 introduz um problema real bastante relevante para o processamento de línguas naturais, a anotação de papéis semânticos. Esse problema é usado para fornecer uma avaliação adicional à técnica de otimização estrutural no Capítulo 5 e a difusão de papéis semânticos baseado em grafos é a investigação principal do Capítulo 6.

CLASSIFICAÇÃO DE ALTO NÍVEL USANDO GRAFO K-ASSOCIADOS ÓTIMO

Neste capítulo, a combinação convexa entre classificadores de alto e baixo nível é investigada mediante uma técnica híbrida em que uma das características salientes é a ausência de parâmetros para ambos classificadores. Em poucas palavras, as associações produzidas pelas técnicas de alto e baixo nível são extraídas a partir do mesmo grafo k -associados ótimo (k AOG). Ademais, o classificador ótimo de Bayes, apresentado em [Bertini et al. \(2011\)](#), é tomado como classificador de baixo nível enquanto um conjunto de medidas de redes complexas, como proposto em [Silva e Zhao \(2012a\)](#), é tomado como classificador de alto nível. Como mostram os experimentos, a técnica resultante é caracterizada pela simplicidade e pela habilidade em detectar a formação de padrão nos dados.

3.1 Descrição da técnica

Seja $\mathcal{X}_{train} = \{x_i, i = 1, \dots, n\}$ um conjunto de dados rotulados onde para cada item x_i é associada uma única classe $l_i \in \mathcal{L}$, e seja $\mathcal{X}_{test} = \{x_i, i = 1, \dots, m\}$ um conjunto de dados cujos rótulos deseja-se prever. Dado que cada item de dado é descrito por d atributos $x_i = \{x_{i1}, \dots, x_{id}\}$ e $\mathcal{X}_{train} \cap \mathcal{X}_{test} = \emptyset$, a técnica proposta pode ser dividida em duas etapas como apresentado a seguir.

Treinamento: A etapa de treinamento da técnica proposta consiste na formação da rede a partir dos dados de treinamento $\mathcal{X}_{train}$ usando o método k AOG tal como apresentado na Subseção [2.3.2.1](#). A rede formada é denotada por $\mathcal{G} = \{V, E\}$, onde V é o conjunto de nós e E o conjunto de arestas. Cada nó $v_i \in V$ denota um item de dado rotulado $x_i \in \mathcal{X}_{train}$ e $e_{ij} \in E$ representa a existência ou não de conexão do vértice v_i para o vértice v_j . A informação necessária para prover as associações de baixo e alto nível na etapa seguinte é extraída do grafo $\mathcal{G}$.

Classificação: Nesta etapa, a classificação final é produzida pela combinação linear convexa das associações obtidas pelos dois classificadores, quando se avalia uma dada instância de teste $y \in \mathcal{X}_{test}$ para um componente α , i.e.

$$\mathcal{M}_y^{(\alpha)} = (1 - \lambda)\mathcal{C}_y^{(\alpha)} + \lambda\mathcal{H}_y^{(\alpha)}, \quad (3.1)$$

onde $\mathcal{C}$ representa o classificador ótimo de Bayes, usado como classificador de baixo nível, e $\mathcal{H}$ representa um conjunto de medidas de redes complexas usadas para prover informação de alto nível. Especificamente, o classificador de baixo nível emprega os valores de *pureza* Φ_α e os vizinhos k_α de cada componente α para calcular a probabilidade de associação de um dado item de teste y em relação aos componentes da rede, tal como descreve a Eq. (2.30); e o classificador de alto nível compreende as associações produzidas pelas medidas de redes complexas selecionadas, as quais são calculadas antes e após a inserção de y nos componentes da rede, de modo que os componentes que apresentam menores variações dessas medidas recebem maior probabilidade de classificação, significando que y possui maior conformidade com o padrão formado por eles. Como apresentado na Subseção 2.3.2.2, λ define as contribuições dos classificadores de baixo e alto nível para a classificação final. No fim, é atribuído a y o rótulo do componente em que ele possui maior valor de associação, i.e.

$$\varphi_y = l \left(\arg \max_{\alpha \in \mathcal{G}} \mathcal{M}_y^{(\alpha)} \right) \in \mathcal{L}. \quad (3.2)$$

A seguir, são detalhadas outras contribuições do trabalho as quais permitem a integração entre o grafo k -associados ótimo e a classificação de alto nível. Algumas dessas contribuições visam, especificamente, tornar o modelo de classificação de alto nível mais simples em termos de parâmetros, como, por exemplo, o uso da rede k AOG e o *coeficiente de influência das medidas de rede*, enquanto outras também representam propostas inovadoras para o aprendizado supervisionado baseado em grafos, as quais podem ser exploradas ou adaptadas em vários outros contextos ou na construção de novas técnicas, tais como a medida de eficiência do componente e a classificação por componentes em vez de classes. Como o classificador de baixo nível usado neste estudo é exatamente aquele descrito na Subseção 2.3.2.1, as próximas subseções estão relacionadas ao classificador de alto nível.

3.1.1 Classificação de alto nível sem parâmetros

A ideia geral por trás da técnica de classificação de alto nível proposta aqui, denotada por $\mathcal{H}_y^{(\alpha)}$, é simples: (i) a rede k AOG provê um conjunto de componentes baseado na otimização da medida de *pureza*; (ii) a técnica de alto nível examina as variações das medidas de rede nestes componentes, considerando a inserção de um item de teste y ; e (iii) as probabilidades associadas a cada componente $\alpha \in \mathcal{G}$ são obtidas. Especificamente, o termo de alto nível representa a conformidade de padrão da nova instância y em relação a cada componente da rede, e é dada por:

$$\mathcal{H}_y^{(\alpha)} = \frac{\sum_{u=1}^Z \delta_y(u) [1 - f_y^{(\alpha)}(u)]}{\sum_{g \in \mathcal{G}} \sum_{u=1}^Z \delta_y(u) [1 - f_y^{(g)}(u)]}, \quad (3.3)$$

onde u está relacionado às medidas de redes complexas utilizadas pela técnica de alto nível, $\delta_y(u) \in [0, 1]$, $\forall u \in \{1, \dots, Z\}$ indica a influência de cada medida de rede no processo de classificação, e $f_y^{(\alpha)}(u)$ provê uma resposta sobre a conformidade de y em relação ao padrão formado pelo componente α , considerando a u -ésima medida de rede. Na equação, o denominador é apenas para normalização.

Em relação ao termo δ , denominado *coeficiente de influência das medidas de rede*, ele é definido de maneira automática na abordagem proposta da seguinte maneira:

$$\delta_y(u) = \frac{1 - \left(\max_{\alpha \in \mathcal{G}} \Delta G_y^{(\alpha)}(u) - \min_{\alpha \in \mathcal{G}} \Delta G_y^{(\alpha)}(u) \right)}{\sum_{u=1}^Z 1 - \left(\max_{\alpha \in \mathcal{G}} \Delta G_y^{(\alpha)}(u) - \min_{\alpha \in \mathcal{G}} \Delta G_y^{(\alpha)}(u) \right)}, \quad (3.4)$$

onde $\Delta G_y^{(\alpha)}(u) \in (0, 1]$ representa a variação capturada por uma medida de rede, quando um item de teste y é inserido em um componente α . Na equação, o coeficiente $\delta_y(u)$ é obtido a partir da diferença entre a maior e a menor variação de cada medida de rede u subtraída por 1. Dessa forma, o objetivo dessa formulação é estabelecer um balanceamento entre todas as medidas de rede empregadas pelo classificador de alto nível, de modo que seja evitado que apenas uma medida de rede domine a decisão de classificação. Note-se que a equação apresentada é válida apenas se $\sum_{u=1}^Z \delta_y(u) = 1$.

Ainda na Eq. (3.3), o termo f é definido de maneira a trabalhar sobre componentes ao invés de classes. Assim, tem-se:

$$f_y^{(\alpha)}(u) = \frac{\Delta G_y^{(\alpha)}(u) p^{(\alpha)}}{\sum_{\alpha} \Delta G_y^{(\alpha)}(u) p^{(\alpha)}}, \quad (3.5)$$

onde $p^{(\alpha)} \in [0, 1]$ é a proporção de instâncias de treinamento que pertencem ao componente α e o denominador da equação é apenas para fins de normalização.

3.1.2 Variação de medidas de redes complexas

A seguir, são apresentadas as medidas de redes complexas usadas para detectar características topológicas dos componentes da rede.

Assortatividade - $\Delta G_y^{(\alpha)}(1)$: a medida de assortatividade é descrita na Subseção 2.1.2.2. A variação de assortatividade causada por um item de teste y em relação a um componente α é dada por:

$$\Delta G_y^{(\alpha)}(1) = \frac{|r'^{(\alpha)} - r^{(\alpha)}|}{\sum_{u \in U} |r'^{(u)} - r^{(u)}|}, \quad (3.6)$$

onde $r^{(\alpha)}$ e $r'^{(\alpha)}$ denotam o valor de assortatividade calculado, respectivamente, antes e após a inserção de y no componente α .

Coeficiente de agrupamento médio - $\Delta G_y^{(\alpha)}(2)$: o coeficiente de agrupamento médio é apresentado na Subseção 2.1.2.3 e sua variação concernente à inserção de um item de teste y em um componente α é calculada por:

$$\Delta G_y^{(\alpha)}(2) = \frac{|CC'^{(\alpha)} - CC^{(\alpha)}|}{\sum_{u \in U} |CC'^{(u)} - CC^{(u)}|}, \quad (3.7)$$

onde $CC^{(\alpha)}$ e $CC'^{(\alpha)}$ denotam o coeficiente de agrupamento médio obtido, respectivamente, antes e após a inserção de y no componente α .

Grau médio - $\Delta G_y^{(\alpha)}(3)$: a medida de grau médio é discutida na Subseção 2.1.2.1 e sua variação em relação à inserção de um item de teste y em um componente α é formalmente definida por:

$$\Delta G_y^{(\alpha)}(3) = \frac{|\langle k'^{(\alpha)} \rangle - \langle k^{(\alpha)} \rangle|}{\sum_{u \in U} |\langle k'^{(u)} \rangle - \langle k^{(u)} \rangle|}, \quad (3.8)$$

onde $\langle k \rangle^{(\alpha)}$ e $\langle k \rangle'^{(\alpha)}$ representam o grau médio obtido, respectivamente, antes e após a inserção de y no componente α .

3.1.3 Medida de eficiência do componente

Na classificação de alto nível proposta em Silva e Zhao (2012a), a inserção de um item de teste nos componentes da rede é dirigida pelo critério ϵN mediante um novo valor de ϵ . Na classificação de alto nível proposta neste capítulo, este parâmetro não é necessário, uma vez que uma nova medida, denominada eficiência do componente, é responsável por realizar a inserção de novos itens de dados a partir da eficiência na troca de informação entre os vértices de cada componente. Especificamente, algumas adaptações conceituais realizadas, a fim de permitir a combinação entre o grafo k -associados ótimo e a classificação de alto nível baseada em medidas de redes complexas, são apresentadas a seguir:

- uma vez que a rede $kAOG$ é constituída por componentes otimizados de acordo com a medida de *pureza*, a variação das medidas de redes, aqui, é calculada sobre componentes ao invés de classes;
- ao contrário do classificador ótimo de Bayes, que calcula probabilidades apenas para componentes cujos vértices são vizinhos do item de teste, a classificação de alto nível calcula a variação das medidas de redes naqueles componentes nos quais a inserção do item de teste aumenta sua eficiência. O conceito de eficiência aqui é derivado do menor caminho entre todos os pares de vértices em um componente e será definido a seguir.

Após a formação da rede k AOG, o valor de eficiência de cada componente α é calculado por meio de:

$$\mathcal{E}^{(\alpha)} = \frac{1}{n^{(\alpha)}} \sum_{i \in \alpha} \xi_i^{(\alpha)}, \quad (3.9)$$

onde $n^{(\alpha)}$ denota o número de vértices no componente α e $\xi_i^{(\alpha)}$ é a eficiência local dos nós que pertencem à α , a qual é dada por:

$$\xi_i^{(\alpha)} = \frac{1}{k_i^{out}} \sum_{i \rightarrow j} D_{i,j} \quad (3.10)$$

onde k_i^{out} denota o grau de saída do vértice i , e $D_{i,j}$ é o peso de cada aresta representando um canal de informação entre os nós i e j . Especificamente, os pesos são definidos como a distância Euclidiana entre os vértices, medida na forma de vetor original.

Calculada a eficiência de cada um dos componentes, suponha-se que um novo item de dado y precisa ser classificado. A técnica de alto nível segue os passos abaixo para associá-lo a uma classe:

1. Verifica-se quais dos componentes da rede têm sua eficiência $\mathcal{E}$ melhorada a partir da conexão temporária de y com seus vértices

$$\mathcal{F}_{y,j} = \mathcal{E}_{j \in \alpha}^{(\alpha)} - D_{y,j}, \quad (3.11)$$

onde $D_{y,j}$ refere-se à distância Euclidiana entre os itens de dados y e j na forma de vetor.

2. Cada vértice $j \in \alpha$ cuja conexão com y aumenta a eficiência de α é, temporariamente, conectado à y através de:

$$\Lambda_y^{(\alpha)} = \Lambda_y^{(\alpha)} \cup \begin{cases} j, & \text{se } \mathcal{F}_{y,j} \geq 0, \\ \emptyset, & \text{caso contrário.} \end{cases} \quad (3.12)$$

Desse modo, as medidas de redes complexas podem ser calculadas sobre α após a inserção de y no componente, bem como as variações dessas medidas, conforme definido na técnica de alto nível. Note-se que, caso $\Lambda_y^{(\alpha)} = \emptyset, \forall \alpha \in \mathcal{G}$, a técnica de alto nível emprega o valor de k_α associado a cada componente na formação da rede k AOG, conectando temporariamente os vértices em α que pertencem ao conjunto de k_α vizinhos de y . Logo, as medidas de redes complexas são calculadas sobre tais componentes, e o processo de classificação segue normalmente.

3.1.4 Análise de complexidade

A complexidade de tempo das principais etapas da técnica de classificação proposta é apresentada a seguir. Seja n o número de nós e m o número de arestas da rede, uma vez que a rede

k AOG é esparsa ($k \ll n$), assume-se que $n = m = \mathcal{O}(N)$ nas análises de tempo computacional apresentadas.

- Construção da rede k AOG: dado um conjunto de treinamento com n itens de dados e d atributos, a complexidade para gerar a matriz de distância correspondente é $n(n-1) \cdot d$, resultando em uma ordem de complexidade de $\mathcal{O}(N^2)$. Outras funções, tais como encontrar os componentes do grafo e calcular a medida de pureza e a eficiência dos componentes apresentam custo de $\mathcal{O}(N)$.
- Classificação de baixo nível: o classificador ótimo de Bayes apresenta complexidade da ordem de $\mathcal{O}(N)$, relacionada ao cálculo da distância entre o caso de teste e os dados de treinamento.
- Classificação de alto nível: três medidas de rede são utilizadas na técnica proposta. A assortatividade possui ordem de complexidade de $\mathcal{O}(N)$; o coeficiente de agrupamento tem complexidade de $\mathcal{O}(N \cdot \langle k \rangle^2)$, com $\langle k \rangle$ denotando o grau médio dos nós; e o grau médio que apresenta ordem de complexidade de $\mathcal{O}(N)$.

Tomando todas as etapas da técnica, o custo computacional é da ordem de $\mathcal{O}(N^2)$, relacionado à etapa de construção da rede. Contudo, esta etapa pode ser reduzida para $\mathcal{O}(N^t)$, com $1.06 \leq t \leq 1.33$, usando o método de bissecção de Lanczos (CHEN; FANG; SAAD, 2009). Outro ponto de interesse na análise é o custo computacional da classificação de alto nível. De fato, na maioria das tarefas de classificação, a complexidade computacional de grande parte das medidas de rede são aceitáveis. Todavia, em problemas de grande escala, este pode ser um critério auxiliar importante na escolha das medidas de rede utilizadas.

3.2 Resultados experimentais

Nesta seção, são apresentados resultados de experimentos para avaliar o desempenho da nova técnica, denominada HL- k AOG a partir daqui. A Subseção 3.2.1 apresenta resultados obtidos sobre bases de dados artificiais os quais destacam as principais características de HL- k AOG. Por sua vez, a Subseção 3.2.2 provê resultados experimentais sobre bases de dados reais. Em todos os experimentos apresentados nesta seção, a distância Euclidiana é usada como medida de dissimilaridade.

3.2.1 Bases de dados artificiais

Bases de dados artificiais apresentando padrões explicitamente caracterizados são usadas para avaliar a técnica proposta. Os exemplos mostram situações particulares nas quais técnicas de classificação de baixo nível apresentam dificuldades para prever corretamente os novos itens de dados. Assim, a subseção serve como uma forma de motivar o uso de HL- k AOG.

Art-triângulo: A Fig. 23 mostra a rede k AOG obtida sobre uma base de dados contendo duas classes, denominada Art-triângulo, uma vez que os itens de dados da classe cinza exibem um claro padrão na forma de triângulo. Como técnicas de baixo nível têm problemas em detectar o padrão de formação das classes, pois consideram apenas as características físicas dos dados, o item de teste y (representado pela cor branca na figura) é, normalmente, associado à classe representada pelos itens de cor preta, tal como acontece nos experimentos utilizando classificadores k vizinhos mais próximos, máquina de vetores de suporte, árvore de decisão e também o classificador ótimo de Bayes no grafo k -associados ótimo. Em outras palavras, tais técnicas não identificam o padrão de triângulo formado pelo componente composto por itens de dado de cor cinza.

Figura 23 – Rede k AOG formada a partir dos dados de treinamento da base de dados Art-triângulo.

A Tabela 4 mostra, em detalhes, como a classificação de alto nível é realizada em HL- k AOG. Considerando o novo item de dado y e a rede apresentada na Fig. 23, o processo de classificação começa por calcular as medidas de rede sobre cada componente antes da inserção do item de teste. No passo seguinte, y é inserido temporariamente nos componentes de acordo com a Eq. (3.12) e calcula-se as medidas de rede sobre estes componentes novamente. Dessa maneira, a variação de cada medida de rede (ΔG_y) pode ser obtida e a Eq. (3.5) calculada.

Tabela 4 – Resultados obtidos a partir das medidas de redes complexas usadas para classificação de alto nível da base de dados Art-triângulo. As medidas de redes são assortatividade, coeficiente de agrupamento e grau médio.

C_α	Assortatividade				Coef. de Agrupamento				Grau médio			
	$r^{(\cdot)}$	$r'^{(\cdot)}$	$\Delta G_y^{(\cdot)}(1)$	$f_y^{(\cdot)}(1)$	$CC^{(\cdot)}$	$CC'^{(\cdot)}$	$\Delta G_y^{(\cdot)}(2)$	$f_y^{(\cdot)}(2)$	$\langle k \rangle^{(\cdot)}$	$\langle k \rangle'^{(\cdot)}$	$\Delta G_y^{(\cdot)}(3)$	$f_y^{(\cdot)}(3)$
C_{gray}	0.977	0.968	0.930	0.869	0.118	0.101	0.408	0.256	6.000	5.667	0.657	0.489
C_{black}	0.850	0.851	0.070	0.131	0.188	0.212	0.592	0.744	8.000	7.826	0.343	0.511

A Tabela 5 exhibe o desempenho da técnica proposta sobre as bases de dados artificiais consideradas. A probabilidade de classificação é apresentada para valores $\lambda = 0.2$ e $\lambda = 0.6$,

i.e., quando a técnica de alto nível fornece menor ou maior contribuição para a classificação. Os resultados mostrados na Tabela 5 destacam algumas propriedades interessantes de HL- k AOG. Em primeiro lugar, é possível perceber que a maior contribuição por parte da técnica de alto nível ($\lambda = 0.6$) permite que o item de teste y seja classificado corretamente, quando se considera o padrão de formação dos dados, ao contrário do que ocorre quando é usada maior contribuição do classificador ótimo de Bayes ($\lambda = 0.2$). Em segundo lugar, a formação da rede, a inserção das instâncias de teste e o peso de cada medida de rede na classificação são independentes de parâmetros. Aliás, o único parâmetro a ser ajustado na técnica proposta é o valor de λ . Em terceiro lugar, HL- k AOG trabalha diretamente sobre os componentes da rede, uma vez que a inserção de um único vértice em uma classe inteira de itens de dados pode tornar difícil a captura de grandes variações das medidas de rede.

Tabela 5 – Predição final obtida para as bases de dados artificiais quando $\lambda = 0.2$ e $\lambda = 0.6$.

Base de dados	α	$\lambda = 0.2$	$\lambda = 0.6$
Art-triângulo	<i>gray</i>	0.439	0.516
	<i>black</i>	0.561	0.483
Art-reta I	<i>gray</i>	0.481	0.586
	<i>black</i>	0.519	0.414
Art-reta II	<i>gray</i> ₁	0.000	0.000
	<i>gray</i> ₂	0.200	0.600
	<i>black</i> ₁	0.400	0.200
	<i>black</i> ₂	0.400	0.200
Art-multiclasse	<i>black</i>	0.800	0.400
	<i>gray</i>	0.000	0.000
	<i>green</i>	0.200	0.600

Figura 24 – Rede k AOG obtida para a base de dados Art-reta I. A base de dados é composta por duas classes de itens de dados (cinza e preta), enquanto itens de teste são representados pela cor branca. Claramente, os itens de dado na cor cinza formam uma reta.

Art-reta I. A Fig. 24 mostra a rede *k*AOG formada para uma base de dados de duas classes, na qual os itens de dados na cor cinza formam um padrão de linha reta. Para fins de clareza, os vizinhos mais próximos do item de teste y (cor branca) pertencem às classes cinza, preta e preta, respectivamente. Dessa forma, o classificador ótimo de Bayes, definido na Eq. (2.30), é incapaz de detectar o padrão apresentado pelo componente dos itens de dado cinza, conforme o processo de classificação apresentado a seguir:

1. A probabilidade a priori de y pertencer a cada componente da rede é descrita por:

$$\begin{aligned} P(y \in C_{gray}) &= 0.5 \\ P(y \in C_{black}) &= 0.5 , \end{aligned}$$

onde a medida de *pureza* normalizada dos componentes é usada como probabilidade a priori ($\Phi_{gray} = 1$ e $\Phi_{black} = 1$).

2. A probabilidade da vizinhança Λ_y dado que $y \in C_{gray}$ e dado que $y \in C_{black}$ resulta em:

$$\begin{aligned} P(\Lambda_y | y \in C_{gray}) &= 1/2 \\ P(\Lambda_y | y \in C_{black}) &= 2/3 . \end{aligned}$$

3. O termo de normalização obtido é:

$$P(\Lambda_y) \approx 0.583 .$$

4. O resultado final do classificador ótimo de Bayes sobre a base de dados Art-reta I é dado por:

$$\begin{aligned} P_y^{(Gray)} &= P(y \in C_{gray} | \Lambda_y) \approx \mathbf{0.43} \\ P_y^{(Black)} &= P(y \in C_{black} | \Lambda_y) \approx \mathbf{0.57} . \end{aligned}$$

É importante ressaltar que outras técnicas de classificação de baixo nível também classificam y como item de dado da cor preta. Por outro lado, a Tabela 5 mostra que, com a maior contribuição do termo de alto nível ($\lambda = 0.6$), a técnica pode atribuir a classe correta para y , i.e., a classe em conformidade com a formação de padrão exibida na Fig. 24.

Art-reta II. A Fig. 25 ilustra outra base de dados interessante, onde o item de teste está muito próximo dos componentes compostos por itens de dados de cor preta. Obviamente, isso resulta que técnicas tradicionais classificam y naquela mesma classe. Por outro lado, a grande distância entre o item de teste e os itens de dados da classe cinza tornam ainda mais difícil a classificação de y naquela classe. Contudo, HL-*k*AOG é capaz de fazê-lo por examinar também as relações estruturais entre os itens de dados tal como apresentado pela Tabela 5.

Figura 25 – Rede k AOG gerada para a base de dados Art-reta II, a qual é composta por itens de dados de duas classes (cores cinza e preta). O item de dado na cor branca representa um item de teste a ser classificado.

Art-multiclasse. A Fig. 26 mostra a rede k AOG obtida para uma base de dados com três classes. A Fig. 26a apresenta um caso duvidoso onde não há clareza sobre a formação de padrão relacionada à instância de teste. Dessa forma, HL- k AOG (independente do valor de λ) atribui a classe preta para y . Por outro lado, na Fig. 26b, é exibida uma informação mais representativa sobre o padrão de formação dos dados a qual permite a classificação de y na classe de cor verde pela técnica proposta, quando usando $\lambda = 0.6$ tal como exibido na Tabela 5.

Figura 26 – Rede k AOG formada para as bases de dados Art-multiclasse. As bases de dados são compostas por itens de dados de três classes (cores cinza, preta e verde). O item de dado na cor branca representa um item de teste a ser classificado em ambas as bases: (a) um caso muito duvidoso para detectar alguma formação de padrão em relação à nova instância; (b) formação de padrão clara relacionada ao item de teste.

Análise sobre λ . Uma análise sobre a variação do valor de λ é conduzida a seguir. Neste experimento, bases de dados artificiais geradas a partir de diferentes níveis de mistura de Gaussianas são consideradas. O objetivo é avaliar a combinação linear convexa das associações de

baixo e alto nível na técnica proposta em problemas caracterizados pelo baixo e elevado grau de mistura entre as classes. A Fig. 27 apresenta as bases de dados bem como o desempenho preditivo obtido pela técnica para diferentes valores de $\lambda \in \{0, 0.1, \dots, 1\}$ em dez execuções usando validação cruzada dez vezes estratificada. Note-se que quando $\lambda = 0$, as associações produzidas pelas medidas redes complexas não são tomadas em conta no processo de classificação, i.e., a classificação é realizada exclusivamente pelo classificador ótimo de Bayes. Pela figura, é possível perceber que no cenário com pouca mistura entre as classes, o uso das medidas de redes não é capaz de melhorar a acurácia obtida pela técnica de baixo nível. Por outro lado, na base de dados com elevado grau de mistura entre as classes, a classificação de alto nível melhora o desempenho da técnica de baixo nível. Tais resultados podem ser explicados pela natureza das bases de dados. A primeira, fortemente caracterizada pelos atributos físicos, favorece as técnicas de baixo nível, as quais são essencialmente baseada nessas informações. Logo, valores menores de λ são suficientes para prover bons resultados. Na segunda, caracterizada por um cenário de sobreposição de classes e ruídos, os atributos topológicos extraídos a partir das medidas de rede possuem maior relevância no processo de classificação, recebendo, portanto, maior valor de λ as associações de alto nível. É importante ressaltar que o desempenho da técnica proposta pode ser ainda melhor utilizando seleção de parâmetros, contudo, o objetivo da investigação apresentada neste capítulo é projetar uma técnica capaz de detectar a formação de padrão nos dados caracterizada pela simplicidade.

Figura 27 – Análise da variação de λ na técnica de alto nível proposta considerando duas bases de dados artificiais geradas com diferentes níveis de mistura entre as classes.

3.2.2 Bases de dados reais

A seguir, são apresentados experimentos conduzidos sobre bases de dados reais disponíveis nos repositórios UCI e KEEL (FRANK; ASUNCION, 2010; ALCALÁ-FDEZ *et al.*, 2011). A Tabela 6 apresenta detalhes sobre as bases de dados, os quais estão relacionados ao número de instâncias, atributos, classes e percentual da classe majoritária.

Tabela 6 – Breve descrição das bases de dados reais: *Nome* representa o nome da base de dados, *#Inst.* é o número de instâncias, *#Attr.* é o número de atributos, *#Classes* é o número de classes e *Classe Maj.* denota o percentual de instâncias na classe majoritária.

Nome	#Inst.	#Attr.	#Classes	Classe Maj.
Iris	150	4	3	33.33%
Glass	214	9	7	35.51%
Balance	625	4	3	46.08%
Monks-2	601	6	2	65.72%
Ecoli	336	7	8	42.56%
Append.	106	7	2	80.19%
Thyroid	215	5	3	69.77%
Sonar	208	60	2	53.37%
Digits	5620	64	10	10.18%
SPECTF	267	44	2	79.40%

Nos experimentos, a técnica híbrida é comparada às seguintes técnicas de classificação: árvore de decisão, aprendizado baseado em instância e máquina de vetores de suporte. A seleção de parâmetros é realizada usando o método de busca em grid. Para árvore de decisão, o algoritmo CART é usado e o seguinte parâmetro é configurado: o número mínimo de itens de dados necessários para estar em um nó folha, o qual é uma maneira efetiva de reduzir a variância do modelo¹. Definido por ms , o parâmetro é otimizado sobre o conjunto $ms \in \{0, 1, 2, 3, 4, 5, 10, 15, 20, 30, 50\}$. No aprendizado baseado instância, o classificador k vizinhos mais próximos (k -NN) é utilizado. O valor de k nos experimentos é otimizado sobre o conjunto $k \in \{1, 3, 5, \dots, 31\}$. Para a máquina de vetores de suporte (SVM), o *kernel* função de base radial (RBF) é previamente definido, de modo a reduzir o espaço de busca por parâmetros. O critério de parada, utilizado no método de otimização Karush-Kuhn-Tucker, considerou violação inferior a 10^{-3} . Para cada base de dados, a seleção de modelo é realizada considerando o parâmetro do *kernel* $\gamma \in \{2^4, 2^3, \dots, 2^{-10}\}$ e o custo $C \in \{2^{12}, 2^{11}, \dots, 2^{-2}\}$. Nos experimentos, os parâmetros são selecionados considerando apenas os dados de treinamento. Ademais, os resultados obtidos pelos algoritmos têm sua média calculada em trinta execuções usando validação cruzada dez vezes estratificada.

Ao contrário de várias técnicas de classificação, HL- k AOG não depende de seleção de parâmetros para formação da rede e para produzir as classificações de alto e baixo nível.

¹ Estabelecer um limite inferior para o número de amostras em um nó folha evita cortes desbalanceados, muitas vezes, influenciados por um ruído ou anomalia nos dados

Tabela 7 – Resultados comparativos obtidos por HL-*k*AOG, CART, k-NN e SVM sobre dez bases de dados reais. Cada célula na tabela é representada por “Acc $\pm$ Std” onde “Acc” e “Std” denotam, respectivamente, a média de acurácia e o desvio padrão em trinta execuções usando validação cruzada estratificada dez vezes. Em HL-*k*AOG, o resultado de classificação é obtido a partir do melhor valor entre $\lambda_1 = 0.2$ e $\lambda_2 = 0.6$.

Base de dados	HL- <i>k</i> AOG	CART	k-NN	SVM
Iris	97.33 $\pm$ 3.52 (λ_1)	93.60 $\pm$ 5.59	96.37 $\pm$ 4.63	96.28 $\pm$ 4.02
Glass	70.78 $\pm$ 9.16 (λ_1)	64.12 $\pm$ 9.33	72.64 $\pm$ 8.09	68.61 $\pm$ 7.78
Balance	95.71 $\pm$ 2.40 (λ_1)	88.20 $\pm$ 4.25	89.77 $\pm$ 1.96	99.97 $\pm$ 0.08
Monks-2	96.53 $\pm$ 2.43 (λ_1)	95.67 $\pm$ 2.48	81.26 $\pm$ 5.02	93.79 $\pm$ 3.22
Ecoli	84.90 $\pm$ 5.73 (λ_2)	80.78 $\pm$ 5.55	85.99 $\pm$ 5.11	87.23 $\pm$ 5.22
Append.	83.54 $\pm$ 7.27 (λ_1)	77.24 $\pm$ 9.95	86.99 $\pm$ 8.71	85.72 $\pm$ 8.15
Thyroid	97.30 $\pm$ 3.16 (λ_1)	96.64 $\pm$ 2.95	93.58 $\pm$ 4.74	97.19 $\pm$ 2.59
Sonar	83.75 $\pm$ 8.07 (λ_1)	74.14 $\pm$ 9.69	81.78 $\pm$ 8.08	86.06 $\pm$ 7.43
Digits	98.75 $\pm$ 0.35 (λ_1)	90.27 $\pm$ 1.27	98.79 $\pm$ 0.37	99.26 $\pm$ 0.33
SPECTF	80.07 $\pm$ 5.42 (λ_2)	75.41 $\pm$ 6.20	77.90 $\pm$ 6.78	78.01 $\pm$ 3.92

Além disso, o único parâmetro λ não influencia na classificação produzida pelas técnicas. Dessa forma, foi fixado $\lambda_1 = 0.2$ tomando como base os resultados apresentados em [Silva e Zhao \(2012a\)](#) e $\lambda_2 = 0.6$, a fim de prover uma porção maior da classificação de alto nível. Na subseção anterior, bases de dados artificiais mostraram situações particulares onde técnicas de baixo nível apresentam dificuldades para classificar corretamente. Nesta seção, são avaliados (i) a combinação linear convexa entre o classificador ótimo de Bayes e o classificador de alto nível, e (ii) a influência de λ no contexto de bases de dados reais. Uma vez que a maioria das técnicas de classificação são, essencialmente, baseadas nos atributos físicos dos dados, não há dúvidas de que essa informação geralmente contribui bastante para uma boa classificação. Porém acredita-se que combinadas às características de alto nível, a predição resultante pode alcançar resultados ainda melhores em alguns problemas reais.

A Tabela 7 mostra o desempenho preditivo das técnicas mencionadas acima sobre as bases de dados descritas na Tabela 6. Os valores de cada célula da tabela representam a acurácia média obtida pela técnica e o seu desvio padrão. De modo a analisar estatisticamente os resultados, o teste estatístico de Friedman é adotado conforme sugerido em [Demšar \(2006\)](#), [Iman e Davenport \(1980\)](#). Tal teste verifica se o desempenho preditivo de múltiplos classificadores sobre várias bases de dados possui diferença. Para conduzir o teste, os classificadores são ranqueados de acordo com seu desempenho em cada base de dados, tal como exibido na Tabela 8. A hipótese nula afirma que todos os algoritmos são equivalentes, portanto, seus ranques devem ser iguais. O resultado do teste de Friedman χ_F^2 e de Iman F_F são dados, respectivamente, por:

$$\chi_F^2 = \frac{12 \times 10}{4 \times 5} \left[1.9^2 + 3.7^2 + 2.5^2 + 1.9^2 - \frac{4 \times 5^2}{4} \right] = 12.96 ,$$

Tabela 8 – Ranque médio com o desempenho das técnicas em cada base de dados.

Dataset	HL- k AOG	CART	kNN	SVM
Iris	1	4	2	3
Glass	2	4	1	3
Balance	2	4	3	1
Monks-2	1	2	4	3
Ecoli	3	4	2	1
Append.	3	4	1	2
Thyroid	1	3	4	2
Sonar	2	4	3	1
Digits	3	4	2	1
SPECTF	1	4	3	2
Av.Rank	1.9 ± 0.83	3.7 ± 0.64	2.5 ± 1.02	1.9 ± 0.83

$$F_F = \frac{(10 - 1) \times 12.96}{10 \times (4 - 1) - 12.96} = 6.85 .$$

Executando o teste para um nível de confiança de 95%, o valor crítico de F , distribuído de acordo com a distribuição F , é $F(3, 27) = 2.96$. Desse modo, a hipótese nula é rejeitada, pois $F_F > F_{0.05}(3, 27) \rightarrow 6.85 > 2.96$, i.e., as técnicas analisadas não são estatisticamente equivalentes. Diante disso, o pós-teste de Nemenyi é adotado para encontrar quais técnicas registram resultados superiores. Novamente, um nível de confiança de 95% é adotado. A diferença crítica CD encontrada é:

$$CD = 2.569 \times \sqrt{\frac{4 \times 5}{6 \times 10}} = 1.483 .$$

Os resultados do pós-teste indicam que HL- k AOG e SVM são estatisticamente equivalentes e que elas superam o algoritmo CART. Este é um resultado promissor, porque HL- k AOG, diferente de outras técnicas tradicionais, é capaz de capturar relações espaciais, estruturais e topológicas entre os dados, além de possuir apenas um parâmetro. Aliás, os resultados na tabela mostram que a maior porção da classificação de alto nível (λ_2) pode melhorar a predição final em algumas bases de dados reais, o que fornece fortes evidências de que tais bases de dados apresentam padrões que podem ser detectados, considerando, principalmente, a estrutura semântica dos dados ao invés de apenas seus atributos físicos.

3.3 Considerações finais

Neste capítulo, foi apresentado HL- k AOG, uma técnica híbrida baseada em rede para a classificação de dados. A técnica explora vantagens de duas técnicas recentes na literatura, o grafo k -associados ótimos e a classificação de alto nível mediante medidas de redes complexas.

Especificamente, a primeira técnica provê uma rede construída sem uso de parâmetros, enquanto a segunda é capaz de capturar relações semânticas globais entre os dados. Como consequência dessa combinação, a técnica resultante é caracterizada pela simplicidade e pela habilidade em detectar a formação de padrão nos dados.

Na etapa de treinamento, a rede k AOG é construída a partir da otimização de uma medida de *pureza*. Na etapa de classificação, a combinação linear convexa entre os classificadores de baixo e alto nível é responsável por fornecer a predição de novos itens de dados. Como classificador de baixo nível, HL- k AOG emprega o classificador ótimo de Bayes. Por sua vez, as associações de alto nível são produzidas pela variação de medidas de redes complexas. Em comum, os classificadores usam a mesma rede k AOG para extrair as informações necessárias para o seu processo de inferência, e a completa ausência de seleção de parâmetros durante esse processo. Em comparações realizadas com técnicas de classificação tradicionais muito usadas, como SVM, por exemplo, a técnica híbrida apresentou desempenho comparável, além de possuir menor número de parâmetros. Ademais, experimentos realizados em bases de dados artificiais enfatizaram que HL- k AOG não sofre com as mesmas limitações de técnicas de baixo nível, tais como a dificuldade em considerar a formação de padrão dos dados.

Outras contribuições apresentadas no capítulo são: (i) desenvolvimento de uma estratégia de inserção temporária de novos itens de dados, baseada na eficiência da troca de informação entre os vértices dos componentes; e (ii) maior sensibilidade às variações das medidas de rede por aplicá-las sobre componentes em vez de diretamente sobre classes. Em algumas bases de dados, as variações das medidas podem ser muito insignificantes devido, por exemplo, ao grande número de itens de dados rotulados para uma classe específica, ou mesmo pelo padrão que estas instâncias exibem. Nesse sentido, a análise sobre componentes é uma estratégia interessante para distinguir o nível de conformidade dos novos itens aos padrões de cada componente. Relembre-se que uma classe pode ser representada por intermédio de vários componentes usando a representação gráfica.

CLASSIFICAÇÃO ORGANIZACIONAL DE DADOS BASEADA NO CONCEITO DE IMPORTÂNCIA EM REDES COMPLEXAS

Neste capítulo, é apresentado um novo conceito de classificação derivado da abstração gráfica provida pelo uso de redes. Enquanto técnicas tradicionais levam em conta apenas os atributos físicos dos dados, técnicas baseadas em redes também são capazes de considerar as relações topológicas entre os itens de dados. Diferente da classificação de alto nível, que verifica a conformidade de padrão por meio de medidas de redes, a técnica proposta a seguir avalia a importância individual dos itens de dado para determinar o rótulo de uma nova instância. Em poucas palavras, o conceito de importância aqui é definido em termos do PageRank, algoritmo usado na *engine* de busca do Google para definir a importância de páginas da web. Ademais, a técnica é capaz de considerar a relação organizacional dos dados combinando atributos espaciais e estruturais mediante um refinamento da medida de eficiência apresentada no capítulo anterior, a *eficiência diferencial espaço-estrutural*.

O PageRank é um algoritmo conhecido na literatura, geralmente utilizado na resolução de problemas de busca, caracterização de redes sociais e recuperação da informação (LANGVILLE; MEYER, 2011; LU; GETOOR, 2003). No contexto de aprendizado de máquina, PageRank está fortemente relacionado ao aprendizado não supervisionado, especialmente agrupamento de dados (TABRIZI *et al.*, 2013; GLEICH; SESHADHRI, 2012). Apesar dos conceitos do PageRank ainda não terem sido aplicados para classificação de dados, há evidências de que eles podem trazer novas características para este tipo de tarefa. Por exemplo, considere os dois métodos de formação da rede, k NN e k AOG. Sabe-se que tais métodos permitem que nós da rede construída possuam o mesmo grau de saída (k), enquanto o grau de entrada de cada nó não é controlado pelos métodos. Isto significa que as conexões de entrada emergem de maneira mais natural. Este é exatamente o conceito por trás da formulação do PageRank, uma vez que ele

mede a importância das páginas da web baseado nas conexões de entrada de cada página, i.e., uma página não é muito importante se ela apenas cita outras, porém não é citada. Pelo contrário, a importância de uma página é dada pelas citações que ela recebe de outras. Dessa forma, os conceitos de importância derivados do PageRank parecem bastante apropriados para projetar técnicas de aprendizado supervisionado baseado em redes.

4.1 Descrição da técnica

Seja $\mathcal{X}_{train} = \{x_i, i = 1, \dots, n\}$ um conjunto de dados rotulados em que para cada item x_i é associada uma única classe $l_i \in \mathcal{L}$, e seja $\mathcal{X}_{test} = \{x_i, i = 1, \dots, m\}$ um conjunto de dados cujos rótulos deseja-se predizer. Dado que cada item de dado é descrito por d atributos $x_i = \{x_{i1}, \dots, x_{id}\}$ e $\mathcal{X}_{train} \cap \mathcal{X}_{test} = \emptyset$, a técnica proposta pode ser dividida em duas etapas, treinamento e classificação, como descrito a seguir.

Treinamento: Nesta etapa, a técnica proposta utiliza um método de formação de rede, como, por exemplo, os métodos gerais apresentados na Subseção 2.3.1, para construir uma rede a partir dos itens de dados rotulados disponíveis em $\mathcal{X}_{train}$. A rede formada é denotada por $\mathcal{G} = \{V, E\}$, onde V é o conjunto de nós e E o conjunto de arestas. Cada nó $v_i \in V$ denota um item de dado rotulado $x_i \in \mathcal{X}_{train}$ e $e_{ij} \in E$ representa uma conexão do vértice v_i para o vértice v_j . A informação necessária para realizar a classificação é dada pelas medidas de *eficiência diferencial espaço-estrutural* e PageRank, as quais são obtidas a partir de $\mathcal{G}$, indicando que a rede construída é vital para a predição e que os padrões dos dados de treinamento serão extraídos a partir dela. Uma vez que a rede é construída e a instância de teste é inserida, os atributos físicos dos dados de treinamento não são mais relevantes para o processo de classificação.

Classificação: Nesta etapa, um novo item de dado $y \in \mathcal{X}_{test}$ é apresentado ao classificador. O item é temporariamente conectado a um conjunto de vértices da rede de acordo com a medida de *eficiência diferencial espaço-estrutural* dos componentes da rede. Em relação à predição, ela acontece por calcular a importância de y em cada classe a partir das conexões temporárias estabelecidas para ele. Finalmente, y é atribuído para a classe na qual ele possui maior valor de *importância*.

O processo de classificação da técnica proposta é ilustrado passo a passo na Fig. 28. A rede formada a partir de $\mathcal{X}_{train}$ é mostrada pela Fig. 28a; a medida de *eficiência* calculada para cada componente é exibida pela Fig. 28b; o valor de *importância* de cada nó rotulado é calculado e mostrado pela Fig. 28c. Posteriormente, uma instância de teste y é conectada a um conjunto de nós de acordo com a medida de *eficiência diferencial espaço-estrutural* (Fig. 28d); a *importância* de y é, então, obtida para cada classe e exibida na Fig. 28e, e a classificação de y na classe em que ele apresenta o maior valor de *importância* é descrito pela Fig. 28f.

Figura 28 – Exemplo ilustrativo do funcionamento da abordagem de classificação proposta. (a) Uma rede $\mathcal{G}$ é formada a partir dos dados de treinamento $\mathcal{X}_{train}$; (b) o valor de *eficiência* de cada componente em $\mathcal{G}$ é calculado; (c) a *importância* de cada nó rotulado é calculada; (d) uma instância de teste y é conectada, temporariamente, a outros nós de acordo com a medida de *eficiência diferencial espaço-estrutural*; (e) a *importância* de y para cada classe é calculada; (f) y é classificado para a classe na qual possui o maior valor de *importância*. Na figura, d significa a distância Euclidiana entre dois nós (dois itens de dados), e é a eficiência local de cada nós, ϵ representa a eficiência de cada componente, i denota a *importância* de um nó e $\mathcal{I}$ é a *importância* recebida por um item de teste em uma dada classe.

4.1.1 Classificação baseada em importância

A seguir, a técnica proposta é descrita, bem como as medidas de redes que fazem parte dela. Inicialmente, suponha-se que uma rede $\mathcal{G}$ seja formada a partir de um conjunto de dados rotulados $\mathcal{X}_{train}$. Suponha-se, também, uma instância de teste $y \in \mathcal{X}_{test}$, a qual precisa ser classificada. De acordo com a Fig. 28, a nova técnica classifica y por calcular a *importância* de sua inserção em cada classe. Formalmente, a *importância* de um item de teste y , denotada por $\mathcal{I}$, em relação à classe $l \in \mathcal{L}$ é dada por:

$$\mathcal{I}_y^{(l)} = \sum_{j \in \Lambda_y^{(l)}} \mathcal{I}_j, \quad (4.1)$$

onde $j \in \mathcal{X}_{train}$ denota um vértice rotulado, $\Lambda_y^{(l)}$ é o conjunto de nós que pertencem à classe l na qual y é temporariamente conectado, e $\mathcal{I}_j$ representa a *importância* do vértice j . De maneira geral, a Eq. (4.1) calcula o valor de *importância* da instância de teste y em relação a cada classe

l . Em sequência, a classe atribuída à y é aquela em que ele recebe o maior valor de *importância*, i.e.

$$\phi_y = \arg \max_{l \in \mathcal{L}} \mathcal{I}_y^{(l)} . \quad (4.2)$$

A *importância* de um nó $j \in \mathcal{X}_{train}$ é quantificada por iterar o seguinte sistema, o qual é equivalente à formulação do PageRank:

$$\mathcal{I}_j^{(t+1)} = \sum_{i \rightarrow j} \beta \cdot \frac{\mathcal{I}_i^{(t)}}{k_i^{out}} + (1 - \beta) \frac{1}{n} , \quad (4.3)$$

onde $i \rightarrow j$ representa uma aresta do nó i para o nó j , $\mathcal{I}_i$ refere-se à *importância* do nó i , k_i^{out} representa o grau de saída do nó i , e n é o número total de nós da rede. Como uma maneira de garantir a convergência dessa formulação, β significa a probabilidade de saltos aleatórios, a qual é definida como 0.85 (BOLDI; SANTINI; VIGNA, 2005). Por razões de clareza, quando $t = 0$, nós definimos a *importância* de cada nó j como $\mathcal{I}_j^{(0)} = 1/n$ como condição inicial. Em relação ao número de iterações do sistema apresentado na Eq. (4.3), nós verificamos a seguinte condição:

$$S = \sum_j | \mathcal{I}_j^{(t+1)} - \mathcal{I}_j^{(t)} | , \quad (4.4)$$

onde $S \leq 10^{-3}$ é o critério de parada para o sistema.

A seguir é definido o conjunto de conexões temporárias para y , denotado por $\Lambda_y^{(l)}$ na Eq. (4.1), o qual é obtido a partir da seguinte formulação:

$$\Lambda_y^{(l)} \cup \{j \mid \mathcal{F}_{y,j} \geq 0 \text{ e } j \in l\} \quad (4.5)$$

onde $\mathcal{F}_{y,j}$, denominado aqui *eficiência diferencial espaço-estrutural*, é uma função que verifica se uma conexão entre y e j aumenta ou diminui a eficiência do componente que contém o nó j . Se ela aumenta a eficiência, então o nó j é adicionado ao conjunto $\Lambda_y^{(l)}$. Por outro lado, se a eficiência de nenhum componente é aumentada por qualquer ligação para y , i.e., $\mathcal{F}_{y,j} < 0$ para todo j , os nós cujas conexões para y menos diminuem a eficiência de um componente são adicionados ao conjunto $\Lambda_y^{(l)}$.

A *eficiência diferencial espaço-estrutural* apresentada aqui mede uma combinação de informações físicas e topológicas na rede e também nos dados originais na forma de vetor. Topologicamente, essa medida quantifica a eficiência em enviar informação entre nós de um componente (LATORA; MARCHIORI, 2001). Fisicamente, a informação flui mediante ligações ponderadas pela distância Euclidiana entre os nós (itens de dados). A formulação matemática de $\mathcal{F}_{y,j}$ é dada por:

$$\mathcal{F}_{y,j} = \mathcal{E}^{(\alpha)} \cdot \gamma - D_{y,j} , \quad (4.6)$$

onde $\mathcal{E}^{(\alpha)}$ é a medida de eficiência do componente α , a qual é dada pela Eq. (3.9) e $D_{y,j}$ refere-se à distância Euclidiana entre itens de dados y e j na forma de vetor. O primeiro termo tem

objetivo de capturar propriedades estruturais dos dados, enquanto o segundo termo captura as características físicas dos dados. Portanto, a medida de *eficiência diferencial espaço-estrutural* combina características físicas e topológicas embutidas nos dados de entrada. Basicamente, a função $\mathcal{F}$ resulta em um valor positivo elevado, se o par de nós y e j está próximo no espaço físico dos dados (tomando os vetores de atributos) e o componente α possui alto valor de medida de *eficiência* no espaço topológico; enquanto $\mathcal{F}$ assume valores baixos (até mesmo negativos), se y e j estão distantes e a medida de eficiência é baixa. Na Eq. (4.6), γ é um parâmetro para balancear os dois termos. A introdução de tal parâmetro permite que a técnica de classificação trabalhe com diferentes métodos de formação de rede. Note-se que quando γ é elevado, a medida topológica possui alta influência, i.e., a medida de *eficiência diferencial espaço-estrutural* enfatiza a captura de propriedades estruturais da rede de dados; por outro lado, quando γ recebe um valor pequeno, a atenção da medida se volta mais para os atributos físicos dos dados de treinamento.

4.1.2 Algoritmo e análise de complexidade

Esta subseção apresenta os principais passos da classificação baseada em *importância* por meio de um algoritmo geral. Na etapa de treinamento, a técnica recebe um conjunto de dados $\mathcal{X}_{train}$ a partir do qual a rede é formada, i.e., $\mathcal{X}_{train} \rightarrow \mathcal{G}$. Em seguida, $\mathcal{G}$ é utilizado para obter os padrões de eficiência de cada componente $\mathcal{G} \rightarrow \mathcal{E}$ e a *importância* de cada nó rotulado $\mathcal{G} \rightarrow \mathcal{I}$. Na etapa de classificação, a técnica recebe um item de dado $y \in \mathcal{X}_{test}$ cuja classe precisa ser predita. Então, o algoritmo seleciona nós da rede para se conectarem temporariamente a y de acordo com os padrões de eficiência $\mathcal{E}$. Posteriormente, a *importância* de y para cada classe é calculada. No fim, y recebe o rótulo da classe onde possua maior *importância*. O Alg. 3 descreve os passos relacionados às etapas de treinamento e teste.

Algoritmo 3: Classificação baseada em *importância*.

Data: $\mathcal{X}_{train}$

Input: $y \in \mathcal{X}_{test}$

Output: ϕ_y

// Etapa de treinamento

1 $\mathcal{G} \leftarrow$ Construa a rede usando $\mathcal{X}_{train}$

2 $\mathcal{E} \leftarrow$ Calcule a *eficiência* dos componentes em $\mathcal{G}$ (Eq. 3.9)

3 $\mathcal{I} \leftarrow$ Calcule o valor de *importância* dos nós em $\mathcal{G}$ (Eq. 4.3)

// Etapa de classificação: para cada item de teste $y \in \mathcal{X}_{test}$

4 $\Lambda_y \leftarrow$ Selecione conexões temporárias para y (Eq. 4.5)

5 $\mathcal{I}_y \leftarrow$ Calcule o valor de *importância* de y para cada classe (Eq. 4.1)

6 $\phi_y \leftarrow$ Classifique y por *importância* (Eq. 4.2)

A seguir, a complexidade de tempo do Alg. 3 é analisada. Para fins de clareza, nós assumimos que o número de nós $n = \mathcal{O}(N)$ e o número de arestas $m = \mathcal{O}(N)$, uma vez que as redes formadas são esparsas por utilizar os métodos de construção kNN ou $kAOG$ ($k \ll N$).

1. A construção da rede, no passo 1 do algoritmo, possui ordem computacional de $\mathcal{O}(N^2)$, uma vez que é necessário calcular as similaridades entre todos os pares de instâncias.
2. O cálculo da *eficiência* do componente, no passo 2, requer uma ordem computacional de $\mathcal{O}(N)$ correspondente a encontrar os componentes da rede e computar a medida de *eficiência*.
3. O valor de *importância* no passo 3 pode ser prontamente calculado, por iterar o sistema apresentado na Eq. (4.3) até que o critério de parada seja satisfeito. Portanto, a ordem computacional desse passo é $\mathcal{O}(N)$, uma vez que o número de iterações $it \ll N$.
4. A ordem de complexidade do passo 4 é $\mathcal{O}(N)$, a qual está relacionada à seleção de vértices a serem conectados temporariamente a y de acordo com a Eq. (4.5).
5. O passo 5 do algoritmo possui complexidade de tempo $\mathcal{O}(N)$ no pior caso, quando todos os nós da rede são selecionados para conexão com y no passo anterior.
6. Finalmente, o passo 6 requer uma busca simples para encontrar em qual classe o item de teste recebe mais *importância*. Levando em conta que o número de classes $c \ll N$, esse passo é constante $\mathcal{O}(1)$.

Colocando todos os passos juntos, a complexidade computacional da técnica proposta é $\mathcal{O}(N^2 + N + N + N + N + 1)$. Tomando o termo de mais alta ordem, ela resulta em $\mathcal{O}(N^2)$. Contudo, tal complexidade pode ser reduzida usando o método de formação de rede denominado bissecção de Lanczos (CHEN; FANG; SAAD, 2009). Nesse caso, o passo 1 possui complexidade de $\mathcal{O}(N^t)$ e a complexidade da técnica é reduzida para $\mathcal{O}(N^t + N + N + N + N + 1)$. Além disso, de acordo com Chen, Fang e Saad (2009), um valor pequeno para o parâmetro t ($1.06 \leq t \leq 1.33$) é suficiente para obter redes de alta qualidade. Assim, a ordem de complexidade computacional da técnica proposta encontra-se entre $\mathcal{O}(N^{1.06})$ e $\mathcal{O}(N^{1.33})$.

4.2 Resultados experimentais

Esta seção apresenta avaliações empíricas para analisar o desempenho da técnica proposta. Ela é dividida em duas subseções: Subseção 4.2.1 provê resultados sobre bases de dados artificiais, a fim de destacar as características particulares da técnica proposta sobre outras técnicas; e Subseção 4.2.2 discute resultados de experimentos realizados em bases de dados reais, mostrando que a técnica proposta é competitiva, quando comparada a várias técnicas de classificação bastante utilizadas. A distância Euclidiana é usada como medida de dissimilaridade em todos os experimentos apresentadas nesta seção.

4.2.1 Bases de dados artificiais

Na Fig. 29, são exibidas duas bases de dados simples com padrões muito claros. Na Fig. 30, são mostradas as redes formadas a partir de ambas bases de dados para uso da técnica proposta. Primeiramente, considera-se a Fig. 29a. Técnicas de classificação tradicional colocam a instância de teste y_1 (denotada como Branco/ $\diamond$) na classe Vermelha/ $\square$. Isso acontece porque há muitos elementos da classe Vermelha/ $\square$ perto dela. Contudo vê-se que a instância de teste y_1 é claramente a continuação de um padrão de Linha formada pelos elementos da classe Azul/ $\circ$. Portanto, y_1 deve ser classificado na classe Azul/ $\circ$ de acordo com o seu significado semântico. Analisando a formação das fronteiras de decisão das classes nas Figs. 31a-31f, é possível perceber que técnicas tradicionais falham em identificar padrões semânticos formados pelos dados. Apesar de uma grande variedade de parâmetros ser testada, a estrutura organizacional dos dados (o padrão de Linha) não é detectada pelas técnicas tradicionais. Por outro lado, a partir do resultado de classificação mostrado pela Fig. 32a, é possível ver que a técnica baseada em *importância* pode classificar corretamente y_1 na classe Azul/ $\circ$, na qual o item de teste forma o padrão de uma Linha com os outros elementos da mesma classe.

Figura 29 – Base de dados composta por itens de dados de duas classes (Azul/ $\circ$ e Vermelho/ $\square$) os quais apresentam padrões claros. (a) Branco/ $\diamond$ e Preto/ $\diamond$ são itens de teste que precisam ser classificados. (b) Preto/ $\triangle$ é uma instância de teste a ser classificada.

Considera-se agora o segundo item de dado y_2 na Fig 29a, denotado por Preto/ $\diamond$. Nesse caso, a técnica proposta e as técnicas tradicionais são capazes de classificar corretamente a instância de teste associando-a à classe Vermelha/ $\square$. Contudo, desta vez, a técnica de classificação de alto nível, proposta em Silva e Zhao (2012a), associa y_2 à classe errada (classe Azul/ $\circ$) sem o termo de baixo nível. Este exemplo mostra que a técnica proposta possui habilidade para capturar características físicas e estruturais dos dados de entrada devido à introdução da medida de *eficiência diferencial espaço-estrutural* na etapa de inserção dos itens de teste.

A seguir, analisam-se os resultados obtidos sobre a base de dados apresentada na Fig. 29b, em que um item de teste representado pelo Preto/ $\triangle$ precisa ser classificado. Como aconteceu no exemplo anterior, técnicas tradicionais também falham em capturar o padrão aqui. Realmente,

Figura 30 – Redes construídas a partir das bases de dados da Fig. 29 contendo padrões de Linha e Triângulo, respectivamente. O método de formação de rede kNN ($k = 3$) foi usado na base de dado Linha e $kAOG$ na base de dados Triângulo. O valor de *importância* de alguns nós também são exibidos na figura.

para técnicas de classificação tradicionais, não há um padrão de Triângulo, apenas um conjunto de pontos esparsamente distribuídos. Esta situação é descoberta por meio das Figs. 31g-31l. Por outro lado, nossa técnica realiza classificação por extrair medidas topológicas a partir da relação entre os dados da rede formada. Dessa forma, ela é capaz de detectar o padrão de formação dos conjuntos de dados sem adição de qualquer técnica tradicional, como exibido pela Fig. 32b.

Em síntese, a técnica descrita neste capítulo explora ambas propriedades físicas e estruturais dos dados de maneira unificada. Mais que isso, o conceito de *importância* dá uma visão intuitiva, porém nova para a classificação de dados.

4.2.1.1 Análise de parâmetros

Nesta subseção, analisa-se o desempenho da técnica por variar os valores de seus parâmetros. Primeiramente, estuda-se o efeito do parâmetro γ na medida de *eficiência diferencial espaço-estrutural* (Eq. 4.6). As fronteiras de decisão geradas pela técnica proposta em função de γ são exibidas pela Fig. 33. Quando γ é pequeno (veja Figs. 33a e 33e), a instância de teste é conectada temporariamente a alguns nós locais e, consequentemente, sua *importância* é influenciada por apenas aqueles nós. Neste caso, padrões de dados globais podem não ser capturados; quando $\gamma = 1.0$ (veja Figs. 33b e 33f), o item de teste é temporariamente conectado de acordo com os padrões de eficiência obtidos pelo método de formação de rede; quando γ assume maiores valores (veja Figs. 33c e 33g), a *importância* da instância de teste é avaliada mais globalmente na rede. Se γ é muito elevado (veja Figs. 33d e 33h), as características locais dos dados se tornam irrelevantes e sua *importância* é influenciada pelos nós mais importantes de

Figura 31 – Fronteiras de decisão formadas por diferentes técnicas de classificação sobre as mesmas bases de dados mostradas na Fig. 29. Itens de dados Azul/○ e Vermelho/□ denotam dados de treinamento enquanto Branco/◇, Preto/◇ e Preto/△ indicam que y_1 , y_2 e y_3 são instâncias de teste a serem classificadas. As fronteiras de decisão consideram apenas características físicas dos dados, as quais tornam a classificação de y_2 fácil. Por outro lado, essas técnicas não são capazes de detectar a formação de padrão dos dados, o que torna difícil classificar corretamente y_1 e y_3 . Note-se que os parâmetros das técnicas tradicionais são otimizados. Mesmo assim, elas ainda são incapazes de classificar os itens de teste y_1 e y_3 na classe Azul/○.

toda a rede.

A próxima análise compreende o desempenho preditivo da técnica proposta quando variando os parâmetro k e γ . Cinco bases de dados artificiais com diferentes níveis de mistura das classes são usadas nessa análise. Os resultados de classificação apresentados consideram dez execuções usando validação cruzada dez vezes estratificada. A Fig. 43 exibe a média de acurácia para cada base de dados em função de $k \in \{1, 2, \dots, 10\}$ e $\gamma \in \{0, 1, \dots, 20\}$. A partir desta figura, é possível perceber que pequenos valores de k e γ normalmente fornecem melhores resultados para baixo níveis de mistura, enquanto valores intermediários de γ obtêm melhores resultados para dados com elevado nível de mistura. Em todos os casos, o parâmetro k possui pequena influência sobre os resultados de classificação, i.e., a técnica não é sensível a k .

4.2.1.2 Análise de tempo computacional

Esta subseção analisa o consumo de tempo da técnica proposta em comparação com algumas técnicas tradicionais. Nos experimentos, cinco bases de dados contendo diferentes números de itens de dados $n \in \{10^2, 10^3, \dots, 10^6\}$ são considerados. As técnicas tradicionais

Figura 32 – Resultados de classificação da técnica proposta sobre as mesmas bases de dados apresentadas na Fig. 31. As instâncias de teste são classificadas de acordo com seus padrões correspondentes pela técnica baseada em *importância*. O valor de *importância* recebido pelos itens de teste em cada classe é mostrado nas subfiguras.

Figura 33 – Análise da abordagem proposta sobre as bases de dados mostradas na Fig. 29. Cada subfigura apresenta as decisões geradas pela técnica proposta com um valor específico do parâmetro γ .

usadas aqui são o algoritmo de árvore de decisão CART, e o algoritmo de máquina de vetores de suporte (SVM) com *kernel* RBF. CART é selecionado devido ao seu baixo consumo de tempo, enquanto SVM em razão de sua alta representatividade entre técnicas de classificação. Para a técnica desenvolvida, têm-se considerado dois valores extremos de k : $k = 1$ e $k = 10$. As simulações são conduzidas em um PC Intel Core i7-2600 com 16GB de RAM utilizando apenas um processador para cada técnica e com programas codificados em Python. A Fig. 35 mostra o tempo de execução em segundos de todas as técnicas sob comparação. Apesar de CART ser um limite inferior, a técnica proposta com ambos valores de k desempenha muito próximo em termos de consumo de tempo. Por outro lado, SVM é incapaz de concluir sua execução dentro de um dia, mesmo com diferentes combinações de parâmetros. É importante notar que códigos otimizados para CART e SVM, os quais estão disponibilizados nos projetos scikit-learn (PEDREGOSA *et*

(a) Bases de dados 2-Moons

(b) Bases de dados 2-Gaussianas

Figura 34 – Análise preditiva da técnica proposta em função dos parâmetros k e γ sobre bases de dados artificiais geradas com diferentes níveis de mistura entre as classes.

al., 2011) e libsvm (CHANG; LIN, 2011), são usados nesse experimento.

Figura 35 – Análise de consumo de tempo em função do número de instâncias na base de dados.

4.2.2 Bases de dados reais

A seguir, são apresentados resultados de classificação usando a técnica proposta em bases de dados reais bem como comparações com técnicas de classificação bastante utilizadas. As simulações foram conduzidas sobre seis bases de dados reais disponíveis no repositório da UCI (FRANK; ASUNCION, 2010) e KEEL (ALCALÁ-FDEZ *et al.*, 2011). A Tabela 15 mostra detalhes sobre as bases de dados. Como pode ser visto na tabela, as bases de dados foram selecionadas de modo a obter diversidade de domínio de dados, bem como considerar diferentes números de classes (de 3 à 11), atributos (de 4 à 64), e tamanhos (de 106 à 5620).

Tabela 9 – Breve descrição das bases de dados reais em termos do número de instâncias (#Inst.), número de atributos (#Attr.), e número de classes (#Classes).

Nome	#Inst.	#Attr.	#Classes
Glass	214	9	7
Iris	150	4	3
Sonar	208	60	2
Opt.Digits	5620	64	10
Thyroid	215	5	3
Vowel	990	13	11

Nos experimentos, a técnica proposta emprega os métodos de formação de rede *k*NN e *k*AOG. Quando a rede é formada pelo método *k*NN, a técnica é chamada PgRkNN, e ela é

denominada PgRkAOG quando usa o método k AOG. Como descrito na Subseção 2.3.2.1, a rede k AOG não requer seleção de parâmetro, uma vez que ela é criada levando em conta a pureza dos componentes da rede. Assim, PgRkAOG toma os padrões de eficiência como obtidos pelo método k AOG, i.e., $\gamma = 1.0$. Por outro lado, PgRkNN possui dois parâmetros: k e γ , uma vez que a rede k NN requer um valor de k para construção da rede. A partir de análises empíricas, PgRkNN é otimizado em função de $k \in \{1, 2, \dots, 6\}$ e $\gamma \in \{0.1, 0.2, \dots, 2.0\}$.

Para cada base de dados, PgRkNN e PgRkAOG são comparados às seguintes técnicas: máquina de vetores de suporte (SVM), florestas aleatórias (RF), k vizinhos mais próximos (k -NN), classificador ótimo de Bayes usando grafo k -associados ótimo (k -AOG) e árvore de decisão (CART). Para experimentos com SVM, o espaço de busca dos parâmetros foi reduzido por adotar a função de base radial (RBF) como *kernel*. O critério de parada empregado para o método de otimização *Karush-Kuhn-Tucker* considerou violação inferior a 10^{-3} . Para cada base de dados, a seleção de modelo é realizada considerando o parâmetro do *kernel* $\gamma \in \{2^4, 2^3, \dots, 2^{-10}\}$ e o custo $C \in \{2^{12}, 2^{11}, \dots, 2^{-2}\}$. Para a técnica k -NN, o valor de k é otimizado sobre o seguinte conjunto $k \in \{1, 2, 3, \dots, 30\}$, o qual é suficiente para prover os melhores resultados para esse algoritmo. O número de árvores na floresta aleatória (RF) é definido como $t = 128$ seguindo o estudo apresentado em Oshiro, Perez e Baranauskas (2012). Todos os parâmetros de SVM, k -NN e PgRkNN são selecionados usando o algoritmo de busca em grid mediante validação cruzada três vezes estratificada. k -AOG, CART e PgRkAOG não precisam de seleção de parâmetro. Finalmente, os resultados obtidos pelos algoritmos têm sua média calculada sobre trinta execuções usando validação cruzada dez vezes estratificada.

A Tabela 10 apresenta o desempenho preditivo das técnicas mencionadas acima sobre as bases de dados reais descritas na Tabela 9. Nesta tabela, cada célula é representada por “Acc. $\pm$ Std.”, onde “Acc.” denota a acurácia média de cada técnica e “Std.” representa o desvio padrão das acurácias. De modo a analisar estatisticamente os resultados, o teste estatístico de Friedman é adotado, conforme sugerido em Demšar (2006), Iman e Davenport (1980). Para conduzir o teste, os classificadores são ranqueados de acordo com seu desempenho em cada base de dados, tal como exibido na Tabela 11. A hipótese nula afirma que todos os algoritmos são equivalentes, portanto, seus ranques devem ser iguais. O resultado do teste de Friedman χ_F^2 e de Iman F_F são dados, respectivamente, por:

$$\chi_F^2 = \frac{12 \times 6}{7 \times 8} \left[3.08^2 + 4.42^2 + 3.75^2 + 3.25^2 + 7.0^2 + 2.67^2 + 3.83^2 - \frac{7 \times 8^2}{4} \right] = 16 ,$$

$$F_F = \frac{(6-1) \times 16}{6 \times (7-1) - 16} = 4 .$$

Executando o teste para um nível de confiança de 95%, o valor crítico de F é dado por $F(6, 30) = 2.42$. Desse modo, a hipótese nula é rejeitada, pois $F_F > F_{0.05}(6, 30) \rightarrow 4 > 2.42$,

Tabela 10 – Resultados comparativos obtidos por SVM, RF, k-NN, k-AOG, CART, PgRkNN e PgRkAOG sobre seis bases de dados reais. “Acc” e “Std.” denotam, respectivamente, a média de acurácia preditiva e o desvio padrão sobre trinta execuções usando validação cruzada dez vezes estratificada. Os resultados de SVM, k-NN e PgRkNN mostram os parâmetros ajustados a partir de seleção de modelo.

Algoritmo	Glass	Iris	Opt.Digits	Sonar	Thyroid	Vowel	Av.Rank
SVM (C, γ)	69.70 ± 7.9 ($2^{11}, 2^3$)	97.09 ± 3.8 ($2^8, 2^3$)	99.31 ± 0.3 ($2^2, 2^2$)	87.50 ± 6.9 ($2^2, 2^3$)	97.18 ± 3.4 ($2^8, 2^2$)	98.20 ± 1.2 ($2^7, 2^3$)	3.08
RF	75.94 ± 7.6	97.09 ± 3.9	98.28 ± 0.5	82.95 ± 7.7	97.01 ± 3.5	95.58 ± 2.0	4.42
k-NN (k)	70.61 ± 8.6 (3)	95.31 ± 4.8 (2)	98.76 ± 0.4 (1)	83.53 ± 7.1 (1)	96.80 ± 3.5 (5)	98.96 ± 1.0 (1)	3.75
k-AOG	70.55 ± 8.4	97.26 ± 3.7	98.76 ± 0.4	83.59 ± 7.4	97.48 ± 3.0	97.16 ± 1.6	3.25
CART	64.82 ± 9.1	94.29 ± 5.25	90.30 ± 1.7	73.84 ± 8.9	96.30 ± 3.7	76.02 ± 3.8	7.00
PgRkNN (k, γ)	71.75 ± 7.9 (3,0.7)	97.20 ± 3.7 (1,0.4)	98.94 ± 0.4 (3,0.6)	82.00 ± 7.5 (3,0.6)	97.55 ± 3.0 (1,0.1)	98.49 ± 1.2 (6,0.1)	2.67
PgRkAOG	69.82 ± 8.6	97.40 ± 3.7	98.53 ± 0.4	83.76 ± 8.1	96.60 ± 4.3	97.27 ± 1.6	3.83

Tabela 11 – Ranque médio com o desempenho das técnicas em cada base de dados.

Alg.	Glass	Iris	Opt.Digits	Sonar	Thyroid	Vowel	Av.Rank
SVM	6	4.5	1	1	3	3	3.08
RF	1	4.5	6	5	4	6	4.42
k-NN	3	6	3.5	4	5	1	3.75
k-AOG	4	2	3.5	3	2	5	3.25
CART	7	7	7	7	7	7	7.00
PgRkNN	2	3	2	6	1	2	2.67
PgRkAOG	5	1	5	2	6	4	3.83

i.e., as técnicas analisadas não são estatisticamente equivalentes. Diante disso, o pós-teste de Nemenyi é adotado para encontrar quais técnicas registram resultados superiores. Novamente, um nível de confiança de 95% é adotado. A diferença crítica CD encontrada é:

$$CD = 2.948 \times \sqrt{\frac{7 \times 8}{6 \times 6}} = 3.677 .$$

Os resultados do pós-teste indicam que SVM, k -AOG e PgRkNN superam o algoritmo CART e seus desempenhos são equivalentes e comparáveis à RF, k -NN e PgRkAOG. Este resultado é bastante atrativo, porque fornece evidências de que a técnica proposta é capaz de lidar com bases de dados fortemente caracterizadas pelos atributos físicos, tais como distância ou distribuição. Ora, as técnicas tradicionais se concentram nas características físicas dos dados, porque tal informação, realmente, contribui para uma boa classificação em uma grande variedade de bases de dados. Contudo, na próxima seção, é exibido um problema de classificação em que melhores resultados preditivos podem ser obtidos por considerar ambas propriedades físicas e topológicas dos dados.

4.2.3 Estudo de caso: detecção de doenças cardíacas

Esta subseção apresenta em detalhes a aplicação da técnica proposta para detectar anormalidades em tomografias cardíacas. Especificamente, duas bases de dados de SPECT (*Single-Photon Emission Computed Tomography*) cardíacos são consideradas. Os dados foram coletados pelo Colégio Médico de Ohio e os passos de pré-processamento são apresentados em [Kurgan et al. \(2001\)](#). Basicamente, as bases de dados consistem do diagnóstico do coração de 267 pacientes classificados em duas categorias: normal ou anormal. A primeira base de dados, chamada “Cardiac Images” aqui, é processada para extrair características que sumarizam as imagens de SPECT originais, resultando em 44 atributos contínuos para cada paciente. A segunda base de dados, denominada “Cardiac Diagnosis”, é composta de 22 atributos binários relacionados ao diagnóstico parcial obtido usando um conjunto de regras definidas por especialistas humanos. Ambas as bases de dados já estão divididas em dados de treinamento e teste,

os quais contêm 80 instâncias rotuladas igualmente distribuídas em ambas as classes e 187 instâncias não rotuladas.

Uma pré-análise dos dados fornece evidências de que as classes são bastante misturadas, quando levando em conta apenas as características físicas dos dados. A Fig. 36 mostra a distribuição das classes após a aplicação do algoritmo de análise de componente principal (PCA) considerando os três componentes mais representativos da base de dados “Cardiac Images”. Na figura, os itens de dados rotulados são representados pelos marcadores Verde/□ e Vermelho/○, os quais denotam as classes de pacientes anormais e normais, respectivamente. Os itens de teste são representados por Azul/△. Assim, o desempenho da técnica proposta é avaliado em um cenário de elevada sobreposição.

Figura 36 – Distribuição das classes da base de dados “Cardiac Images”. A figura mostra os primeiros três componentes obtidos a partir do algoritmo de análise de componente principal (PCA). As instâncias rotuladas são representadas por marcadores quadrados (cor verde) e círculos (cor vermelha), os quais denotam as classes de pacientes com coração anormal e normal, respectivamente. As instâncias não rotuladas são representadas por pequenos triângulos (cor azul).

A Tabela 12 mostra o desempenho preditivo de PgRkNN e PgRkAOG sobre as bases de dados em comparação com as mesmas técnicas de classificação tradicionais empregadas na Tabela 7. Como na subseção anterior, todos os parâmetros de SVM, k-NN e PgRkNN são otimizados empregando o algoritmo de busca em grid sobre um processo de validação cruzada três vezes estratificada considerando apenas os dados de treinamento. De acordo com a tabela, em ambas as bases de dados, PgRkAOG supera as outras técnicas em termos de acurácia preditiva, seguido por PgRkNN e as bem conhecidas técnicas SVM e RF. Ao mesmo tempo, k-NN, k-AOG, e CART obtêm baixo desempenho preditivo. Embora k-AOG use o mesmo método de construção da rede que PgRkAOG, o primeiro faz previsões tomando uma estratégia fortemente baseada nas características físicas dos dados, enquanto o segundo classifica as instâncias de teste por *importância*. Em essência, o desempenho das técnicas propostas são atrativos, indicando que levar em conta outras propriedades dos dados, além das características físicas, pode, realmente,

resultar em melhoria de desempenho, mesmo em cenários com classes bastante sobrepostas e com número de instâncias razoáveis, situações em que abordagens de classificação tradicionais, normalmente, apresentam dificuldades. Além disso, a classificação baseada em *importância* possui duas características particulares em relação às técnicas de classificação de dados existentes: (i) ela não negligencia a relevância espacial de cada item de dado no processo de classificação, e (ii) ela leva em conta as ligações recebidas pelos nós (grau de entrada), uma vez que é mais fácil fazer conexões do que recebê-las, tal como acontece nas páginas da Internet.

Tabela 12 – Resultados preditivos sobre as bases de dados “Cardiac Images” e “Cardiac Diagnosis”. SVM, RF, k-NN, k-AOG e CART denotam máquina de vetores de suporte, floresta aleatória, k-vizinho mais próximo, classificador do grafo k-associado ótimo e árvore de classificação e regressão, respectivamente. Os termos PgRkNN e PgRkAOG referem-se à abordagem proposta usando os métodos de formação de rede *k*NN e *k*AOG, respectivamente. Resultados de SVM, k-NN e PgRkNN mostram os parâmetros ajustados por meio de seleção de modelo. Os melhores resultados estão em negrito.

Algoritmo	C. Images	C. Diagnosis
SVM (C, γ)	77.54 ($2^0, 2^{-9}$)	81.28 ($2^{-1}, 2^{-2}$)
RF	76.47	77.01
k-NN (k)	62.03 (6)	59.89 (7)
k-AOG	56.68	65.24
CART	65.77	69.52
PgRkNN (k, γ)	79.14 (4, 1.3)	81.82 (1, 0.5)
PgRkAOG	91.44	87.70
CLIP3 (KURGAN <i>et al.</i> , 2001)	77.00	80.00
SVM (CIECHOLEWSKI, 2011)	77.54	80.75

Outra comparação pode ser realizada com trabalhos correlatos da literatura. Na Tabela 12, “CLIP3 (KURGAN *et al.*, 2001)” e “SVM (CIECHOLEWSKI, 2011)” denotam resultados de classificação obtidos em Kurgan *et al.* (2001) e Ciecholewski (2011), respectivamente. Apesar da diferença relacionada ao processo de otimização de parâmetros desses algoritmos, seus desempenhos preditivos são superados pela técnica proposta.

A seguir, apresenta-se uma análise das redes formadas, as quais são ilustradas na Fig. 37. Como o número de atributos dos dados é maior que três, usa-se o algoritmo *Fruchterman-Reingold* para gerar as redes (FRUCHTERMAN; REINGOLD, 1991). Na figura, itens de dados Vermelho/□ estão relacionados à classe normal e Azul/○ à classe anormal. A Fig. 37a exibe a rede usada por PgRkAOG sobre a base de dados “Cardiac Images”, a qual é esparsa e ajustada aos dados, visto que é construída por otimizar a *pureza* dos componentes. Por outro lado, a Fig. 37b exibe o grafo explorado por PgRkNN, o qual é mais denso que o anterior, e fornece várias possibilidades para investigar relações complexas entre os itens de dados. Por exemplo, a rede em PgRkNN deriva duas propriedades distintas para cada componente (classe): a classe anormal (instâncias ○) é um tipo de subrede exponencial, cuja maioria dos nós têm aproximadamente o mesmo grau de entrada; e a classe normal (instâncias □) possui alguma característica de

Figura 37 – Redes exploradas por (a) PgRkAOG e (b) PgRkNN para conduzir o processo de classificação sobre a base de dados “Cardiac Images”.

redes livres de escala, ao passo que muitos nós apresentam pequeno valor de grau de entrada, enquanto poucos nós apresentam um valor maior (tais nós são denominados *hubs*). Acredita-se que essas duas propriedades podem ser entendidas pela natureza dos dados, já que a classe anormal compreende um variedade de diferentes doenças cardíacas, que inclui isquemia, infarto e outras; e a classe normal compreende um padrão mais comum, o qual está associado com os *hubs*.

Outra análise interessante é provida pela Fig. 38, que mostra o processo de classificação de um item de teste da base de dados “Cardiac Images”, representado como item de dado Preto/ $\triangle$, pela técnica proposta. Após a formação da rede, tal como mostrado pela Fig. 37a, a técnica seleciona conexões temporárias para a instância de teste, considerando informações físicas e estruturais da rede, e a nova instância é associada à classe onde possua o maior valor de *importância* (0.065): classe anormal (Azul/ $\circ$). Contudo a mesma instância representa um caso de equívoco para técnicas de classificação tradicionais, as quais a classificam erroneamente por levarem em conta apenas as características físicas dos dados. Por exemplo, na mesma figura, exibem-se os seis vizinhos mais próximos da instância de teste, os quais são utilizados pelo algoritmo k-NN para classificar o paciente na classe normal (Vermelho/ $\square$). Assim, acredita-se que a técnica proposta também contribui para a tarefa de classificação de dados por considerar a estrutura organizacional dos dados além das características físicas. Nesse sentido, ela pavimenta um caminho para se realizar uma classificação um pouco mais próxima à humana por meio da máquina.

Figura 38 – Um item de teste da base de dados “Cardiac Images” representado pelo marcador Preto/ $\triangle$. A instância de teste denota um caso de equívoco para a classificação tradicional, uma vez que é erroneamente classificada quando considerando apenas as características físicas. Por exemplo, na figura são mostrados os seis vizinhos mais próximos da nova instância, os quais são analisados pelo algoritmo k-NN para classificar o paciente na classe normal (Vermelha/ $\square$). Por outro lado, o item de teste é corretamente predito pela técnica proposta a qual leva em conta a classe na qual ele possui maior *importância* (0.065): classe anormal (Azul/ $\circ$).

4.3 Resultados experimentais com outra medida de similaridade

Em todas as análises apresentadas até então neste capítulo, a mesma medida de similaridade foi usada para definir as conexões entre os itens de dados e também ponderá-las. As conexões da rede são criadas entre vértices com maior similaridade e a eficiência do componente captura padrões de conexões nos componentes a partir dos pesos das arestas. Dessa maneira, a medida de similaridade tem a sua influência na técnica de classificação proposta. Enquanto a distância Euclidiana é, de longe, a medida mais usada na literatura geral, sabe-se que há outras medidas tão importantes quanto. A seguir, é apresentada a aplicação da técnica para a classificação de imagens adotando a medida de Bhattacharyya.

4.3.1 Estudo de caso: reconhecimento de padrões invariantes

As bases de dados empregadas nos experimentos fazem parte da coleção ETH-80, a qual compreende um total de 3280 imagens divididas em 8 categorias: Maçã, Carro, Vaca, Xícara, Cachorro, Cavalo, Pêra e Tomate, conforme os exemplos apresentados na Fig. 39. Cada categoria contém 10 objetos com diferentes variações dentro da classe, mas, ainda assim, pertencendo à mesma categoria. As variações de cada um desses objetos são representadas por meio de 41 imagens cuidadosamente capturadas, conforme exemplo exposto na Fig. 40.

Figura 39 – Exemplo de todas as categorias da base de dados ETH-80. Cada categoria é composta por dez objetos distintos.

(a) Categoria Maçã - Objeto 2

(b) Categoria Maçã - Objeto 7

Figura 40 – Exemplo de dois objetos pertencentes à categoria Maçã da base de dados ETH-80. Cada objeto é representado por meio de 41 imagens em diferentes ângulos e rotações.

O pré-processamento das bases de dados envolveu os seguintes passos: o tamanho das imagens foi reduzido de 128×128 para 32×32 de modo a tornar o processamento mais rápido; os atributos das imagens foram extraídos de seu histograma; e a similaridade entre as imagens é calculada usando o coeficiente de Bhattacharyya. A validação cruzada estratificada com dez

pastas é usada nos experimentos e os resultados são obtidos pela média de 10 execuções. Os parâmetros da técnica baseada em importância são selecionados da mesma maneira como descrita na seção anterior.

Ainda sobre os experimentos, o desempenho preditivo da técnica é avaliado sobre cada categoria separadamente. Como as categorias compreendem vários objetos bastante similares, a tarefa é considerada difícil. A Tabela 13 apresenta os resultados obtidos pela técnica baseada em importância com o método de formação de rede k NN. Como um modo de prover alguma base de comparação, os resultados registrados em [Cupertino \(2013\)](#) a partir de uma técnica baseada em facilidade de acesso ([CUPERTINO; ZHAO; CARNEIRO, 2015](#)), usando caminhada aleatória, são tomados como referência (RWLP). Os resultados obtidos mostram que a abordagem de classificação baseada em importância alcança bom desempenho preditivo na detecção de padrões invariantes e que as etapas de construção da rede bem como do cálculo de eficiência dos componentes capturam informações relevantes para o processo de classificação usando o coeficiente de Bhattacharyya. Logo, a técnica também desempenha bem, quando empregada com medidas de distância (dissimilaridade) diferentes da Euclidiana.

Tabela 13 – Resultados preditivos sobre as diferentes categorias que compõe a base de dados ETH-80. RWLP e PgRkNN denotam a técnica de classificação por facilidade de acesso usando caminhada aleatória (*random walk limiting probabilities*) e a técnica de classificação baseada no conceito de *importância* em redes complexas, respectivamente. Os resultados de ambas as técnicas mostram os parâmetros ajustados por meio de seleção de modelo.

Categoria	RWLP(α, τ)	PgRkNN(k, γ)
Maçã	86.8 $\pm$ 0.8 (0.5,1)	90.9 $\pm$ 0.6 (8,0.3)
Carro	89.1 $\pm$ 0.6 (0.5,1)	95.5 $\pm$ 0.2 (1,0.1)
Vaca	65.3 $\pm$ 1.3 (0.5,41)	75.2 $\pm$ 0.7 (7,0.2)
Xícara	81.1 $\pm$ 1.0 (0.4,47)	96.2 $\pm$ 0.6 (9,0.5)
Cachorro	82.1 $\pm$ 1.0 (0.5,39)	91.8 $\pm$ 0.5 (10,0.1)
Cavalo	75.4 $\pm$ 0.9 (0.45,1)	91.1 $\pm$ 0.8 (8,0.1)
Pêra	74.3 $\pm$ 0.9 (0.55,47)	80.6 $\pm$ 0.9 (3,0.1)
Tomate	89.9 $\pm$ 0.8 (0.95,1)	94.4 $\pm$ 0.6 (6,0.1)

4.4 Considerações finais

Neste capítulo foi apresentada uma técnica de classificação de dados baseada em medidas de redes complexas. A técnica é capaz de sobrepujar limitações de classificadores tradicionais, trabalhando sobre uma combinação de propriedades distintas dos dados, além dos atributos físicos. Em poucas palavras, a técnica extrai propriedades físicas e topológicas da rede construída via medidas de redes complexas, capturando a *importância* de cada nó rotulado para, posteriormente, classificar um dado item de teste na classe em que ele recebe o maior valor de *importância*. O conceito de *importância* usado, aqui, foi derivado da formulação do PageRank.

Vários experimentos computacionais foram conduzidos em bases de dados artificiais e reais, de modo a avaliar as características e a eficiência da técnica proposta. Os resultados revelaram que a mesma é capaz de tirar vantagem de propriedades espaciais e estruturais da rede, analisando as relações entre os dados. Tais características refletiram em bom desempenho preditivo, quando comparada com técnicas de classificação bastante usadas, como, por exemplo, SVM e RF. Ademais, a aplicação da classificação por *importância* na detecção de doenças cardíacas revelou que considerar propriedades distintas dos dados pode melhorar o desempenho de classificação, mesmo em cenários com elevado nível de mistura entre as classes.

Na literatura, a técnica de classificação de alto nível é a que mais se relaciona com a técnica proposta neste capítulo. Em uma breve comparação, esta última apresenta novidades significantes nos níveis tecnológico e conceitual. Do ponto de vista tecnológico, a técnica desenvolvida não exige a adição de uma técnica de classificação tradicional ao modelo, i.e., a técnica considera atributos físicos e propriedades da rede por si própria, o que reduz consideravelmente o tempo computacional e a complexidade da seleção de parâmetro. No nível conceitual, a abordagem fornece uma nova visão sobre o processo de classificação, o qual pode considerar divisão do espaço, conformidade de padrão ou caracterização de importância.

COMPUTAÇÃO BIOINSPIRADA PARA OTIMIZAÇÃO ESTRUTURAL DE REDES E SUAS APLICAÇÕES

Este capítulo apresenta um *framework* projetado para realizar otimização estrutural de redes. Ao contrário de trabalhos correlatos, os quais definem um valor específico de conexões para todos os nós (ou grupos de nós), a investigação apresentada a seguir propõe que as conexões entre os nós sejam atualizadas iterativamente, usando uma técnica de otimização bioinspirada. O *framework* deve construir a rede enquanto otimiza uma função de qualidade orientada à tarefa, por exemplo, a acurácia na tarefa de classificação de dados. Especificamente, o processo de otimização é avaliado em duas funções de qualidade: a classificação baseada em importância e a classificação híbrida usando técnicas de baixo e alto nível. Como mostrado pelos experimentos, o processo de otimização conduzido pelo *framework* é capaz de melhorar o desempenho preditivo de técnicas baseadas em rede.

5.1 Descrição geral da técnica

O *framework* de otimização estrutural proposto neste capítulo é dividido em duas etapas. A primeira é denominada etapa de otimização (ou treinamento), a qual compreende a otimização da rede. Nesta fase, um algoritmo de otimização por enxame de partículas (PSO) é usado para construir a rede a partir dos dados de treinamento X_{train} de acordo com a otimização de uma dada função de qualidade f sobre um dado conjunto de dados de validação X_{valid} . Na segunda etapa, denominada etapa de teste, a melhor solução na etapa de treinamento é equipada por f para classificar novos itens de dados $y \in X_{test}$. A função f é orientada à tarefa investigada. Por exemplo, considerando a tarefa de classificação, f pode representar a acurácia obtida pela técnica baseada em rede. Os principais passos do *framework* geral para otimização estrutural são

ilustrados pela Fig. 41. Inicialmente, o algoritmo de otimização cria um enxame de partículas, em que cada partícula representa uma rede $\mathcal{G}_i$. Então, em cada iteração t , as partículas são avaliadas e atualizadas (Δ) usando a função de qualidade f . No fim, o algoritmo retorna a partícula com o melhor valor de qualidade, o qual contém a rede a ser adotada na etapa de teste para classificar quaisquer novos itens de dados.

Figura 41 – *Framework* geral para otimização estrutural no aprendizado baseado em rede. A ilustração enfatiza as duas etapas da técnica: otimização (também denominada etapa de treinamento) e teste.

5.1.1 Algoritmo de otimização por enxame de partículas

O algoritmo de otimização por enxame de partículas baseado em aprendizado social (SL-PSO) é um paradigma de inteligência de enxames proposto recentemente para problemas de otimização de grande escala (CHENG; JIN, 2015). Ao contrário de outros algoritmos tradicionais de otimização por enxame de partículas (KENNEDY; EBERHART, 1995; KENNEDY, 2006), SL-PSO não armazena o histórico de melhores posições globais nem locais. Em SL-PSO, o enxame é ordenado de acordo com os valores de qualidade das partículas, e consequentemente, cada partícula aprende a partir de quaisquer outras partículas melhores no enxame atual, como segue:

$$s_{i,j}(t+1) = \begin{cases} s_{i,j}(t) + vel_{i,j}(t+1), & \text{se } p_i(t) \leq Prob_i^L \\ s_{i,j}(t), & \text{caso contrário} \end{cases} \quad (5.1)$$

onde t é o contador da geração, $s_{i,j}(t)$ é a j -ésima ($j \in \{1, 2, 3, \dots, D\}$) dimensão no vetor de posição da partícula i ($i \in \{1, 2, 3, \dots, m\}$), com m e D denotando o tamanho do enxame e o

número de variáveis de decisão, respectivamente; $s_{i,j}(t)$ é atualizado de acordo com a velocidade $vel_{i,j}(t+1)$ baseada sobre uma *probabilidade de aprendizado* $Prob_i^L$ definida a priori para cada partícula i no enxame ordenado; ε é um parâmetro conhecido como fator de influência social; $F_1(t)$, $F_2(t)$ e $F_3(t)$ são três coeficientes gerados aleatoriamente entre $[0, 1]$. Em detalhes, $vel_{i,j}(t+1)$ é gerado da seguinte maneira:

$$vel_{i,j}(t+1) = F_1(t) \cdot vel_{i,j}(t) + F_2(t) \cdot I_{i,j}(t) + F_3(t) \cdot \varepsilon \cdot S_{i,j}(t), \quad (5.2)$$

onde $vel_{i,j}(t+1)$ consiste de três componentes: o *componente da inércia* $F_1(t) \cdot vel_{i,j}(t)$, o *componente da imitação* $I_{i,j}(t)$ e o *componente de influência social* $S_{i,j}(t)$. O componente da inércia é similar àquele no PSO tradicional, enquanto os componentes de imitação e de influência social são inspirados das teorias de aprendizado social, tal como:

$$\begin{cases} I_{i,j}(t) = s_{k,j}(t) - s_{i,j}(t), \\ S_{i,j}(t) = \bar{s}_j(t) - s_{i,j}(t). \end{cases} \quad (5.3)$$

onde $s_{k,j}(t)$ é a j -ésima dimensão no vetor de posição da partícula k , conhecida como demonstrador, a qual tem melhor valor de qualidade que a partícula i ; $\bar{s}_j(t) = \frac{\sum_{i=1}^m s_i^j}{m}$ é a posição média do enxame na geração t .

Os resultados experimentais em [Cheng e Jin \(2015\)](#) demonstram que SL-PSO apresenta desempenho robusto em comparação com vários algoritmos estados-da-arte para otimização em grande escala. No trabalho, o algoritmo é avaliado sobre 47 funções *benchmark* com escalas variando de 30 à 1000 dimensões.

5.1.2 Framework de otimização estrutural bioinspirado

O processo de otimização estrutural no *framework* proposto é conduzido pelo algoritmo SL-PSO. Os principais passos do algoritmo são apresentados no Alg. 4. Em cada iteração, SL-PSO avalia e atualiza as partículas de acordo com as melhores partículas, de modo que cada partícula pode aprender de quaisquer outras partículas mais bem avaliadas. O algoritmo termina quando o número máximo pré-definido de iterações é atingido.

5.1.2.1 Representação geral

Para ser manipulado tanto pelo algoritmo bioinspirado quanto pela técnica baseada em rede adotada em f , uma representação computacional eficiente foi projetada para as redes. Dado um enxame de m partículas $P = \{\mathcal{P}_1, \dots, \mathcal{P}_m\}$, cada partícula $\mathcal{P}_k \in P$ é representada como segue:

$$\mathcal{P}_k = \{v_1, \dots, v_n\}, \quad (5.4)$$

Algoritmo 4: Framework para otimização estrutural de redes

```

1 begin
2   Gere uma matriz de mapeamento (usando MapLab ou MapAll) para determinar o
   conjunto de possíveis configurações da rede
3   Inicie cada partícula  $\mathcal{P}_i \in P$  por escolher aleatoriamente uma configuração da rede a
   partir do conjunto gerado no passo anterior
4   while critério de parada não é satisfeito do
5     Calcule o valor de qualidade  $F_i$  para cada partícula  $\mathcal{P}_i$ 
6     Ordene  $P$  de acordo com os valores de qualidade  $F$ 
7     Atualize a melhor solução  $\mathcal{G}^*$ 
8     Atualize as partículas de acordo com as Eqs.(5.1)-(5.3)
9   return  $\mathcal{G}^*$ ;

```

onde n significa o número de itens de dados (ou nós). Cada nó $v_i \in \mathcal{P}_k$ denota um item de dado rotulado $x_i \in X_{train}$ e suas ligações são representadas por:

$$v_i = \{e'_{i1}, \dots, e'_{iq}\}, \quad (5.5)$$

com q denotando o grau de saída máximo e e'_{ij} denotando a probabilidade de uma ligação entre dois nós. Os valores assumidos por e'_{ij} são contínuos e variam entre $[0, 1]$, a fim de serem manipulados por SL-PSO. Por outro lado, as propriedades topológicas da rede são capturadas por meio de conexões, i.e., valores binários denotados por e_{ij} , os quais são obtidos como:

$$e_{ij} = \begin{cases} 1, & \text{se } e'_{ij} \geq 0.5, \\ 0, & \text{caso contrário.} \end{cases} \quad (5.6)$$

Para calcular a qualidade de uma dada partícula $\mathcal{P}_k$, ela é convertida em uma rede $\mathcal{G}_k = \{V_k, E_k\}$, onde $V_k = \{1, \dots, n\}$ denota os vértices associados a cada item de dado e as arestas são dadas por:

$$E_k \cup \{(i, M_{ij}) \mid M_{ij} > 0\}, \quad (5.7)$$

onde M é uma matriz $n \times q$, a qual contém em cada linha i os nós que recebem ligações do vértice i . Formalmente, M é dado por:

$$M_{ij} = e_{ij} \cdot Map_{ij}. \quad (5.8)$$

onde Map é a heurística de mapeamento utilizada, a qual possui em cada linha i os vértices que podem ser conectados por v_i . Note-se que, apesar de Map manter todas as conexões possíveis a partir de cada vértice, a existência ou não de qualquer conexão é dada pela probabilidade de conexão e' , a qual é manipulada pelas partículas no *framework* de otimização.

5.1.2.2 Heurísticas de mapeamento

A heurística de mapeamento Map tem por objetivo reduzir a complexidade do *framework* proposto. Dada uma rede com n nós, a complexidade do espaço de busca tende a $\mathcal{O}(n^2)$, uma vez

que ela é dada pelo número total de arestas possíveis. Para o *framework* de otimização, contudo, visto vez que n pode ser tão grande como centenas ou mesmo milhares, uma complexidade de $\mathcal{O}(n^2)$ é, inviavelmente, cara. Para resolver esse problema, foi projetado o conceito de heurística de mapeamento, o qual cria um espaço subdimensional de tamanho q baseado nos atributos dos dados de treinamento X_{train} . Consequentemente, a heurística reduz a complexidade do espaço de busca de $\mathcal{O}(n^2)$ para $\mathcal{O}(n)$ ($\mathcal{O}(n \cdot q)$, $q \ll n$), onde q é o número máximo de arestas possíveis a partir de um dado nó. Duas heurísticas de mapeamento são propostas aqui: *MapLab* e *MapAll*. A primeira enfatiza a estrutura das classes por meio das conexões, o que pode ser uma característica muito útil em bases de dados caracterizadas por padrões muito bem definidos. A segunda é mais preocupada em considerar a mistura entre as classes através da estrutura da rede, o que é uma característica interessante, ao lidar com bases de dados caracterizadas por padrões complexos e repletas de ruídos. A seguir, ambas heurísticas são descritas em detalhes.

MapLab: A heurística de mapeamento *MapLab* compreende os seguintes passos:

- Calcule a similaridade entre cada par de itens de dados (nós);
- Selecione para cada nó v_i seus q nós mais similares que pertencem à classe l_i ;
- Dado que $1 \leq z \leq q$, crie a matriz $Map_{n \times q}$:

$$Map_{iz} = v_z. \quad (5.9)$$

MapAll: A heurística de mapeamento *MapAll* compreende os seguintes passos:

- Calcule a similaridade entre cada par de itens de dados;
- Selecione para cada vértice v_i seus q nós mais similares;
- Dado que $1 \leq z \leq q$, crie a matriz $Map_{n \times q}$:

$$Map_{iz} = \begin{cases} v_z, & \text{se } l_i = l_z, \\ 0, & \text{caso contrário.} \end{cases} \quad (5.10)$$

A Fig. 42 traz um exemplo ilustrando a aplicação das duas heurísticas. Em princípio, analisa-se as duas matrizes obtidas. Para um dado vértice v_j , *MapLab* define os vértices $\{8, 2, 7, 1, 3\}$ como possíveis conexões (Fig. 42a), enquanto *MapAll* define $\{8, 2, 7, 0, 1\}$ (Fig. 42b). Note-se que o valor 0 em *MapAll* significa que o vértice v_4 não pertence à mesma classe que v_j , logo tal conexão não ocorrerá. Por outro lado, *MapLab* considera apenas vértices da mesma classe que v_j quando estabelecendo os seus q vizinhos. Considerando-se agora a representação adotada para o processo de otimização, cada posição do vetor v_j representa a probabilidade de ligação, denotada por e'_{jk} em SL-PSO: $\{0.23, 0.65, 0.83, 0.51, 0.39\}$. Posteriormente, tal representação é transformada em um vetor binário usando a Eq. (5.6), a qual define a existência ou

não das conexões de v_j na rede: $\{0, 1, 1, 1, 0\}$. A partir de tal representação, uma multiplicação simples entre o vetor binário v_j e a heurística de mapeamento Map_j fornece os vértices que são conectados por v_j (Eq. 5.8), tal como ilustrado pelas Figs. 42a e 42b.

Figura 42 – Exemplo ilustrativo sobre a conversão do vetor de probabilidade (manipulado por SL-PSO) para rede (na qual as propriedades estruturais são exploradas), e sobre as heurísticas de mapeamento propostas. (a) *MapLab* considera apenas vizinhos mais próximos de mesmo rótulo para criar a matriz *Map*; (b) *MapAll* leva em conta itens de dados de todas as classes para criar a matriz *Map*.

5.1.2.3 Função de qualidade

Na etapa de otimização, cada partícula $\mathcal{G}_i$ é avaliada por uma função de qualidade f sobre um conjunto de dados de validação. Na etapa de teste, a melhor partícula $\mathcal{G}^*$ é usada por f para realizar a classificação de todos os novos itens de dados um por um. Dessa maneira, a função de qualidade f está relacionada à tarefa a ser desempenhada pelo grafo otimizado. Por exemplo, nos experimentos apresentados na Subseção 5.2, f emprega a técnica de classificação baseada em importância.

5.1.2.4 Análise de complexidade

Esta subsecção provê uma análise sobre a complexidade de tempo do *framework* de otimização. Para simplicidade, a análise computacional é realizada para cada um dos módulos seguintes:

- Otimização da rede: SL-PSO é empregado para conduzir o processo de otimização no *framework* proposto. A complexidade de tempo do algoritmo bioinspirado é $\mathcal{O}(md)$ considerando que $d > m$, onde m e d significam o número de partículas e dimensões, respectivamente. Como $d = qn$, onde n denota o número de itens de dados em X_{train} e q o número máximo de possíveis conexões, e $q \ll n$, a complexidade de tempo de SL-PSO no *framework* proposto encontra-se em torno de $\mathcal{O}(mn)$.
- Matriz de mapeamento: A complexidade de tempo relacionada às heurísticas de mapeamento *MapLab* e *MapAll* é $\mathcal{O}(n^2)$, uma vez que a dissimilaridade entre cada par de instâncias precisa ser calculada antes da busca pelos vizinhos mais próximos.
- Função de qualidade: A ordem de complexidade da função de qualidade f é orientada à tarefa. Supondo que a técnica baseada em *importância* é adotada, a complexidade é $\mathcal{O}(V + E)$, onde $V = n$ denota o número de vértices e E é o número de arestas. Visto que as redes construídas pelo *framework* proposto são esparsas ($q \ll n$), considera-se $\mathcal{O}(E) = \mathcal{O}(V) = \mathcal{O}(n)$ e, conseqüentemente, a complexidade de tempo é $\mathcal{O}(n)$.

É conhecido que a função de qualidade é, normalmente, o passo com maior custo computacional de algoritmos bioinspirados. Ela compreende a avaliação de cada solução candidata durante o processo iterativo até que o critério de parada seja satisfeito. Dessa forma, um algoritmo sequencial para o *framework* proposto tem ordem de complexidade de $\mathcal{O}(n^2 + gmn)$, com g denotando o número de iterações, m o número de partículas e n o maior valor entre o número de nós e arestas. Apesar desta ser uma complexidade de tempo maior que do método de construção kNN , o qual normalmente é de $\mathcal{O}(n^2 + n \log n)$, ela permite a aplicação do *framework* de otimização em um grande número de problemas de classificação. Além disso, a complexidade pode ser reduzida por tomar vantagem de computação distribuída, característica própria da maior parte dos algoritmos baseados em população.

5.2 Aplicação do framework para a classificação baseada em importância

Esta seção investiga o uso do framework para a otimização estrutural de redes usando a técnica de classificação baseada em importância como função de qualidade. Especificamente, é esperado que o framework construa a rede que otimize o desempenho preditivo em cada base

de dados. A fim de realizar uma avaliação criteriosa, várias bases de dados artificiais e reais são empregadas, além de comparações com métodos de formação de redes bastante usados na literatura, bem como técnicas de classificação tradicionais.

5.2.1 Função de qualidade

A função de qualidade adotada aqui consiste no uso da técnica de classificação baseada em importância. A função f toma os padrões de eficiência dos componentes tal como obtido pelos métodos de formação da rede ($\gamma = 1$). Basicamente, a técnica baseada em importância tira vantagem de atributos físicos e topológicos dos dados e emprega a medida PageRank para classificar novos itens de dados nos componentes da rede onde eles recebem maior importância. Mais detalhes sobre a técnica podem ser encontrados no Capítulo 4.

5.2.2 Resultados em bases de dados artificiais

Esta subseção apresenta análises empíricas do *framework* de otimização que enfatizam benefícios em obter a rede mediante um processo de otimização em comparação com outros métodos de construção.

As bases de dados artificiais consideradas aqui são mostradas pela Fig. 43, onde é possível perceber diferentes níveis de mistura entre as classes. Cada simulação é executada adotando um processo de validação cruzada dez vezes estratificada. Neste processo, os itens de dados são divididos em 10 pastas e, em cada execução, nove pastas são usadas como dados de treinamento e uma pasta como dados de teste, resultando em um total de 10 execuções. Os dados de treinamento são divididos em cada execução: 75% como subtreinamento (X_{train}) e 25% como validação (X_{valid}). Desse modo, os dados de teste não participam do processo de aprendizado. Os resultados de cada execução são obtidos pela média sobre cinco execuções distintas (variando a semente aleatória), uma vez que o *framework* de otimização é uma técnica não determinística.

Em relação aos parâmetros, $kAOG$ não requer seleção de parâmetro e kNN é otimizado sobre o seguinte conjunto $k \in \{1, 2, \dots, 30\}$. O *framework* proposto, o qual é chamado BIO (Otimização Bioinspirada) a partir daqui, emprega heurística de mapeamento *MapAll* e possui três parâmetros a serem definidos, todos eles relacionados ao algoritmo SL-PSO: o tamanho da população do enxame m , o número de iterações gen e o número máximo de possíveis ligações q . Usa-se $m = gen = 100$ e $q = 3$ nos resultados providos aqui.

A Tabela 14 registra os resultados obtidos usando dois métodos de formação da rede e o *framework* de otimização sobre cinco bases de dados com níveis distintos de mistura entre as classes. A partir da tabela, é possível observar que o *framework* de otimização é comparado favoravelmente em relação à kNN e $kAOG$, quando considerando bases de dados com maior nível de mistura das classes. Apesar de $kAOG$ desempenhar bem nas duas primeiras bases de dados, seu desempenho fica pior nas seguintes, indicando que a seleção de parâmetro desempenha um

Figura 43 – Bases de dados artificiais geradas com níveis de mistura distintos entre as classes: (a) 10%; (b) 25%; (c) 50%; (d) 75%; (e) 90%.

papel importante em lidar com a complexidade daquelas bases de dados, tal como exemplificado por k NN o qual tem melhor desempenho, embora superado por BIO.

5.2.3 Resultados em bases de dados reais

Esta subsecção provê resultados experimentais, a fim de avaliar o desempenho do *framework* proposto sobre bases de dados reais. A Subsecção 5.2.3.1 apresenta uma visão geral sobre os experimentos; A Subsecção 5.2.3.2 mostra uma análise detalhada sobre a convergência do *framework* bem como sobre seu desempenho usando ambas heurísticas de mapeamento propostas neste capítulo; As Subsecções 5.2.3.3 e 5.2.3.4 analisam os resultados do *framework* de otimização em comparação com métodos de formação da rede direcionados e não direcionados, respectivamente. O desempenho preditivo do *framework* também é comparado com técnicas de

Tabela 14 – Resultados obtidos pelo *framework* proposto (BIO) sobre as cinco bases de dados mostradas pela Fig. 43, caracterizadas por diferentes níveis de mistura entre as classes. BIO é comparado com métodos de construção da rede bastante usados na literatura, *k*NN e *k*AOG.

Datasets	<i>k</i> AOG	<i>k</i> NN	BIO
TwoMoons-10	100.0 $\pm$ 0.00	99.50 $\pm$ 0.15	99.80 $\pm$ 0.06
TwoMoons-25	88.00 $\pm$ 7.14	89.00 $\pm$ 6.24	89.80 $\pm$ 6.79
TwoMoons-50	74.50 $\pm$ 11.06	81.50 $\pm$ 9.23	80.70 $\pm$ 7.99
TwoMoons-75	71.50 $\pm$ 6.73	74.50 $\pm$ 8.50	77.20 $\pm$ 6.42
TwoMoons-90	67.80 $\pm$ 11.00	70.50 $\pm$ 12.54	72.20 $\pm$ 11.04

classificação tradicionais bastante utilizadas na Subseção 5.2.3.5. Finalmente, a Subseção 5.2.3.6 exibe detalhes sobre o processo de otimização bem como um exemplo de erro de classificação para técnicas tradicionais o qual é corretamente abordado pelo *framework* proposto.

5.2.3.1 Experimentos

A seguir, são fornecidos detalhes sobre as bases de dados nas quais o *framework* de otimização é aplicado, bem como sobre a condução dos experimentos.

Bases de dados A Tabela 15 traz uma breve descrição em termos do número de instâncias, atributos e classes das bases de dados reais consideradas. A seleção foi feita de maneira a envolver diversidade sobre domínios de dados, atributos e classes. Uma descrição detalhada sobre as bases de dados podem ser encontradas em (HASLINGER *et al.*, 2004; FRANK; ASUNCION, 2010). Como uma preparação dos dados, bases de dados, nas quais os atributos preditivos não estão em escala similar, foram padronizados de modo a apresentar média zero e desvio padrão unitário, do contrário, cada item de dado é normalizado pela norma l_2 . Eventuais atributos categóricos foram tratados como numéricos e a distância Euclideana foi usada em todos os experimentos como medida de dissimilaridade.

Experimentos Os experimentos foram executados exatamente da mesma maneira como apresentado na Subseção 5.2.2, i.e., usando o processo de validação cruzada dez vezes estratificada com aquela mesma porcentagem de itens de dados divididos em subtreinamento e validação.

Parâmetros do *framework* Nos experimentos, BIO emprega valores de m e gen iguais aos da Subseção 5.2.2, e q é otimizado sobre o conjunto $q \in \{3, 4, 5\}$, oferecendo uma boa relação entre desempenho preditivo e complexidade do espaço de busca.

Teste estatístico De modo a analisar estatisticamente os resultados obtidos nos experimentos, um teste estatístico que compara duas técnicas sobre várias bases de dados foi adotado, o teste de Postos Sinalizados de Wilcoxon. O teste é uma alternativa não paramétrica ao teste t pareado, quando suas suposições, tal como distribuição normal, não podem ser garantidas.

Tabela 15 – Descrição das bases de dados reais em termos do número de instâncias (*#Inst.*), atributos (*#Attr.*) e classes (*#Classes*).

Name	#Inst.	#Attr.	#Classes
Glass	214	9	7
Hayes-roth	160	4	3
Iris	150	4	3
Leukemia	100	50	2
Red-Wine	1599	11	6
Seed	210	7	3
Sonar	208	60	2
Teaching	151	5	3
Vehicle	846	18	4
Vowel	990	13	11

Basicamente, o teste de Wilcoxon estabelece um ranque a partir das diferenças de desempenho entre duas técnicas em cada base de dados, ignorando os sinais, e compara os ranques para as diferenças positivas e negativas (DEMŠAR, 2006). Um nível de confiança de 95% ($\alpha = 0.05$) foi considerado nos testes realizados.

5.2.3.2 Heurísticas de mapeamento e análises de convergência

Esta subseção analisa, em detalhes, a convergência do *framework* de otimização, considerando as duas heurísticas de mapeamento, *MapLab* e *MapAll*.

Análise de convergência As Tabelas 16 e 17 mostram os resultados obtidos por BIO durante a etapa de otimização (treinamento), usando as heurísticas *MapLab* e *MapAll*, respectivamente. Naquelas tabelas, a acurácia média e o desvio padrão são apresentados para cada base de dados, por considerar uma, dez e cem iterações, respectivamente. Em ambas tabelas, é possível observar uma expressiva melhoria em termos de desempenho preditivo em quase todas as bases de dados. A base de dados “Iris” é uma exceção, pois o desempenho preditivo máximo é alcançado na primeira iteração. Neste caso, não há espaço para o processo de otimização.

Resultados detalhados sobre o processo de convergência também são providos. A Fig. 44 mostra a acurácia média obtida por BIO, durante a etapa de otimização, usando ambas heurísticas de mapeamento. Pela figura, é possível concluir que uma população de cem partículas através de cem iterações é normalmente suficiente para assegurar uma boa convergência para o processo de otimização. Além disso, também é possível ver que o desempenho com *MapAll* é um pouco melhor que com *MapLab* na etapa de treinamento.

Resultados preditivos Apesar dos resultados apresentados na etapa de otimização, no qual as duas heurísticas de mapeamento desempenham bem em termos de convergência, a acurácia preditiva do *framework* de otimização estrutural precisa ser avaliada pelos resultados

Tabela 16 – Análise de convergência usando a heurística de mapeamento *MapLab* na etapa de otimização. Os resultados representam a acurácia média obtida mediante a função de qualidade f para uma, dez e cem iterações, respectivamente.

Base de dados	1gen	10gen	100gen
Glass	57.67 ± 3.63	66.15 ± 2.88	83.11 ± 1.88
Hayes-roth	68.97 ± 2.76	77.14 ± 2.81	84.70 ± 2.29
Iris	100.0 ± 0.00	-	-
Leukemia	91.52 ± 1.82	93.86 ± 1.47	95.15 ± 1.37
Red-Wine	35.16 ± 2.43	45.86 ± 1.54	63.40 ± 1.26
Seed	99.79 ± 0.63	99.79 ± 0.63	99.79 ± 0.63
Sonar	74.85 ± 3.75	79.66 ± 3.87	85.96 ± 2.93
Teaching	76.52 ± 4.44	82.56 ± 3.92	91.44 ± 4.20
Vehicle	70.62 ± 0.96	73.42 ± 0.63	83.47 ± 0.66
Vowel	52.24 ± 3.49	57.39 ± 3.74	68.77 ± 5.25

Tabela 17 – Análise de convergência usando a heurística de mapeamento *MapAll* na etapa de otimização. Os resultados representam a acurácia média obtida por meio da função de qualidade f para uma, dez e cem iterações, respectivamente.

Base de dados	1gen	10gen	100gen
Glass	72.12 ± 3.04	77.45 ± 3.38	82.22 ± 2.48
Hayes-roth	69.35 ± 3.47	77.41 ± 2.76	79.62 ± 2.91
Iris	100.0 ± 0.00	-	-
Leukemia	95.07 ± 0.81	96.02 ± 1.04	96.11 ± 1.04
Red-Wine	56.30 ± 1.25	60.49 ± 1.14	71.91 ± 0.83
Seed	99.79 ± 0.63	99.79 ± 0.63	99.79 ± 0.63
Sonar	75.57 ± 3.88	81.62 ± 2.44	88.60 ± 1.94
Teaching	80.05 ± 4.49	86.89 ± 5.24	91.05 ± 5.13
Vehicle	72.77 ± 0.72	76.22 ± 1.24	85.70 ± 1.25
Vowel	55.86 ± 2.39	62.74 ± 2.04	72.34 ± 3.46

Figura 44 – Análises da convergência na etapa de otimização do *framework* proposto usando as heurísticas de mapeamento *MapLab* e *MapAll*. Cada subfigura exibe a acurácia média obtida por meio da função de qualidade f em função do número de iterações. Os resultados mostram que cem partículas e cem iterações, geralmente, são suficientes para assegurar a convergência de ambas heurísticas. Em termos de desempenho preditivo, a heurística de mapeamento *MapAll* é um pouco melhor.

na etapa de teste, onde itens de dados completamente desconhecidos precisam ser rotulados a partir da exploração dos atributos da rede otimizada. A Tabela 18 registra uma comparação entre os resultados obtidos por BIO na etapa de teste usando ambas heurísticas de mapeamento. Os resultados na tabela revelam que o desempenho de *MapAll* é, estatisticamente, melhor que *MapLab*. Tais resultados são interessantes e fornecem evidências sobre a importância em considerar a complexidade da configuração das classes quando lidando com bases de dados reais. Nos experimentos computacionais demonstrados então, BIO é sempre equipado com a heurística de mapeamento *MapAll*.

Tabela 18 – Acurácia preditiva obtida pelo *framework* proposto quando usando as heurísticas de mapeamento *MapLab* e *MapAll* na etapa de teste.

Base de dados	BIO (<i>MapLab</i>)	BIO (<i>MapAll</i>)
Glass	60.82 ± 9.07	64.55 ± 7.13
Hayes-roth	55.34 ± 15.17	58.99 ± 11.81
Iris	98.67 ± 4.00	98.67 ± 4.00
Leukemia	84.22 ± 8.59	86.47 ± 9.69
Red-Wine	49.10 ± 3.24	53.89 ± 3.86
Seed	91.05 ± 8.50	90.86 ± 7.30
Sonar	60.88 ± 12.68	63.06 ± 11.85
Teaching	56.11 ± 16.13	55.52 ± 17.40
Vehicle	65.79 ± 3.22	67.80 ± 3.52
Vowel	58.36 ± 6.78	62.87 ± 6.99

5.2.3.3 Comparação com métodos de formação de dígrafos

Esta subseção apresenta a comparação do desempenho de BIO em relação a métodos de construção da rede bem conhecidos, como *kNN* e *kAOG*. As técnicas também são comparadas com um *baseline*, denominado RM (abreviação para rede mapeada aleatoriamente), o qual gera um conjunto de redes aleatoriamente a partir de uma heurística de mapeamento. Os três métodos considerados formam dígrafos tal como o *framework* de otimização, e a mesma função de qualidade f é usada sobre todas as redes.

Parâmetros A seguir, são dados detalhes sobre os experimentos e parâmetros de cada técnica:

- *kNN* possui um único parâmetro, o qual está associado com o número de vizinho mais próximos a considerar, quando conectando um dado vértice. Nos experimentos computacionais aqui, k é otimizado sobre o seguinte conjunto $k \in \{1, 2, \dots, 30\}$.
- *kAOG* é um método de formação da rede não paramétrico. Relembre-se que ele constrói o grafo por meio da otimização da medida de *pureza*.

- *RM* (grafo mapeado aleatoriamente) gera um conjunto de redes cujas conexões são determinadas aleatoriamente a partir da heurística de mapeamento *MapAll*. Enquanto *BIO* otimiza a rede no decorrer das iterações, *RM* estabelece as conexões aleatoriamente sem qualquer processo de atualização. *RM* possui dois parâmetros: o número de redes a serem geradas R e o número máximo de possíveis ligações q , os quais foram definidos como $R = 100$ e $q = 3$. Como esse é um método não determinístico, os resultados de cada execução são calculados pela média sobre cinco execuções distintas (variando a semente geradora de números aleatórios).

Resultados A acurácia média obtida por f , usando cada método de formação da rede, é exibida na Tabela 19. Pela tabela, é possível observar que o processo de otimização conduzido em *BIO* é capaz de obter uma melhoria considerável em termos de desempenho preditivo para algumas bases de dados, tais como “Sonar” e “Teaching”, quando comparado com os outros métodos. Os resultados do teste estatístico de Wilcoxon, exibidos na mesma tabela, revelam que não há diferença estatística entre *kNN*, *kAOG* e *RM* e que todos eles são superados pelo *framework* de otimização bioinspirado.

Tabela 19 – Acurácia preditiva do *framework* de otimização (*BIO*) em comparação métodos de formação da rede bem conhecidos (*kNN* e *kAOG*) e outro método estocástico (*RM*). As últimas linhas denotam os resultados do teste estatístico de Postos Sinalizados de Wilcoxon: (i) os símbolos “>” e “<” indicam uma rejeição da hipótese nula, onde o algoritmo correspondente à coluna é, respectivamente, melhor ou pior que o algoritmo correspondente à linha; (ii) e o símbolo “~” sugere que ambos os algoritmos desempenham igualmente bem, i.e., falha na rejeição da hipótese nula para um nível de confiança de 95%.

Bases de dados	<i>BIO</i>	<i>kNN</i>	<i>kAOG</i>	<i>RM</i>
Glass	64.55 ± 7.13	61.28 ± 11.27	64.55 ± 11.34	63.07 ± 10.06
Hayes-roth	58.99 ± 11.81	54.24 ± 13.12	53.54 ± 13.43	56.79 ± 11.88
Iris	98.67 ± 4.00	98.67 ± 4.00	98.67 ± 4.00	95.33 ± 5.21
Leukemia	86.47 ± 9.69	83.85 ± 11.54	88.07 ± 7.29	80.05 ± 8.70
Red-Wine	53.89 ± 3.86	53.43 ± 3.93	50.98 ± 2.44	48.41 ± 3.07
Seed	90.86 ± 7.30	90.00 ± 9.15	85.71 ± 12.05	90.00 ± 9.39
Sonar	63.06 ± 11.85	57.33 ± 15.72	59.16 ± 14.06	52.29 ± 12.55
Teaching	55.52 ± 17.40	46.10 ± 11.07	47.98 ± 14.16	51.15 ± 17.89
Vehicle	67.80 ± 3.52	66.20 ± 3.16	66.42 ± 3.97	69.00 ± 4.57
Vowel	62.87 ± 6.99	56.87 ± 11.36	59.70 ± 10.85	60.40 ± 9.57
<i>kNN</i>	>	-	-	-
<i>kAOG</i>	>	~	-	-
<i>RM</i>	>	~	~	-

5.2.3.4 Comparação com métodos de formação de rede não direcionada

Esta subseção compara o desempenho preditivo de BIO com métodos de formação de rede não direcionadas, os quais incluem o k NN simétrico e o k NN mútuo. Ambos os métodos são amplamente utilizados em tarefas do aprendizado não supervisionado e semissupervisionado, tais como agrupamento de dados e propagação de rótulos.

Parâmetros Como, na rede k NN, k é otimizado sobre o seguinte conjunto $k \in \{1, 2, \dots, 30\}$ em Sk NN e Mk NN.

Resultados A Tabela 20 apresenta o desempenho preditivo obtido pelos métodos de formação de rede em termos da acurácia de classificação média. Novamente, os resultados dos experimentos revelam que a rede otimizada gerada por BIO é capaz de obter uma melhoria considerável para algumas bases de dados, tais como “Hayes-roth”, “Sonar” e “Vowel”, quando comparada aos métodos de construção Sk NN e Mk NN. Os resultados dos testes estatísticos exibidos na mesma tabela indicam que não há diferença entre Sk NN e Mk NN, e que ambos são superados por BIO.

Tabela 20 – Acurácia preditiva do *framework* de otimização (BIO) em comparação com métodos de formação de rede bastante utilizados na literatura (Sk NN e Mk NN). Consulte-se a legenda da Tabela 19 para uma explicação detalhada sobre os resultados dos testes estatísticos.

Bases de dados	BIO	Sk NN	Mk NN
Glass	64.55 ± 7.13	53.76 ± 14.60	64.85 ± 10.49
Hayes-roth	58.99 ± 11.81	53.26 ± 12.71	50.96 ± 15.55
Iris	98.67 ± 4.00	98.00 ± 4.27	96.00 ± 6.11
Leukemia	86.47 ± 9.69	81.94 ± 10.92	86.07 ± 10.07
Red-Wine	53.89 ± 3.86	51.82 ± 5.89	52.80 ± 3.59
Seed	90.86 ± 7.30	88.09 ± 8.04	89.52 ± 9.94
Sonar	63.06 ± 11.85	50.45 ± 9.06	57.07 ± 12.50
Teaching	55.52 ± 17.40	53.37 ± 10.73	55.32 ± 12.92
Vehicle	67.80 ± 3.52	67.38 ± 3.51	69.02 ± 4.25
Vowel	62.87 ± 6.99	55.66 ± 7.98	55.56 ± 10.90
Sk NN	>	-	-
Mk NN	>	~	-

5.2.3.5 Comparação com classificadores tradicionais

Esta subseção compara o desempenho de BIO com técnicas de classificação tradicionais, tais como classificador k vizinhos mais próximos (k -NN), classificador ingênuo de Bayes (NB) e máquina de vetores de suporte (SVM).

Parâmetros A seguir, são fornecidos detalhes sobre os parâmetros de cada classificador:

Tabela 21 – Acurácia preditiva do *framework* de otimização (BIO) em comparação com técnicas de classificação tradicionais bem conhecidas (k -NN, NB e SVM). Uma explicação detalhada sobre os resultados dos testes estatísticos é fornecida na legenda da Tabela 19.

Bases de dados	BIO	k -NN	NB	SVM
Glass	64.55 ± 7.13	61.86 ± 9.40	42.53 ± 12.23	61.59 ± 13.20
Hayes-roth	58.99 ± 11.81	57.23 ± 9.93	45.94 ± 10.57	57.91 ± 12.67
Iris	98.67 ± 4.00	97.33 ± 4.42	96.66 ± 4.47	95.33 ± 6.00
Leukemia	86.47 ± 9.69	87.16 ± 10.58	89.07 ± 9.27	88.98 ± 10.45
Red-Wine	53.89 ± 3.86	52.69 ± 6.55	51.30 ± 5.68	57.56 ± 6.42
Seed	90.86 ± 7.30	89.52 ± 8.19	88.09 ± 10.91	89.52 ± 11.02
Sonar	63.06 ± 11.85	51.14 ± 14.12	58.59 ± 21.37	54.94 ± 10.97
Teaching	55.52 ± 17.40	51.41 ± 14.51	48.38 ± 9.74	49.54 ± 13.39
Vehicle	67.80 ± 3.52	69.72 ± 5.17	49.20 ± 4.63	82.86 ± 1.59
Vowel	62.87 ± 6.99	52.93 ± 9.33	56.36 ± 7.47	72.12 ± 11.55
k -NN	>	-	-	-
NB	>	~	-	-
SVM	~	~	~	-

- k -NN possui um único parâmetro k , o qual é otimizado sobre o seguinte conjunto $k \in \{1, 2, \dots, 30\}$.
- NB é uma técnica de classificação não paramétrica. Assume-se que a probabilidade dos atributos é Gaussiana nos experimentos computacionais realizados aqui.
- SVM possui muitos parâmetros para serem ajustados. Nos experimentos, o espaço de busca dos parâmetros foi reduzido por considerar a função de base radial como *kernel*. O critério de parada utilizado foi o *Karush-Kuhn-Tucker* com violação inferior a 10^{-3} . Os parâmetros do *kernel* e de custo são otimizados sobre os seguintes conjuntos $\gamma \in \{2^4, 2^3, \dots, 2^{-10}\}$ e $C \in \{2^{12}, 2^{11}, \dots, 2^{-2}\}$, respectivamente.

Resultados A Tabela 21 apresenta a acurácia média obtida para cada técnica. De modo a analisar os resultados, o teste de Wilcoxon foi realizado para comparar cada par de algoritmos. Os resultados do teste estatístico, os quais são exibidos na mesma tabela, revelam que não há diferença estatística entre k -NN, NB e SVM, e que os classificadores k -NN e NB são superados por BIO.

5.2.3.6 Análises das redes geradas

Esta subseção mostra algumas análises das redes obtidas pelo *framework* proposto. A Fig. 45 exibe uma comparação entre redes formadas pelo método k NN e BIO. As redes são construídas sobre a base de dados “Teaching”, considerando a primeira execução nos experimentos. Como

Figura 45 – Comparação entre redes formadas por diferentes métodos: (a) rede obtida por kNN ; (b) rede obtida por BIO considerando uma iteração; (c) rede obtida por BIO ao término do processo de otimização (após cem iterações). As informações extraídas a partir de um componente cujos vértices são os mesmos nas três redes enfatizam uma organização estrutural bastante distinta. Por exemplo, nas redes geradas em kNN , BIO-1gen e BIO-optimized, os nós mais importantes no componente analisado registram pontuação 0.047 (id. 33), 0.014 (id. 39) e 0.066 (id. 30), respectivamente.

o número de atributos é maior que três, o plano Fruchterman-Reingold (FRUCHTERMAN; REINGOLD, 1991) é usado para gerar os grafos. A rede k NN é apresentada pela Fig. 45a, e as soluções obtidas por BIO, na primeira iteração e no fim do processo de otimização, são exibidas pelas Figs. 45b e 45c, respectivamente. Nas redes da figura, cada item de dado é representado por um nó e cada cor denota uma classe. Como apontado anteriormente, a rede otimizada apresenta o melhor desempenho em termos de acurácia preditiva. Como meio de oferecer uma análise mais detalhada, informações foram extraídas de um componente composto pelos mesmos nós nas três redes. Além disso, os nós são ilustrados em diferentes tamanhos de acordo com sua *importância*, e um número de identificação (id.) é exibido para cada um deles. Nas redes k NN, BIO-1gen e BIO-otimizada, os nós mais importantes do componente analisado possuem pontuação 0.047 (id. 33), 0.014 (id. 39) e 0.066 (id. 30), respectivamente.

Outra análise é ilustrada pela rede otimizada mostrada pela Fig. 46. A figura expõe a rede obtida para a base de dados “Sonar” após o processo de otimização. Itens de dados, na forma de círculo e quadrado, representam as duas classes existentes. Os itens de dados contornados por pontilhados denotam itens de teste particulares, os quais são erroneamente preditos pelos classificadores tradicionais (k -NN, NB e SVM). Por outro lado, o *framework* proposto é capaz de atribuir o rótulo correto para cada um deles, uma vez que a classificação final é dada por examinar as estruturas físicas e topológicas dos dados através da rede otimizada.

Figura 46 – Rede otimizada obtida para a base de dados “Sonar”. Itens de dados no formato de círculos e quadrados representam as duas classes de dados existentes na base. Itens de teste representados circulos por pontilhados denotam casos de erro para as três técnicas tradicionais.

5.3 Aplicação do framework para classificação de alto nível

Nesta seção, o framework para otimização estrutural de redes é investigado no contexto da classificação de alto nível (SILVA; ZHAO, 2012a). Especificamente, o framework de otimização bioinspirada, apresentado a seguir, combina classificação de baixo e alto nível. Dele, é esperado a construção da rede e a seleção de parâmetros do modelo híbrido enquanto conduz a otimização de uma função de qualidade denotada pela acurácia preditiva. A fim de analisar o desempenho da nova abordagem, experimentos computacionais são realizados sobre dois problemas reais, o reconhecimento de dígitos manuscritos e a anotação de papéis semânticos.

5.3.1 Função de qualidade e representação adicional

A função de qualidade f calcula a acurácia preditiva obtida pelo modelo híbrido descrito pela Eq. (2.37). Especificamente para a etapa de otimização, também foi projetada algumas estratégias de otimização estrutural baseadas em medidas de redes, como segue:

$$f^{(\cdot)} = (1 - \alpha)f + \alpha f^{(\cdot)}, \quad (5.11)$$

onde $\alpha = 0.001$ provê um viés bastante suave na função de qualidade, e $f^{(\cdot)}$ pode ser:

- f^{AC+} , quando se busca-se aumentar a proximidade média enquanto maximiza o desempenho preditivo;
- f^{r+} , onde se busca aumentar os padrões de assortatividade enquanto maximiza a acurácia preditiva.

Em processos de otimização baseado em população, é esperado que exista alguma informação do gradiente que guie a convergência das soluções candidatas em direção ao ótimo global. Dessa forma, a Eq. (5.11) fornece um viés suave para algumas características da rede, tais como os padrões de assortatividade quando usando f^{r+} .

Em termos de representação, além do grafo $\mathcal{G}_i$, uma partícula $\mathcal{P}_i$ é também composta pelos parâmetros da técnica de classificação híbrida, denotados por $\mathcal{R}_i$:

$$\mathcal{R}_i = \{\varepsilon, \delta, \lambda\}, \quad (5.12)$$

onde:

- ε é a distância usada para estabelecer conexões temporárias de um novo item de dado na rede. Na representação proposta, ε é otimizado entre o menor e o maior valor de distância em X_{train} ;

- δ denota a contribuição de cada medida empregada pelo classificador de alto nível. Ela considera um número Z de medidas de rede tal que $\sum_{h=1}^Z \delta(h) = 1$;
- $\lambda \in [0, 1]$ representa a contribuição de cada classificação, de baixo e alto nível, no modelo híbrido.

5.3.1.1 Medidas de redes complexas

As seguintes medidas de redes complexas são empregadas para produzir as associações de alto nível: proximidade média, assortatividade e coeficiente de agrupamento. Como a assortatividade e o coeficiente de agrupamento já foram empregados no Capítulo 3 (ver Subseção 3.1.2), apenas a variação da proximidade média é definida abaixo.

Proximidade média - $\Delta G_y^{(c)}(1)$: a medida de proximidade média é discutida na Subseção 2.1.2.4 e sua variação em relação à inserção de um item de teste y em uma classe c é definida por:

$$\Delta G_y^{(c)}(1) = \frac{|AC'^{(c)} - AC^{(c)}|}{\sum_{u \in U} |AC'^{(u)} - AC^{(u)}|}, \quad (5.13)$$

onde $AC^{(\alpha)}$ e $AC'^{(\alpha)}$ denotam a proximidade média obtida, respectivamente, antes e após a inserção de y na classe c .

5.3.2 Resultados experimentais

O desempenho da técnica proposta é analisado a seguir. A Subseção 5.3.2.1 apresenta um exemplo ilustrativo sobre como o *framework* proposto para classificação de alto nível funciona; A Subseção 5.3.2.2 discute resultados de experimentos sobre bases de dados reais, mostrando que BIO não apenas melhora o desempenho de um classificador estado-da-arte rigorosamente tunado, mas também alcança melhor resultado que um método de formação de rede bastante usado.

5.3.2.1 Exemplo ilustrativo

Considere-se uma demonstração da ideia geral do *framework* proposto mediante o exemplo ilustrativo mostrado na Fig. 29b. Relembre-se que, na figura, há itens de dados rotulados de duas classes e um novo item y cujo rótulo é desconhecido. Pela figura, é certo que as classes apresentam claro padrão de formação. Ao contrário de técnicas tradicionais que falham na classificação de y , medidas de redes complexas podem detectar a formação de padrão dos dados por analisar também sua estrutura topológica. A Tabela 22 exibe as probabilidades associadas a cada classe por um classificador tradicional $\mathcal{C}$ (algoritmo de regressão logística foi usado aqui), por um classificador baseado em medidas de redes complexas $\mathcal{H}$ e a combinação entre eles $\mathcal{M}$, após o processo de otimização conduzido pelo *framework*. Os parâmetros ajustados pelo *framework* são apresentados na Tabela 23.

Tabela 22 – Resultados de classificação obtidos pelo *framework* de otimização no exemplo ilustrativo.

Alg.	Classe Cinza	Classe Preto
$\mathcal{C}$ (Regressão logística)	0.093	0.907
$\mathcal{H}$ (Medidas de rede)	0.796	0.204
$\mathcal{M}$ (Combinação)	0.605	0.395

Tabela 23 – Parâmetros otimizados pelo *framework* no exemplo ilustrativo.

Parâmetro	Valor
ε	0.135
w_1	0.431
w_2	0.239
w_3	0.330
λ	0.729

A Fig. 47 mostra a inserção temporária de y nas classes da rede, quando a variação das medidas de redes são calculadas. Os resultados das medidas antes e após a inserção de y são exibidos na Tabela 24. Pela tabela, é possível observar que BIO ajusta as conexões e os parâmetros de modo a detectar o padrão.

Figura 47 – Inserção temporária do item de teste y nas classes de dados do exemplo ilustrativo.

5.3.2.2 Resultados em bases de dados reais

Dados A seguir, são fornecidos detalhes sobre as bases de dados reais, nas quais BIO é avaliado, bem como sobre a condução dos experimentos. A Tabela 29 traz uma breve descrição das bases em termos de número de itens de dados, atributos e classes. Como uma preparação

Tabela 24 – Classificação de alto nível por meio de medidas de redes complexas no exemplo ilustrativo. Valores calculados de proximidade, assortatividade e coeficiente de agrupamento para cada classe.

Classe	Medidas de redes complexas					
	Proximidade		Assortatividade		Coef. de agrupamento	
	$AC^{(\cdot)}$	$AC'^{(\cdot)}$	$r^{(\cdot)}$	$r'^{(\cdot)}$	$CC^{(\cdot)}$	$CC'^{(\cdot)}$
Cinza	0.115	0.114	0.064	0.031	0.036	0.035
Preto	0.015	0.080	-0.064	-0.136	0.296	0.438

dos dados, cada vetor de atributo foi normalizado para ter uma magnitude de um, e a distância Euclidiana foi usada em todos os experimentos como medida de dissimilaridade.

Tabela 25 – Breve descrição das bases de dados reais em termos do número de instâncias, número de atributos e número de classes.

Nome	#Instâncias	#Atributos	#Classes
AlphaDigits	1404	320	36
PBbr-do	397	2118	8

Experimentos e parâmetros Cada experimento foi realizado usando processo de validação cruzada dez vezes estratificada. Em cada execução, os dados de treinamento são divididos em 75% para subtreinamento (X_{train}) e 25% para validação (X_{valid}). Como mencionado anteriormente, BIO possui apenas três parâmetros para serem ajustados, os quais são relacionados ao algoritmo SL-PSO. Nos experimentos computacionais, foi usado $m = 100$, $gen = 100$ e $q = 5$.

Classificador de baixo nível Uma vez que o modelo híbrido é também composto por um classificador tradicional, o classificador de regressão logística (LR) (FAN *et al.*, 2008), também chamado de classificador de máxima entropia, foi selecionado por ser uma técnica estado-da-arte para a tarefa de anotação automática de papéis semânticos (BAZRAFSHAN; GILDEA, 2013). Os parâmetros de LR foram, rigorosamente, tunados usando o método de busca em grid sobre um elevado número de combinações: $p = \{l1, l2\}$ para a norma usada na penalização e $C = \{2^{-2}, 2^{-1}, \dots, 2^{12}\}$ para a força da regularização.

Reconhecimento de dígitos manuscritos O primeiro problema abordado é o reconhecimento de dígitos manuscritos (HDR), o qual é uma tarefa de classificação bem conhecida. A base de dados usada foi a *Binary Alphanum*, a qual está disponível *online*¹ e contém dígitos binários 20×16 , que vão de “0” a “9” e letras capitais de “A” a “Z”, com 39 imagens de cada classe. A Fig. 48 mostra algumas amostras de imagem da base de dados. Nos experimentos, cada imagem é representada como um vértice na rede.

¹ <<http://www.cs.nyu.edu/~roweis/data.html>>

Figura 48 – Imagens da base de dados “AlphaDigits”.

Anotação de papéis semânticos O segundo problema abordado é a anotação automática de papéis semânticos (SRL), o qual é a tarefa de, automaticamente, identificar e classificar os argumentos de um predador com papéis que indicam relações semânticas entre os argumentos, como quem fez o quê para quem, onde, quando e como (veja a Seção 2.4 para mais detalhes sobre SRL). Especificamente, o foco é a tarefa de classificação de argumentos usando o Prop-Bank.br (DURAN; ALUÍSIO, 2012). Para o experimento, foram selecionadas todas as sentenças relacionadas aos predicadores “fazer”, um dos verbos mais frequentes no corpus, e os atributos foram extraídos usando um conjunto de características da literatura² (GILDEA; JURAFSKY, 2002; ALVA-MANCHEGO; ROSA, 2012; PRADHAN; WARD; MARTIN, 2008). Como um passo de pré-processamento, classes de argumentos com menos de dez itens de dados foram excluídas. Os metadados da base de dados obtida, referida como “PBbr-do”, podem ser vistos na Tabela 29. Nos experimentos, cada argumento foi mapeado como um vértice na rede.

Resultados preditivos A Tabela 26 apresenta a acurácia média obtida para as duas bases reais usadas nos experimentos. Na tabela, $\mathcal{C}(LR)$ denota o classificador de regressão logística e $\mathcal{M}(\cdot)$ é a classificação combinada entre as associações providas por LR e aquelas providas por medidas de redes complexas, dada pela Eq. (2.38), onde f quer dizer que as redes são otimizadas pelo *framework* proposto; e kNN denota a rede obtida pelo método de construção kNN onde apenas os parâmetros do classificador de alto nível são otimizados pelo algoritmo bioinspirado. Os resultados na tabela revelam que (i) BIO é capaz de melhorar o desempenho de classificação alcançado por um algoritmo estado-da-arte, rigorosamente tunado; e (ii) na tarefa de SRL, a rede obtida por meio do processo de otimização provê melhor resultado que aquela obtida pela rede kNN , enquanto a diferença entre os resultados obtidos por f e kNN na tarefa de HDR é pequena. Ao contrário de “PBbr-do”, “AlphaDigits” é uma base de dados muito bem balanceada onde

² As características usadas foram: FirstForm+FirstPostag, FirstLemma, Head, HeadLemma, TopSequence, PostagSequence, PredLemma+PhraseType, LastForm+LastPostag, PredLemma+Path, FirstPostag, LeftHead, RightHead, VoicePosition, LeftHeadPostag, RightPhrase, e PredLemma.

a configuração das classes reflete menor complexidade. Por exemplo, o número de classes em “AlphaDigits” é quatro vezes maior que em “PBbr-do”, porém a diferença entre o desempenho das técnicas é em torno de 6%. Acredita-se que essa possa ser uma informação sobre quando usar a otimização estrutural e quando a rede kNN possa ser preferível.

Tabela 26 – Acurácia preditiva do classificador de regressão logística (LR), do *framework* com otimização da rede ($\mathcal{M}(f)$) e sem otimização da rede ($\mathcal{M}(kNN)$), com $k = q$.

Alg.	AlphaDigits	PBbr-do
LR	68.68 ± 5.04	74.80 ± 4.97
$\mathcal{M}(f)$	69.72 ± 4.66	76.82 ± 3.89
$\mathcal{M}(kNN)$	69.65 ± 4.74	75.50 ± 5.49

A Tabela 28 apresenta os resultados de outro experimento onde f^{r+} e f^{AC+} , os quais são definidos na Eq. (5.11), descrevem a otimização da rede combinando a acurácia preditiva com as medidas de assortatividade e proximidade média, respectivamente. Como resultado, é possível observar que a otimização dessas medidas, em conjunto com a acurácia preditiva, é superada pela função de qualidade, considerando apenas a acurácia preditiva. Isto fornece evidências de que ambas medidas, quando compõe a função de qualidade, levam o algoritmo a convergir para soluções ótimas locais com menor habilidade de generalização.

Tabela 27 – Acurácia preditiva do *framework* de otimização usando diferentes composições da função de qualidade.

Alg.	PBbr-do
$\mathcal{M}(f)$	76.82 ± 3.89
$\mathcal{M}(f^{r+})$	76.35 ± 5.27
$\mathcal{M}(f^{AC+})$	76.32 ± 5.04

Além da técnica LR, BIO também é avaliado com outro classificador tradicional, o classificador ingênuo de Bayes (NB). Os resultados são apresentados na Tabela 28, os quais revelam que a abordagem proposta é capaz de melhorar, consideravelmente, a classificação provida por NB.

Tabela 28 – Acurácia preditiva usando o classificador ingênuo de Bayes (NB) e o *framework* de otimização. A probabilidade dos atributos foi assumida ser Gaussiana nos experimentos.

Alg.	AlphaDigits	PBbr-do
NB	50.23 ± 4.40	56.20 ± 8.00
$\mathcal{M}$	55.58 ± 3.85	60.58 ± 7.53

5.4 Considerações finais

A ideia por trás das investigações conduzidas neste capítulo é simples. A hipótese investigada aqui sugere que redes obtidas mediante algum processo de otimização de suas conexões apresentam melhor desempenho do que aquelas geradas heurísticamente. De fato, os resultados experimentais revelam as vantagens da otimização estrutural para a classificação baseada em redes. Em comparações com métodos de construção de rede bastante usados, o *framework* de otimização, normalmente, gera redes que propiciam melhores resultados, tal como revelado pelos teste estatísticos. Em comparações com técnicas de classificação tradicionais, os resultados obtidos evidenciam um melhor desempenho preditivo por parte do *framework* de otimização estrutural, mesmo sobre algoritmos do estado-da-arte rigorosamente tunados. Em suma, todos estes resultados sugerem que a aplicação do *framework* para outras tarefas do aprendizado de máquinas, tais como redução de dimensionalidade, agrupamento de dados, detecção de anomalia, etc, podem trazer ganhos consideráveis em termos de desempenho.

DIFUSÃO DE PAPÉIS SEMÂNTICOS BASEADA EM GRAFOS

Este capítulo investiga a aplicação de técnicas baseadas em rede para a difusão de papéis semânticos, a qual nada mais é do que a tarefa de anotação de papéis semânticos no contexto semissupervisionado. Para lidar com a tarefa, um *framework*, que varia em função de diferentes métodos de construção da rede e estratégias de difusão dos rótulos, é analisado. Especificamente, o foco da investigação é a classificação de argumentos usando o PropBank-br. Além de ser um corpus de língua portuguesa, outra motivação importante para a escolha do PropBank-br é a pouca quantidade de dados anotados, a qual representa um cenário difícil para aprendizado de máquina em que técnicas baseadas em grafos podem ser avaliadas em bases de dados com distribuições muito arbitrárias e desbalanceadas.

6.1 Descrição do framework

Dado um conjunto de argumentos $\mathcal{X} = \{a_1, \dots, a_l, a_{l+1}, \dots, a_n\}$ e um conjunto de papéis semânticos $\mathcal{L} = 1, \dots, c$, os l primeiros argumentos possuem papéis semânticos conhecidos $\{y_1, \dots, y_l\} \in \mathcal{L}$, e os argumentos restantes ($u = n - l$) são desprovidos de tal informação. Uma vez que $l \ll u$, i.e., a grande maioria dos argumentos não possui papel semântico, o objetivo é predizê-los a partir dos argumentos já rotulados.

A fim de realizar a difusão de papéis semânticos, o seguinte *framework* baseado em grafo é adotado:

1. Construa um grafo não direcionado $\mathcal{G}$ a partir de $\mathcal{X}$;
2. Obtenha uma matriz ponderada $\mathbf{W}$ para as conexões em $\mathcal{G}$ usando uma medida de similaridade $\mathcal{K}$;

3. Calcule o Laplaciano normalizado do grafo $\mathbf{L}$ a partir de $\mathbf{W}$;
4. Forneça os papéis semânticos aos argumentos sem rótulos usando a matriz de saída $\mathbf{F}$ obtida a partir de um método de propagação de rótulos.

6.1.1 Construção da rede

Considere-se um grafo não direcionado $\mathcal{G} = (V, E)$ onde cada nó $v_i \in V$ representa um argumento $a_i \in \mathbf{X}$. Permita-se $\mathbf{D}$ ser uma matriz de distância onde $\mathbf{D}_{ij} = \delta(a_i, a_j)$ e $k\text{NN}_i$ ser o conjunto dos k vizinhos mais próximos de a_i , a matriz de adjacência $\mathbf{A}$ da rede $k\text{NN}$ é, então, obtida por meio da Eq. (2.25).

Uma vez que as conexões no grafo $k\text{NN}$ são direcionadas, i.e., a existência de conexão de v_i para v_j não implica o contrário, dois métodos são, frequentemente, usados para tornar a rede não direcionada: o $k\text{NN}$ simétrico ($Sk\text{NN}$) e $k\text{NN}$ mútuo ($Mk\text{NN}$), os quais são obtidos por meio das Eqs. (2.26) e (2.27), respectivamente.

Uma suposição do aprendizado semissupervisionado baseado em grafos é que os itens de dados não rotulados encontram-se em um subgrafo conectado, no qual existe, pelo menos, um item de dado rotulado. Como meio de garantir estritamente tal suposição, os métodos de formação de rede $Sk\text{NN}$ e $Mk\text{NN}$ também são avaliados combinados com a árvore geradora mínima (MST) do grafo. Permita-se $\mathcal{G}$ representar um grafo completamente conectado, cujos pesos das arestas são dados por uma matriz de distância $\mathbf{D}$. A MST de $\mathcal{G}$ é o subconjunto de arestas $E' \subset E$ o qual conecta todos os nós em V , minimizando o seguinte valor:

$$\sum_{i,j} \mathbf{D}_{ij}. \quad (6.1)$$

6.1.2 Geração da matriz ponderada e do Laplaciano do grafo

Apesar da possibilidade de gerar uma matriz ponderada completamente conectada, a etapa de construção do grafo é importante para promover esparsificação em $\mathbf{W}$, o que reduz consideravelmente a complexidade computacional do modelo e, normalmente, melhora o desempenho preditivo. Logo, uma matriz ponderada simétrica $\mathbf{W} \subset \mathbb{R}^{n \times n}$ pode ser obtida por:

$$\mathbf{W}_{ij} = \mathbf{A}_{ij} \mathcal{K}(x_i, x_j), \quad (6.2)$$

onde $\mathcal{K}(x_i, x_j)$ denota uma função de similaridade. A função de similaridade usada neste estudo é o *kernel* RBF:

$$\mathcal{K}(x_i, x_j) = \exp\left(-\frac{\delta(x_i, x_j)}{2\sigma^2}\right), \quad (6.3)$$

onde $\delta(\cdot)$ é uma função de distância baseada em vetor tal como a norma l_2 $\|x_i - x_j\|^2$, e σ é o parâmetro da largura de banda do *kernel*.

A partir da matriz ponderada $\mathbf{W}$, o Laplaciano normalizado $\mathbf{L}$ é obtido:

$$\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}, \quad (6.4)$$

onde $\mathbf{I}_n$ é a matriz identidade e $\mathbf{D}$ é a matriz diagonal a qual contém o grau dos vértices. Como a literatura (JOHNSON; ZHANG, 2007) sugere que o Laplaciano normalizado pode levar a melhores resultados do que a Laplaciano combinatorial ($\Delta = \mathbf{D} - \mathbf{W}$), $\mathbf{L}$ é usado na formulação dos algoritmos de difusão de rótulos apresentados a seguir.

6.1.3 Propagação de rótulos

O processo de difusão de rótulos é conduzido pelas seguintes técnicas baseadas em grafo: campos Gaussianos e funções harmônicas (GFHF) (ZHU; GHAMRANI; LAFFERTY, 2003) e consistência local e global (LGC) (ZHOU *et al.*, 2004). Dado o Laplaciano $\mathbf{L}$ e considere-se $\mathbf{Y} \in \mathbb{B}^{n \times c}$ ser uma matriz de rótulos onde $Y_{ij} = 1$, se e somente se x_i possui papel semântico $y_i = j$, GFHF e LGC retornam a matriz de saída $\mathbf{F}$ por propagação de rótulos.

Campos Gaussianos e funções harmônicas GFHF obtém a matriz de saída $\mathbf{F}$ como segue:

$$\mathbf{F}_{\mathcal{U}} = -\mathbf{L}_{\mathcal{U}\mathcal{U}}^{-1} \mathbf{L}_{\mathcal{U}\mathcal{L}} \mathbf{Y}_{\mathcal{L}}. \quad (6.5)$$

Consistência local e global LGC obtém a matriz de saída $\mathbf{F}$ como segue:

$$\mathbf{F}(t+1) = \alpha \mathbf{L} \mathbf{F}(t) + (1 - \alpha) \mathbf{Y}, \quad (6.6)$$

onde α define a quantidade relativa de informação a partir dos vizinhos e da informação de rótulos inicial.

6.2 Resultados experimentais

A seguir, são apresentados resultados experimentais usando os métodos de construção de grafos e os algoritmos de propagação de rótulos descritos anteriormente. O objetivo é avaliar o desempenho preditivo da difusão de rótulos na tarefa de APS. No estudo, cada experimento foi realizado de maneira transdutiva, partindo de diferentes números de argumentos, inicialmente, rotulados. Especificamente, o número de argumentos rotulados l é dependente do número de classes c existente na base de dados $\mathcal{X}$, i.e., $l = \eta c$ com $\eta \in \{1, 2, 3\}$. Para cada valor de l testado, 30 execuções são realizadas. Em cada execução, os l dados rotulados são obtidos, aleatoriamente, considerando todos os argumentos existentes na base de dados, porém assegurando que exista pelo menos um item rotulado por papel semântico. Finalmente, a taxa de erro médio sobre as execuções é usada para avaliar a qualidade do processo de difusão dos papéis semânticos.

6.2.1 Bases de dados

Como os papéis semânticos são atribuídos em função do verbo neste estudo, e a fim de obter um maior número de argumentos não rotulados, três dos verbos mais frequentes no PropBank-br, a saber “dar”, “fazer” e “dizer”, são usados nos experimentos computacionais. Para tanto, foram selecionadas todas as sentenças do córpus relacionadas a tais verbos. Os atributos de cada argumento foram extraídos usando um conjunto de características da literatura¹ (GILDEA; JURAFSKY, 2002; ALVA-MANCHEGO; ROSA, 2012; PRADHAN; WARD; MARTIN, 2008). Como um pré-processamento, classes de argumentos com menos que dez itens foram excluídas. A Tabela 29 apresenta uma breve descrição das bases de dados obtidas, denominadas aqui “PBbr-dar”, “PBbr-fazer” e “PBbr-dizer”, em termos de número de instâncias/argumentos e classes. Nos experimentos, cada argumento é mapeado como um vértice na rede. Como uma preparação dos dados, cada vetor de atributo foi normalizado para ter uma magnitude de um, e a distância Euclideana foi usada em todos os experimentos como medida de distância. De modo a evitar o problema da maldição da dimensionalidade, a técnica de análise de componente principal (PCA) é executada, a fim de reduzir o número de dimensões dos dados para 100 (valor definido empiricamente).

Tabela 29 – Breve descrição das bases de dados extraídas do PropBank-br em termos do número de argumentos e classes.

Nome	#Instâncias	#Classes
PBbr-dar	148	3
PBbr-fazer	397	8
PBbr-dizer	506	5

6.2.2 Parâmetros e baseline

Os seguintes parâmetros são empregados nos experimentos computacionais apresentadas aqui: os valores de k para formação da rede kNN são otimizados sobre o seguinte conjunto $k \in \{1, 2, \dots, 60\}$; a largura de banda σ no *kernel* Gaussiano $\mathcal{K}$ é definida como sugerido em (JEBARA; WANG; CHANG, 2009) por $\sigma = \bar{d}_k/3$, onde $\bar{d}_k$ é a distância Euclidiana média entre cada argumento e seus k -ésimos vizinhos mais próximos; e o parâmetro α em LGC é escolhido do seguinte conjunto $\{0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 1.0\}$. Como discutido em muitos trabalhos na literatura, o classificador 1NN é usado para fins de comparação.

¹ As características usadas foram: FirstForm+FirstPostag, FirstLemma, Head, HeadLemma, TopSequence, PostagSequence, PredLemma+PhraseType, LastForm+LastPostag, PredLemma+Path, FirstPostag, LeftHead, RightHead, VoicePosition, LeftHeadPostag, RightPhrase, e PredLemma.

6.2.3 Resultados e discussões

Os resultados obtidos nos experimentos são organizados em três grupos de experimentos de acordo com o número de pontos rotulados, considerando $\eta \in \{1, 2, 3\}$. Note-se que $\eta = 1$ é o experimento mais relevante para o contexto semissupervisionado da tarefa de SRL, uma vez que ele sugere que, aproximadamente, uma sentença por verbo é rotulada, enquanto $\eta > 1$ significa duas ou mais sentenças rotuladas por verbo.

O primeiro grupo de experimentos é conduzido usando um argumento rotulado por papel semântico ($\eta = 1$) em cada base de dados e seus resultados são apresentados na Tabela 30. A primeira linha da tabela denota o classificador 1NN (*baseline*), enquanto as outras se referem a cada combinação entre métodos de formação da rede e de propagação de rótulos. Pela tabela, é possível inferir: (i) LGC-SkNN registra o melhor desempenho preditivo nas bases de dados PBbr-dar e PBbr-fazer, enquanto GFHF-SkNN desempenha muito bem sobre PBbr-dizer; (ii) o desempenho preditivo do *baseline* está muito longe de ambos os métodos de propagação GFHF e LGC, quando eles são combinados com os métodos de construção do grafo SkNN e SkNN+MST.

Tabela 30 – Resultados comparativos em termos de erro médio e desvio padrão (em trinta execuções) usando um item rotulado por classe ($\eta = 1$). O melhor resultado para cada base de dados é destacado.

Alg.	PBbr-dar (3)	PBbr-fazer (8)	PBbr-dizer (5)
1NN	45.31 (11.37)	63.52 (9.80)	35.41 (14.97)
GFHF-SkNN	41.26 (11.20)	50.21 (4.50)	29.41 (19.84)
GFHF-SkNN+MST	41.54 (9.96)	50.21 (4.50)	30.95 (15.03)
GFHF-MkNN	43.72 (9.05)	50.75 (4.52)	34.55 (7.93)
GFHF-MkNN+MST	43.84 (8.70)	50.34 (4.53)	34.42 (7.16)
LGC-SkNN	37.65 (11.20)	47.02 (2.90)	30.31 (19.36)
LGC-SkNN+MST	37.65 (11.20)	47.78 (3.79)	30.82 (16.26)
LGC-MkNN	39.61 (9.69)	48.93 (3.03)	34.47 (7.96)
LGC-MkNN+MST	39.72 (10.38)	47.13 (3.30)	33.59 (4.01)

A Tabela 31 mostra o desempenho preditivo dos algoritmos no segundo grupo de experimentos onde $\eta = 2$. De acordo com a tabela: (i) LGC-* (os quatro métodos de construção do grafo desempenham de maneira similar) apresenta os melhores resultados nas bases de dados PBbr-dar e LGC-SkNN sobre PBbr-fazer; (ii) GFHF/LGC-SkNN+MST registra bom desempenho sobre PBbr-dizer; (iii) dentre os métodos de construção do grafo, MkNN possui o pior desempenho em todas as bases de dados.

Os resultados do terceiro grupo de experimentos, o qual considera $\eta = 3$, é exibido na Tabela 32. Nos resultados, é possível perceber: (i) como no grupo de experimentos anterior, LGC-* possui melhor desempenho preditivo sobre PBbr-dar, LGC-SkNN sobre PBbr-fazer e LGC-SkNN+MST sobre PBbr-dizer; (ii) o método MkNN registra o pior desempenho sobre as

Tabela 31 – Resultados comparativos em termos de erro médio e desvio padrão (em trinta execuções) usando $\eta = 2$ (o número de argumentos rotulados é mostrado ao lado do nome da base). O melhor resultado para cada base de dados é destacado.

Alg.	PBbr-dar (6)	PBbr-fazer (16)	PBbr-dizer (10)
1NN	32.98 (8.96)	47.23 (3.77)	22.65 (13.15)
GFHF-SkNN	31.78 (10.15)	44.28 (4.84)	23.47 (10.71)
GFHF-SkNN+MST	31.73 (9.58)	45.74 (4.09)	21.46 (12.36)
GFHF-MkNN	32.91 (9.21)	46.02 (4.28)	26.16 (7.77)
GFHF-MkNN+MST	32.14 (9.92)	45.82 (4.12)	23.56 (12.10)
LGC-SkNN	30.96 (9.43)	43.57 (4.18)	22.91 (10.83)
LGC-SkNN+MST	30.96 (9.43)	44.82 (4.73)	21.70 (14.49)
LGC-MkNN	30.96 (9.52)	46.63 (3.24)	26.04 (7.81)
LGC-MkNN+MST	30.89 (9.36)	44.98 (5.67)	22.00 (12.67)

três bases de dados novamente; (iii) com o aumento de número de argumentos, inicialmente, rotulados ($\eta > 1$), o classificador 1NN tem seu desempenho melhorado, porém ele ainda está distante do desempenho de LGC nas bases de dados PBbr-dar e PBbr-fazer.

Tabela 32 – Resultados comparativos em termos de erro médio e desvio padrão (em trinta execuções) usando $\eta = 3$ (o número de argumentos rotulados é mostrado ao lado do nome da base). O melhor resultado para cada base de dados é destacado.

Alg.	PBbr-dar (9)	PBbr-fazer (24)	PBbr-dizer (15)
1NN	33.76 (9.51)	43.22 (6.26)	15.48 (6.57)
GFHF-SkNN	29.50 (8.64)	42.48 (5.07)	17.05 (8.69)
GFHF-SkNN+MST	29.50 (8.64)	43.13 (4.28)	15.80 (8.26)
GFHF-MkNN	30.36 (8.71)	43.44 (4.24)	20.98 (7.79)
GFHF-MkNN+MST	29.86 (8.73)	43.34 (4.28)	15.84 (8.37)
LGC-SkNN	28.94 (8.36)	41.38 (5.07)	16.71 (8.69)
LGC-SkNN+MST	28.94 (8.36)	44.03 (5.09)	15.12 (7.35)
LGC-MkNN	29.18 (8.74)	45.29 (4.87)	19.81 (8.17)
LGC-MkNN+MST	29.02 (8.91)	43.31 (5.48)	16.00 (7.92)

A seguir, o desempenho preditivo de ambos métodos de propagação de rótulos GFHF e LGC é analisado em função dos métodos de construção de grafo e a variação de seus parâmetros k . Como LGC possui também o parâmetro α , emprega-se α de acordo com o melhor desempenho obtido pela técnica. Mais uma vez, é considerada a divisão dos experimentos nos três grupos de experimentos. As Figs. 49, 50 e 51 mostram as taxas de erro médio das combinações de algoritmos para $\eta = 1$, $\eta = 2$ e $\eta = 3$, respectivamente. Pela análise das figuras, é possível concluir que:

SkNN $\times$ MkNN: independentemente do número de argumentos, inicialmente, rotulados l e dos algoritmos de propagação de rótulos, MkNN possui o pior desempenho geral sobre as bases de dados PBbr-dar bem como sobre PBbr-dizer. SkNN também é melhor que MkNN sobre a base de dados PBbr-fazer salvo o caso quando $\eta = 1$.

MST $\times$ not MST: para $\eta > 1$, o grafo MST melhora o desempenho geral de ambos métodos de formação de rede SkNN e MkNN sobre as bases de dados PBbr-dar e PBbr-dizer.

PBbr-fazer: as menores taxas de erro médio obtidas para PBbr-fazer normalmente são para pequenos valores de k . Uma suposição é que isso acontece porque a base de dados é muito desbalanceada com 2 de 8 classes de papéis semânticos, contendo mais de 68% do número total de argumentos. Nesse sentido, LGC possui alguma dificuldade para diminuir sua taxa de erro médio com o aumento de k , enquanto GFHF não. Tal característica precisa ser mais investigada.

Figura 49 – Erro médio de (a) GFHF e (b) LGC sobre as bases de dados PBbr-dar (esquerda), PBbr-fazer (centro) e PBbr-dizer (direita) com um argumento rotulado por papel semântico.

Finalmente, a Fig. 52 mostra um comparativo em função dos métodos de propagação de rótulos. Nos resultados, SkNN, método de formação de rede que obteve melhores resultados nos experimentos anteriores, foi adotado para a construção do grafo. Apesar de GFHF dispensar o uso de seleção de parâmetros (σ é definido heurísticamente), seu desempenho preditivo é, normalmente, pior que do algoritmo LGC. Desse modo, a combinação de LGC com o método de formação de rede SkNN parece uma escolha mais adequada para a difusão de papéis semânticos no PropBank-br.

Figura 50 – Erro médio de (a) GFHF e (b) LGC sobre as bases de dados PBbr-dar (esquerda), PBbr-fazer (centro) e PBbr-dizer (direita) com 6, 16 e 10 argumentos rotulados, respectivamente.

Figura 51 – Erro médio de (a) GFHF e (b) LGC sobre as bases de dados PBbr-dar (esquerda), PBbr-fazer (centro) e PBbr-dizer (direita) com 9, 24 e 15 argumentos rotulados, respectivamente.

Figura 52 – Erro médio de GFHF e LGC nas bases de dados PBbr-dar (esquerda), PBbr-fazer (centro) e PBbr-dizer (direita) com diferentes números de argumentos previamente rotulados: (a) $\eta = 1$, (b) $\eta = 2$ e (c) $\eta = 3$.

6.3 Considerações finais

Neste capítulo, a anotação automática de papéis semânticos é investigada no contexto do aprendizado semissupervisionado baseado em grafos. Especificamente, quatro métodos de construção da rede e dois métodos de propagação de rótulos são usados em um *framework*, cujo desempenho é avaliado sobre a classificação de argumentos no corpus de língua Portuguesa denominado PropBank-br. Experimentos computacionais usando diferentes números de argumentos rotulados foram realizadas. Os resultados e análises gráficas indicam que os métodos k vizinhos mais próximos simétrico ($SkNN$) e consistência local e global (LGC) são uma combinação promissora para a difusão de papéis semânticos no PropBank-br. Além disso, os resultados mostram que a combinação dos métodos de formação da rede com árvores geradoras mínimas podem melhorar o desempenho geral das técnicas, à medida que o número de argumentos previamente rotulados aumenta.

CONCLUSÕES

Ante as características salientes da representação em rede, as quais permitem analisar informações estruturais, dinâmicas e funcionais dos dados, além da habilidade de detectar estruturas espaciais e topológicas intrínsecas às relações mapeadas pela rede, o principal objetivo das investigações da tese foi o projeto e desenvolvimento de métodos e heurísticas de aprendizado baseados em teorias de redes complexas para a classificação de dados. Como resultado, contribuições relevantes para as áreas de classificação, sistemas de aprendizado e redes complexas foram discutidas neste trabalho. A seguir, é apresentado um resumo das principais contribuições, oportunidades de futuras investigações e uma lista dos artigos gerados durante o desenvolvimento desta tese.

7.1 Principais contribuições

Primeiramente, esta tese demonstrou que é possível extrair associações de baixo e alto nível a partir de uma mesma rede. Consequentemente, a conformidade de padrão dos dados pôde ser verificada mediante uma técnica simples, caracterizada pela ausência de parâmetros nas etapas de formação da rede e na produção das predições de baixo e alto nível. Mesmo simples, a técnica foi capaz de prover novidades para o processo de classificação e evidenciou o aumento de desempenho preditivo oriundo da classificação de alto nível em algumas bases de dados reais. Ademais, outra contribuição importante refere-se à detecção dos padrões de formação dos dados sobre cada componente da rede ao invés de considerar todos os itens de dados de cada classe, a qual propicia maior sensibilidade para as associações de alto nível.

Em segundo lugar, a tese propôs um novo conceito de classificação baseado na importância dos dados pelo qual um novo item de dado recebe o rótulo da classe onde ele recebe maior valor de importância. O conceito de importância proposto é derivado da formulação do Page-Rank. Além disso, a técnica baseada em importância também é capaz de capturar propriedades espaciais e estruturais dos dados usando uma nova medida denominada *eficiência diferencial*

espaço-estrutural. Por considerar propriedades distintas dos dados, além dos atributos físicos, os resultados experimentais refletiram em bom desempenho preditivo mesmo em cenários com elevado nível de mistura entre as classes. Outros experimentos também revelaram a habilidade em capturar relações semânticas entre os dados e o baixo custo computacional da técnica em comparação com outras técnicas tradicionais bastante utilizadas. Ademais, a nova abordagem - classificando dados por *importância* - apresenta uma visão completamente diferente sobre o processo de classificação, a qual sugere, por sua vez, que a classificação de dados pode ser realizada de formas distintas e em diferentes níveis, por exemplo, divisão do espaço de dados como é feito por técnicas tradicionais; conformidade de padrão como é feito pelas técnicas de classificação de alto nível; classificação por *importância*, como mencionado neste trabalho; e outras a serem descobertas. Nesse sentido, outra contribuição importante desta tese é a generalização do conceito de classificação de dados.

Em terceiro lugar, esta tese provou a suposição inicial de que redes geradas a partir de um processo de otimização estrutural podem apresentar melhor desempenho que outras obtidas heurísticamente em aplicações do aprendizado baseado em grafos. Para tanto, foi projetado um *framework* de otimização bioinspirado onde as conexões da rede são atualizadas iterativamente enquanto uma função de qualidade orientada à tarefa desejada é otimizada. Dois tipos de tarefas foram examinadas nesta tese, a classificação baseada em importância e o *framework* híbrido para classificação de alto nível. Na classificação baseada em importância, os resultados de vários experimentos computacionais provaram estatisticamente que o processo de otimização é benéfico, em termos de desempenho preditivo, quando comparado aos métodos de construção da rede mais usados na literatura. No *framework* híbrido para classificação de alto nível, os resultados evidenciaram que a otimização das conexões é capaz de melhorar os resultados obtidos por um algoritmo estado-da-arte para a tarefa de anotação de papéis semânticos, o qual foi, rigorosamente, tunado. De fato, em ambos os casos, o processo de otimização estrutural teve contribuição relevante na melhoria do desempenho preditivo, o que sugere que a aplicação do *framework* para outras tarefas do aprendizado baseado em grafos, tais como redução de dimensionalidade, agrupamento de dados, detecção de anomalia, etc, podem trazer ganhos consideráveis em termos de desempenho.

Finalmente, a análise e comparação de diferentes técnicas baseadas em grafos para a tarefa de difusão de papéis semânticos, usando textos jornalísticos da língua Portuguesa, também foi realizada. Em um cenário caracterizado por distribuições arbitrárias e desbalanceamento de classes, diferentes métodos de construção da rede e estratégias de propagação de rótulos são comparados. Os resultados evidenciam a superioridade das técnicas baseadas em grafo sobre um *baseline* bastante usado, especialmente, quando o número de argumentos rotulados é pequeno. Entre as técnicas baseadas em grafos, a combinação entre a rede k vizinhos mais próximos simétrico e a estratégia de propagação consistência local e global se mostrou promissora para a tarefa.

7.2 Oportunidades de trabalhos futuros

A seguir, são apresentadas oportunidades de trabalhos futuros derivadas das contribuições desta tese.

1. Em relação à conformidade de padrão, as seguintes oportunidades de pesquisa merecem destaque:
 - A fim de detectar a conformidade de padrão nos dados, as variações das medidas de redes precisam ser calculadas. Nesse sentido, um problema em aberto é: quais medidas de rede utilizar? Nos experimentos apresentados nesta tese, quatro medidas foram avaliadas: proximidade, assortatividade, coeficiente de agrupamento e grau. Contudo a literatura de redes complexas possui outras medidas de rede capazes de caracterizar de maneira diferente as relações entre os dados, tais como modularidade, excentricidade, *Katz*, etc. Desse modo, um estudo sistemático apresentando de maneira detalhada as principais vantagens e limitações de um grande conjunto de medidas em tarefas do aprendizado baseado em grafos é de grande interesse para a comunidade científica.
 - Uma novidade da técnica híbrida é que a conformidade de padrão é verificada sobre os componentes da rede ao invés das classes inteiras. Se essa característica tem por vantagem tornar a variação nas medidas de redes mais sensíveis aos padrões expressos em cada componente, ela tem por desvantagem a dificuldade em combinar a classificação de alto nível com técnicas tradicionais, as quais são incapazes de fornecer predições associadas a cada componente. Nesse sentido, uma oportunidade de trabalho futuro é avaliar e desenvolver métodos que permitam tomar classificações por componentes também por meio de técnicas que não sejam baseadas em redes. Por exemplo, pode-se criar uma camada adicional de rótulos baseada nos componentes existentes na rede.
2. Em relação à técnica de classificação baseada em importância, sugerem-se os seguintes tópicos como continuação da pesquisa:
 - A medida de eficiência proposta nesta tese é capaz de considerar propriedades espaciais e estruturais dos dados. Contudo ela pode ser definida a partir de várias outras medidas de rede, por exemplo, considerando o menor caminho, o coeficiente de agrupamento, etc. Uma investigação sistemática nesse sentido pode melhorar ainda mais o desempenho preditivo da técnica, além de agregar novidades ao processo de classificação.
 - A técnica baseada em importância foi projetada para o aprendizado supervisionado indutivo, porém outra abordagem interessante consiste na sua extensão para o aprendizado transdutivo, onde em vez de calcular a importância de apenas um item de

dado por vez, a importância pode ser verificada, considerando a inserção temporária de várias instâncias na rede.

- Na classificação baseada em importância, a formulação do PageRank é explorada para produzir um conceito de classificação a partir da caracterização de importância de novos itens de dados nos componentes da rede. Contudo além do conceito de importância e de conformidade de padrão, é possível extrair outros conceitos da representação em redes para a classificação.
3. Em relação à otimização estrutural no aprendizado baseado em redes, as seguintes investigações são sugeridas:
- Nos estudos sobre otimização estrutural de redes apresentados na tese, a função de qualidade esteve relacionada à classificação baseada em importância e à técnica híbrida para classificação de alto nível. Contudo, pelos resultados obtidos, também é esperado que o *framework* proposto alcance bom desempenho em outras tarefas do aprendizado baseado em grafos, tais como redução de dimensionalidade, detecção de anomalia, agrupamento de dados, etc. Desse modo, investigações abordando novas aplicações são uma interessante oportunidade de pesquisa futura.
 - O *framework* de otimização estrutural é conduzido por uma técnica bioinspirada, recentemente, proposta para otimização em grande escala. Porém a avaliação de outras técnicas e heurísticas de otimização é bem vinda, pois pode trazer novos benefícios para o processo de construção iterativo da rede como, por exemplo, menor complexidade computacional.
 - Duas heurísticas de mapeamento foram propostas para o *framework* de otimização estrutural. Essas heurísticas têm por objetivo reduzir a complexidade do espaço de busca de modo que o *framework* possa ser eficiente em termos de complexidade computacional. Uma vez que as medidas propostas nesta tese são baseadas no método k vizinhos mais próximos, sugere-se como oportunidade futura o desenvolvimento de outras heurísticas de mapeamento baseadas em outros métodos, como, por exemplo, a vizinhança de raio ε .
 - Motivado pelo crescimento recente e pelo potencial de técnicas baseadas em grafos em problemas de aprendizado, sugere-se a investigação de métodos automáticos de projeção de técnicas baseadas em grafos como uma oportunidade futura bastante promissora. Dada a variedade de métodos de construção da rede, medidas de rede e técnicas baseada em redes, uma ferramenta capaz de gerar, automaticamente, diferentes algoritmos baseados em rede é bastante promissora, uma vez que tal ela poderia cobrir uma variedade muito grande de padrões.
4. Em relação à difusão de papéis semânticos, a seguinte oportunidade de pesquisa é destacada:

- Nesta tese, quatro métodos de formação da rede e duas estratégias de propagação de rótulo foram avaliadas na difusão de papéis semânticos. Como trabalho futuro, sugere-se a inclusão de outros métodos de formação da rede e propagação de rótulos nos experimentos. Além disso, é importante compará-los com outras abordagens para classificação semissupervisionada.

7.3 Lista de artigos gerados

As contribuições desta tese foram publicadas em artigos científicos de conferências e revistas. A seguir esses artigos são apresentados.

- CARNEIRO, M. G.; ZHAO, L. Improving semantic role labeling with complex network measures (submetido). In: **IEEE International Joint Conference on Neural Networks**. [S.l.: s.n.].
- CARNEIRO, M. G.; ZHAO, L.; JIN, Y. Bio-inspired structural optimization for network-based data classification (submetido). **IEEE Transactions on Cybernetics**.
- CARNEIRO, M. G.; ZHAO, L. Organizational data classification based on importance concept of complex networks (aceito condicionalmente: revisões simples). **IEEE Transactions on Neural Networks and Learning Systems**.
- CARNEIRO, M. G.; ZHAO, L.; ROSA, J. L. G. Graph-based semi-supervised learning for semantic role diffusion. In: **Symposium on Knowledge Discovery, Mining and Learning**. [S.l.: s.n.], 2016. p. 1–8.
- CARNEIRO, M. G.; ZHAO, L.; CHENG, R.; JIN, Y. Network structural optimization based on swarm intelligence for highlevel classification. In: **IEEE International Joint Conference on Neural Networks**. [S.l.: s.n.], 2016. p. 3737–3744.
- CUPERTINO, T. H.; ZHAO, L.; CARNEIRO, M. G. Network-based supervised data classification by using an heuristic of ease of access. **Neurocomputing**, Elsevier, v. 149, p. 86–92, 2015.
- CARNEIRO, M. G.; ROSA, J. L. G.; LOPES, A. A.; ZHAO, L. Network-based data classification: combining k-associated optimal graphs and high-level prediction. **Journal of the Brazilian Computer Society**, Springer, v. 20, n. 1, p. 1–14, 2014.
- CARNEIRO, M. G.; CUPERTINO, T. H.; ZHAO, L. K-associated optimal network for graph embedding dimensionality reduction. In: **IEEE International Joint Conference on Neural Networks**. [S.l.: s.n.], 2014. p. 1660–1666.

- CUPERTINO, T. H.; CARNEIRO, M. G.; ZHAO, L. Dimensionality reduction with the k-associated optimal graph applied to image classification. In: **IEEE International Conference on Imaging Systems and Techniques**. [S.l.: s.n.], 2013. p. 366–371.
- CARNEIRO, M. G.; ZHAO, L. High level classification totally based on complex networks. In: **IEEE BRICS Congress on Computational Intelligence**. [S.l.: s.n.], 2013. p. 507–514.
- CARNEIRO, M. G.; ROSA, J. L.; LOPES, A. A.; ZHAO, L. Classificação de alto nível utilizando grafo k-associados ótimo. In: **International Workshop on Web and Text Intelligence**. [S.l.: s.n.], 2012. p. 1–10.

O doutorando também escreveu e colaborou em artigos relacionados a outros tópicos durante o período da tese, especificamente no tópico de escalonamento usando autômatos celulares, onde os seguintes artigos foram gerados.

- CARVALHO, T. I.; CARNEIRO, M. G.; OLIVEIRA, G. M. B. A comparison of forecast dynamical behavior parameter and direct verification of lattice's configuration using a cellular automata model to task scheduling. In: **International Conference on Cellular Automata for Research and Industry**. [S.l.: s.n.], 2016. p. 258–268.
- CARVALHO, T. I.; CARNEIRO, M. G.; OLIVEIRA, G. M. B. A hybrid strategy to evolve cellular automata rules with a desired dynamical behavior applied to the task scheduling problem. In: **Brazilian Conference on Intelligent Systems**. [S.l.: s.n.], 2016. p. 1–6.
- CARNEIRO, M. G.; OLIVEIRA, G. M. B. Synchronous cellular automata-based scheduler initialized by heuristic and modeled by a pseudo-linear neighborhood. **Natural Computing**, Springer, v. 12, n. 3, p. 339–351, 2013.
- CARNEIRO, M. G.; OLIVEIRA, G. M. B. SCAS-IS: Knowledge extraction and reuse in multiprocessor task scheduling based on cellular automata. In: **Brazilian Symposium on Neural Networks**. [S.l.: s.n.], 2012. p. 142–147.
- CARNEIRO, M. G.; OLIVEIRA, G. M. B. Scas-h: Synchronous cellular automata-based scheduler with initialization heuristic to task scheduling. In: **Automata & JAC**. [S.l.: s.n.], 2012. p. 52–61.

REFERÊNCIAS

AGGARWAL, C. C. **Data Classification: Algorithms and Applications**. [S.l.]: Taylor & Francis, 2014. Citado na página 1.

ALBERT, R.; ALBERT, I.; NAKARADO, G. L. Structural vulnerability of the north american power grid. **Physical Review E**, American Physical Society, v. 69, n. 2, p. 025103, 2004. Citado na página 2.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, American Physical Society, v. 74, n. 1, p. 47–97, 2002. Citado 18 vezes nas páginas 2, 3 e 14.

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Internet: Diameter of the world-wide web. **Nature**, Nature Publishing Group, v. 401, n. 6749, p. 130–131, 1999. Citado na página 2.

ALCALÁ-FDEZ, J.; FERNÁNDEZ, A.; LUENGO, J.; DERRAC, J.; GARCÍA, S. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. **Multiple-Valued Logic and Soft Computing**, Old City Publishing, v. 17, n. 2-3, p. 255–287, 2011. Citado 3 vezes nas páginas 76 e 92.

ALPAYDIN, E. **Introduction to Machine Learning**. [S.l.]: The MIT Press, 2009. Citado 5 vezes nas páginas 34 e 38.

ALVA-MANCHEGO, F. E. **Anotação automática semissupervisionada de papéis semânticos para o português do Brasil**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013. Citado 7 vezes nas páginas 16, 58, 59 e 61.

ALVA-MANCHEGO, F. E.; ROSA, J. L. G. Semantic role labeling for brazilian portuguese: A benchmark. In: **Advances in Artificial Intelligence – IBERAMIA 2012**. [S.l.]: Springer, 2012. p. 481–490. Citado 13 vezes nas páginas 9, 61, 62, 126 e 132.

AMANCIO, M. A.; DURAN, M. S.; ALUÍSIO, S. M. Automatic question categorization: a new approach for text elaboration. **Procesamiento de Lenguaje Natural**, Sociedad Espanola para el Procesamiento del Lenguaje Natural, v. 46, p. 43–50, 2011. Citado na página 61.

ARAUJO, B.; ZHAO, L. Data heterogeneity consideration in semi-supervised learning. **Expert Systems with Applications**, Elsevier, v. 45, p. 234–247, 2016. Citado na página 43.

BARABASI, A.-L. **Linked: How everything is connected to everything else and what it means for Business, Science, and Everyday Life**. [S.l.]: Plume, 2003. Citado na página 2.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citado 6 vezes nas páginas 16, 25, 29 e 30.

BARABÁSI, A.-L.; ALBERT, R.; JEONG, H. Scale-free characteristics of random networks: the topology of the world-wide web. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 281, n. 1, p. 69–77, 2000. Citado na página 14.

BARABÁSI, A.-L.; JEONG, H.; NÉDA, Z.; RAVASZ, E.; SCHUBERT, A.; VICSEK, T. Evolution of the social network of scientific collaborations. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 311, n. 3, p. 590–614, 2002. Citado na página 2.

BARRAT, A.; BARTHÉLEMY, M.; VESPIGNANI, A. **Dynamical Processes on Complex Networks**. [S.l.]: Cambridge University Press, 2008. Citado na página 2.

BAZRAFSHAN, M.; GILDEA, D. Semantic roles for string to tree machine translation. In: **Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2013. v. 2, p. 419–423. Citado 3 vezes nas páginas 61 e 125.

BEARMAN, P. S.; MOODY, J.; STOVEL, K. Chains of affection: The structure of adolescent romantic and sexual networks¹. **American Journal of Sociology**, JSTOR, v. 110, n. 1, p. 44–91, 2004. Citado na página 2.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. **Neural Computation**, MIT Press, v. 15, n. 6, p. 1373–1396, 2003. Citado 8 vezes nas páginas 41 e 43.

BELKIN, M.; NIYOGI, P.; SINDHWANI, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. **Journal of Machine Learning Research**, JMLR.org, v. 7, p. 2399–2434, 2006. Citado na página 44.

BERTINI, J. R.; ZHAO, L.; MOTTA, R.; LOPES, A. A. A nonparametric classification method based on k-associated graphs. **Information Sciences**, Elsevier, v. 181, n. 24, p. 5435–5456, 2011. Citado 14 vezes nas páginas 16, 5, 41, 43, 44, 45, 47, 48 e 65.

BETHARD, S.; LU, Z.; MARTIN, J. H.; HUNTER, L. Semantic role labeling for protein transport predicates. **BMC Bioinformatics**, BioMed Central, v. 9, n. 1, p. 1–15, 2008. Citado na página 60.

BICK, E. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. [S.l.]: Aarhus University Press, 2000. Citado na página 57.

- BICK, E. Automatic semantic role annotation for portuguese. In: **Workshop on Information and Human Language Technology**. [S.l.: s.n.], 2007. p. 1713–1716. Citado na página 61.
- BISHOP, C. M. **Pattern recognition and machine learning**. [S.l.]: Springer, 2006. Citado 16 vezes nas páginas 1, 2, 34 e 38.
- BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: **ACM Annual Conference on Computational Learning Theory**. [S.l.: s.n.], 1998. p. 92–100. Citado 3 vezes nas páginas 7 e 40.
- BOCCALETTI, S.; IVANCHENKO, M.; LATORA, V.; PLUCHINO, A.; RAPISARDA, A. Detecting complex network modularity by dynamical clustering. **Physical Review Letters**, American Physical Society, v. 75, p. 045102, 2007. Citado 14 vezes nas páginas 3 e 14.
- BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4, p. 175–308, 2006. Citado 15 vezes nas páginas 3 e 14.
- BOLDI, P.; SANTINI, M.; VIGNA, S. Pagerank as a function of the damping factor. In: **International Conference on World Wide Web**. [S.l.: s.n.], 2005. p. 557–566. Citado na página 84.
- BOLLOBÁS, B. **Modern graph theory**. [S.l.]: Springer, 1998. Citado na página 15.
- BORNHOLDT, S.; SCHUSTER, H. G. (Ed.). **Handbook of Graphs and Networks: From the Genome to the Internet**. [S.l.]: John Wiley & Sons, Inc., 2003. Citado na página 2.
- BREIMAN, L. **Classification and regression trees**. [S.l.]: Chapman & Hall/CRC, 1984. Citado 5 vezes nas páginas 4 e 36.
- BREIMAN, L. Random forests. **Machine Learning**, Kluwer Academic Publishers, v. 45, n. 1, p. 5–32, 2001. Citado na página 4.
- BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer Networks and ISDN Systems**, Elsevier, v. 30, n. 1, p. 107–117, 1998. Citado 7 vezes nas páginas 3, 14 e 23.
- BULLMORE, E.; SPORNS, O. Complex brain networks: graph theoretical analysis of structural and functional systems. **Nature Reviews Neuroscience**, Nature Publishing Group, v. 10, n. 3, p. 186–198, 2009. Citado na página 2.
- CANCHO, R. F. i; SOLÉ, R. V. Optimization in complex networks. In: **Statistical Mechanics of Complex Networks**. [S.l.]: Springer, 2003. p. 114–126. Citado 7 vezes nas páginas 16, 9, 32 e 33.

CARNEIRO, M. G.; CUPERTINO, T. H.; ZHAO, L. K-associated optimal network for graph embedding dimensionality reduction. In: **IEEE International Joint Conference on Neural Networks**. [S.l.: s.n.], 2014. p. 1660–1666. Citado 5 vezes nas páginas 45, 51 e 52.

CARNEIRO, M. G.; OLIVEIRA, G. M. B. Scas-h: Synchronous cellular automata-based scheduler with initialization heuristic to task scheduling. In: **Automata & JAC**. [S.l.: s.n.], 2012. p. 52–61. Nenhuma citação no texto.

CARNEIRO, M. G.; OLIVEIRA, G. M. B. SCAS-IS: Knowledge extraction and reuse in multiprocessor task scheduling based on cellular automata. In: **Brazilian Symposium on Neural Networks**. [S.l.: s.n.], 2012. p. 142–147. Nenhuma citação no texto.

CARNEIRO, M. G.; OLIVEIRA, G. M. B. Synchronous cellular automata-based scheduler initialized by heuristic and modeled by a pseudo-linear neighborhood. **Natural Computing**, Springer, v. 12, n. 3, p. 339–351, 2013. Nenhuma citação no texto.

CARNEIRO, M. G.; ROSA, J. L.; LOPES, A. A.; ZHAO, L. Classificação de alto nível utilizando grafo k-associados ótimo. In: **International Workshop on Web and Text Intelligence**. [S.l.: s.n.], 2012. p. 1–10. Nenhuma citação no texto.

CARNEIRO, M. G.; ROSA, J. L. G.; LOPES, A. A.; ZHAO, L. Network-based data classification: combining k-associated optimal graphs and high-level prediction. **Journal of the Brazilian Computer Society**, Springer, v. 20, n. 1, p. 1–14, 2014. Nenhuma citação no texto.

CARNEIRO, M. G.; ZHAO, L. Improving semantic role labeling with complex network measures (submetido). In: **IEEE International Joint Conference on Neural Networks**. [S.l.: s.n.]. Nenhuma citação no texto.

CARNEIRO, M. G.; ZHAO, L. Organizational data classification based on importance concept of complex networks (aceito condicionalmente: revisões simples). **IEEE Transactions on Neural Networks and Learning Systems**. Nenhuma citação no texto.

CARNEIRO, M. G.; ZHAO, L. High level classification totally based on complex networks. In: **IEEE BRICS Congress on Computational Intelligence**. [S.l.: s.n.], 2013. p. 507–514. Nenhuma citação no texto.

CARNEIRO, M. G.; ZHAO, L.; CHENG, R.; JIN, Y. Network structural optimization based on swarm intelligence for highlevel classification. In: **IEEE International Joint Conference on Neural Networks**. [S.l.: s.n.], 2016. p. 3737–3744. Citado na página 62.

CARNEIRO, M. G.; ZHAO, L.; JIN, Y. Bio-inspired structural optimization for network-based data classification (submetido). **IEEE Transactions on Cybernetics**. Nenhuma citação no texto.

CARNEIRO, M. G.; ZHAO, L.; ROSA, J. L. G. Graph-based semi-supervised learning for semantic role diffusion. In: **Symposium on Knowledge Discovery, Mining and Learning**. [S.l.: s.n.], 2016. p. 1–8. Nenhuma citação no texto.

CARVALHO, T. I.; CARNEIRO, M. G.; OLIVEIRA, G. M. B. A comparison of forecast dynamical behavior parameter and direct verification of lattice's configuration using a cellular automata model to task scheduling. In: **International Conference on Cellular Automata for Research and Industry**. [S.l.: s.n.], 2016. p. 258–268. Nenhuma citação no texto.

CARVALHO, T. I.; CARNEIRO, M. G.; OLIVEIRA, G. M. B. A hybrid strategy to evolve cellular automata rules with a desired dynamical behavior applied to the task scheduling problem. In: **Brazilian Conference on Intelligent Systems**. [S.l.: s.n.], 2016. p. 1–6. Nenhuma citação no texto.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, v. 2, n. 3, p. 27, 2011. Citado na página 92.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. **Semi-Supervised Learning**. [S.l.]: MIT Press, 2006. Citado 17 vezes nas páginas 2, 5, 34, 39, 40 e 43.

CHEN, J.; FANG, H.-r.; SAAD, Y. Fast approximate k nn graph construction for high dimensional data via recursive lanczos bisection. **Journal of Machine Learning Research**, JMLR.org, v. 10, p. 1989–2012, 2009. Citado 3 vezes nas páginas 70 e 86.

CHENG, R.; JIN, Y. A social learning particle swarm optimization algorithm for scalable optimization. **Information Sciences**, Elsevier, v. 291, p. 43–60, 2015. Citado 2 vezes nas páginas 104 e 105.

CIECHOLEWSKI, M. Support vector machine approach to cardiac spect diagnosis. In: **Combinatorial Image Analysis**. [S.l.]: Springer, 2011. p. 432–443. Citado na página 97.

CLAUSET, A.; NEWMAN, M. E. J.; MOORE, C. Finding community structure in very large networks. **Physical Review E**, American Physical Society, v. 70, p. 066111, 2004. Citado 9 vezes nas páginas 3, 22 e 30.

COHEN, R.; HAVLIN, S. **Complex networks: structure, robustness and function**. [S.l.]: Cambridge University Press, 2010. Citado 14 vezes nas páginas 2, 3, 9 e 32.

COLLOBERT, R.; WESTON, J.; BOTTOU, L.; KARLEN, M.; KAVUKCUOGLU, K.; KUKSA, P. P. Natural language processing (almost) from scratch. **Journal of Machine Learning Research**, JMLR.org, v. 12, p. 2461–2505, 2011. Citado 7 vezes nas páginas 61 e 62.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Kluwer Academic Publishers, v. 20, n. 3, p. 273–297, 1995. Citado 4 vezes nas páginas 4 e 40.

- COSTA, L. d. F.; JR, O. N. O.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. R. V.; ANTQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, Taylor & Francis, v. 60, n. 3, p. 329–412, 2011. Citado na página 3.
- COSTA, L. d. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in Physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007. Citado na página 18.
- CUPERTINO, T. H. **Machine learning via dynamical processes on complex networks**. Tese (Doutorado) — Universidade de São Paulo, 2013. Citado na página 101.
- CUPERTINO, T. H.; CARNEIRO, M. G.; ZHAO, L. Dimensionality reduction with the k-associated optimal graph applied to image classification. In: **IEEE International Conference on Imaging Systems and Techniques**. [S.l.: s.n.], 2013. p. 366–371. Citado 4 vezes nas páginas 51 e 52.
- CUPERTINO, T. H.; HUERTAS, J.; ZHAO, L. Data clustering using controlled consensus in complex networks. **Neurocomputing**, Elsevier, v. 118, p. 132–140, 2013. Citado na página 43.
- CUPERTINO, T. H.; ZHAO, L.; CARNEIRO, M. G. Network-based supervised data classification by using an heuristic of ease of access. **Neurocomputing**, Elsevier, v. 149, p. 86–92, 2015. Citado 3 vezes nas páginas 45, 50 e 101.
- DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, JMLR.org, v. 7, p. 1–30, 2006. Citado 5 vezes nas páginas 77, 93 e 113.
- DESCHACHT, K.; MOENS, M.-F. Semi-supervised semantic role labeling using the latent words language model. In: **ACL Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2009. p. 21–29. Citado na página 62.
- DIESTEL, R. **Graph Theory**. [S.l.]: Springer, 2006. Citado na página 15.
- DILLON, G. **Introduction to contemporary linguistic semantics**. [S.l.]: Prentice-Hall, 1977. Citado na página 55.
- DOROGOVTSSEV, S. N.; MENDES, J. F. Evolution of networks. **Advances in physics**, Taylor & Francis, v. 51, n. 4, p. 1079–1187, 2002. Citado na página 14.
- DOWTY, D. Thematic proto-roles and argument selection. **Language**, Linguistic Society of America, v. 67, n. 3, p. 547–619, 1991. Citado na página 55.
- DUDA, R.; HART, P.; STORK, D. **Pattern Classification**. [S.l.]: Wiley, 2007. Citado 4 vezes nas páginas 1 e 39.

- DURAN, M. S.; ALUÍSIO, S. M. Propbank-br: a brazilian treebank annotated with semantic role labels. In: **International Conference on Language Resources and Evaluation**. [S.l.: s.n.], 2012. p. 1862–1867. Citado 3 vezes nas páginas 9, 56 e 126.
- ERDÖS, P.; RÉNYI, A. On random graphs I. **Publ. Math. Debrecen**, v. 6, p. 290–297, 1959. Citado na página 26.
- FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. In: **ACM SIGCOMM Computer Communication Review**. [S.l.: s.n.], 1999. v. 29, n. 4, p. 251–262. Citado na página 2.
- FAN, R.-E.; CHANG, K.-W.; HSIEH, C.-J.; WANG, X.-R.; LIN, C.-J. Liblinear: A library for large linear classification. **The Journal of Machine Learning Research**, JMLR.org, v. 9, p. 1871–1874, 2008. Citado na página 125.
- FILIPPOVA, K.; SURDEANU, M.; CIARAMITA, M.; ZARAGOZA, H. Company-oriented extractive summarization of financial news. In: **ACL European Chapter of the Association for Computational Linguistics**. [S.l.: s.n.], 2009. p. 246–254. Citado na página 55.
- FILLMORE, C.; JOHNSON, C.; PETRUCK, M. Background to framenet. **International Journal of Lexicography**, Oxford Journals, v. 16, n. 3, p. 235–250, 2003. Citado na página 56.
- FILLMORE, C. J. The case for case. In: **Universals in Linguistic Theory**. [S.l.]: Holt, Rinehart and Winston, 1968. p. 1–88. Citado na página 55.
- FILLMORE, C. J. Frame semantics and the nature of language. In: **Annals of the New York Academy of Sciences**. [S.l.: s.n.], 1976. v. 280, p. 20–32. Citado na página 55.
- FONSECA, E. R. **Uma abordagem conexcionista para anotação de papéis semânticos**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013. Citado na página 61.
- FONSECA, E. R.; ROSA, J. L. G. A two-step convolutional neural network approach for semantic role labeling. In: **IEEE International Joint Conference on Neural Networks**. [S.l.: s.n.], 2013. p. 2955–2961. Citado 8 vezes nas páginas 9, 61 e 62.
- FORTUNATO, S. Community detection in graphs. **Physics Reports**, Elsevier, v. 486, n. 3–5, p. 75 – 174, 2010. Citado 34 vezes nas páginas 3, 5, 14, 23 e 43.
- FRANK, A.; ASUNCION, A. **UCI Machine Learning Repository**. [S.l.]: University of California, Irvine, School of Information and Computer Sciences, 2010. Citado 5 vezes nas páginas 76, 92 e 112.
- FRAWLEY, W. **Linguistic semantics**. [S.l.]: Lawrence Erlbaum Associates, 1992. Citado na página 55.

- FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, American Sociological Association, v. 40, n. 1, p. 35–41, 1977. Citado na página 21.
- FREEMAN, L. C. Centrality in social networks conceptual clarification. **Social networks**, Elsevier, v. 1, n. 3, p. 215–239, 1979. Citado na página 21.
- FRUCHTERMAN, T. M.; REINGOLD, E. M. Graph drawing by force-directed placement. **Software: Practice and experience**, Wiley Online Library, v. 21, n. 11, p. 1129–1164, 1991. Citado 2 vezes nas páginas 97 e 121.
- FÜRSTENAU, H.; LAPATA, M. Semi-supervised semantic role labeling via structural alignment. **Computational Linguistics**, MIT Press, v. 38, n. 1, p. 135–171, 2012. Citado na página 62.
- GALLAGER, R. G. **Discrete stochastic processes**. [S.l.]: Springer Science & Business Media, 2012. Citado na página 51.
- GIESE, M. A.; POGGIO, T. Neural mechanisms for the recognition of biological movements. **Nature Reviews Neuroscience**, Nature Publishing Group, v. 4, n. 3, p. 179–192, 2003. Citado na página 3.
- GILDEA, D.; JURAFSKY, D. Automatic labeling of semantic roles. **Computational Linguistics**, MIT Press Cambridge, v. 28, n. 3, p. 245–288, 2002. Citado 15 vezes nas páginas 60, 62, 126 e 132.
- GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citado 4 vezes nas páginas 16, 30 e 31.
- GLEICH, D. F.; SESHADHRI, C. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In: **ACM SIGKDD Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2012. p. 597–605. Citado na página 81.
- GOLDENBERG, A.; ZHENG, A. X.; FIENBERG, S. E.; AIROLDI, E. M. A survey of statistical network models. **Foundations and Trends® in Machine Learning**, Now Publishers Inc., v. 2, n. 2, p. 129–233, 2010. Citado na página 3.
- GOLDMAN, S.; ZHOU, Y. Enhancing supervised learning with unlabeled data. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2000. p. 327–334. Citado na página 7.
- GROSS, J.; YELLEN, J. **Graph Theory and Its Applications**. [S.l.]: Chapman & Hall/CRC, 2006. Citado na página 15.
- HACIOGLU, K. Semantic role labeling using dependency trees. In: **ACL International Conference on Computational Linguistics**. [S.l.: s.n.], 2004. p. 1273–1276. Citado na página 60.

HARTMANN, N. S. **Anotação automática de papéis semânticos de textos jornalísticos e de opinião sobre árvores sintáticas não revisadas**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015. Citado 5 vezes nas páginas 9 e 61.

HASAN, M. A.; CHAOJI, V.; SALEM, S.; ZAKI, M. Link prediction using supervised learning. In: **Workshop on Link Analysis, Counter-terrorism and Security**. [S.l.: s.n.], 2006. Citado na página 44.

HASLINGER, C.; SCHWEIFER, N.; STILGENBAUER, S.; DÖHNER, H.; LICHTER, P.; KRAUT, N.; STRATOWA, C.; ABSEHER, R. Microarray gene expression profiling of b-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and vh mutation status. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 22, n. 19, p. 3937–3949, 2004. Citado na página 112.

HAYKIN, S. S. **Neural networks and learning machines**. [S.l.]: Pearson, 2009. Citado 3 vezes nas páginas 4, 7 e 37.

HE, S.; GILDEA, D. **Self-training and co-training for semantic role labeling: Primary report**. [S.l.], 2006. Citado na página 62.

HEEKEREN, H. R.; MARRETT, S.; BANDETTINI, P.; UNGERLEIDER, L. A general mechanism for perceptual decision-making in the human brain. **Nature**, Nature Publishing Group, v. 431, n. 7010, p. 859–862, 2004. Citado na página 3.

HSU, C.-W.; LIN, C.-J. A comparison of methods for multiclass support vector machines. **IEEE Transactions on Neural Networks**, IEEE, v. 13, n. 2, p. 415–425, 2002. Citado na página 37.

IMAN, R. L.; DAVENPORT, J. M. Approximations of the critical region of the fbietkan statistic. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 9, n. 6, p. 571–595, 1980. Citado 4 vezes nas páginas 77 e 93.

JACKENDOFF, R. **Semantic Interpretation in Generative Grammar**. [S.l.]: MIT Press, 1972. Citado na página 55.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, Elsevier, v. 31, n. 8, p. 651–666, 2010. Citado 5 vezes nas páginas 38 e 43.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM computing surveys**, v. 31, n. 3, p. 264–323, 1999. Citado 5 vezes nas páginas 34 e 38.

JEBARA, T.; WANG, J.; CHANG, S.-F. Graph construction and b-matching for semi-supervised learning. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2009. p. 441–448. Citado 2 vezes nas páginas 43 e 132.

- JENSEN, D.; NEVILLE, J.; GALLAGHER, B. Why collective inference improves relational classification. In: **ACM International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2004. p. 593–598. Citado na página 44.
- JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABÁSI, A.-L. The large-scale organization of metabolic networks. **Nature**, Nature Publishing Group, v. 407, n. 6804, p. 651–654, 2000. Citado na página 2.
- JOHANSSON, R. **Dependency-based semantic analysis of natural-language text**. Tese (Doutorado) — Lund University, 2008. Citado na página 60.
- JOHNSON, R.; ZHANG, T. On the effectiveness of laplacian normalization for graph semi-supervised learning. **Journal of Machine Learning Research**, v. 8, p. 1489–1517, 2007. Citado na página 131.
- JOLLIFFE, I. **Principal component analysis**. [S.l.]: Wiley Online Library, 2002. Citado na página 39.
- KALJAH, R. S. Z. Adapting self-training for semantic role labeling. In: **ACL Student Research Workshop**. [S.l.: s.n.], 2010. p. 91–96. Citado na página 62.
- KENNEDY, J. **Swarm Intelligence**. [S.l.]: Springer, 2006. Citado na página 104.
- KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: **IEEE International Conference on Neural Networks**. [S.l.: s.n.], 1995. p. 1942–1948. Citado na página 104.
- KIDWELL, D. **Lazy Learning**. [S.l.]: Springer, 2013. Citado 2 vezes nas páginas 4 e 37.
- KIPPER, K.; DANG, H. T.; PALMER, M. Class-based construction of a verb lexicon. In: **AAAI Conference On Artificial Intelligence**. [S.l.]: Association for the Advancement of Artificial Intelligence, 2000. p. 691–696. Citado na página 56.
- KURGAN, L. A.; CIO, K. J.; TADEUSIEWICZ, R.; OGIELA, M.; GOODENDAY, L. S. Knowledge discovery approach to automated cardiac spect diagnosis. **Artificial Intelligence in Medicine**, Elsevier, v. 23, n. 2, p. 149–169, 2001. Citado 4 vezes nas páginas 95 e 97.
- LANG, J.; LAPATA, M. Unsupervised semantic role induction via split-merge clustering. In: **Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2011. v. 1, p. 1117–1126. Citado na página 62.
- LANG, J.; LAPATA, M. Unsupervised semantic role induction with graph partitioning. In: **ACL Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2011. p. 1320–1331. Citado na página 62.
- LANGVILLE, A. N.; MEYER, C. D. **Google's PageRank and beyond: The science of search engine rankings**. [S.l.]: Princeton University Press, 2011. Citado 6 vezes nas páginas 3 e 81.

- LATORA, V.; MARCHIORI, M. Efficient behavior of small-world networks. **Physical Review Letters**, American Physical Society, v. 87, p. 198701, 2001. Citado na página 84.
- LI, W.; HUANG, H.; PENG, F. Trajectory classification in circular restricted three-body problem using support vector machine. **Advances in Space Research**, Elsevier, v. 56, n. 2, p. 273–280, 2015. Citado na página 1.
- LIBBRECHT, M. W.; NOBLE, W. S. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, Nature Publishing Group, v. 16, n. 6, p. 321–332, 2015. Citado na página 1.
- LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. **Journal of the American Society for Information Science and Technology**, Wiley Online Library, v. 58, n. 7, p. 1019–1031, 2007. Citado na página 44.
- LILJEROS, F.; EDLING, C. R.; AMARAL, L. A. N.; STANLEY, H. E.; ÅBERG, Y. The web of human sexual contacts. **Nature**, Nature Publishing Group, v. 411, n. 6840, p. 907–908, 2001. Citado na página 2.
- LU, Q.; GETOOR, L. Link-based classification. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2003. p. 496–503. Citado 4 vezes nas páginas 44 e 81.
- MACSKASSY, S. A.; PROVOST, F. A simple relational classifier. In: **International Workshop on Multi-Relational Data Mining**. [S.l.: s.n.], 2003. p. 64–76. Citado na página 44.
- MACSKASSY, S. A.; PROVOST, F. Classification in networked data: A toolkit and a univariate case study. **Journal of Machine Learning Research**, JMLR. org, v. 8, p. 935–983, 2007. Citado na página 44.
- MAIER, M.; LUXBURG, U. V.; HEIN, M. Influence of graph construction on graph-based clustering measures. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2008. p. 1025–1032. Citado na página 8.
- MÀRQUEZ, L.; CARRERAS, X.; LITKOWSKI, K. C.; STEVENSON, S. Semantic role labeling: an introduction to the special issue. **Computational Linguistics**, MIT Press, v. 34, n. 2, p. 145–159, 2008. Citado na página 56.
- MCDOWELL, L. K.; GUPTA, K. M.; AHA, D. W. Cautious collective classification. **Journal of Machine Learning Research**, JMLR.org, v. 10, n. Dec, p. 2777–2836, 2009. Citado na página 44.
- MILGRAM, S. The small world problem. **Psychology Today**, New York, v. 2, n. 1, p. 60–67, 1967. Citado na página 27.
- MITCHELL, T. **Machine Learning**. [S.l.]: McGraw-Hill, 1997. Citado 10 vezes nas páginas 1, 34, 36 e 38.

- MIZRUCHI, M. S. **The American corporate network, 1904-1974**. [S.l.]: Sage Publications, Inc, 1982. Citado na página 2.
- NEAPOLITAN, R. E. **Learning Bayesian Networks**. [S.l.]: Prentice-Hall, Inc., 2003. Citado 3 vezes nas páginas 4 e 37.
- NEWMAN, M. **Networks: An Introduction**. [S.l.]: Oxford University Press, Inc., 2010. Citado 26 vezes nas páginas 2, 3, 8, 14, 25 e 32.
- NEWMAN, M. E. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 98, n. 2, p. 404–409, 2001. Citado na página 2.
- NEWMAN, M. E. Assortative mixing in networks. **Physical review letters**, American Physical Society, v. 89, n. 20, p. 208701, 2002. Citado na página 19.
- NEWMAN, M. E. The structure and function of complex networks. **SIAM Review**, SIAM, v. 45, n. 2, p. 167–256, 2003. Citado 20 vezes nas páginas 2, 3, 14 e 18.
- NEWMAN, M. E. Communities, modules and large-scale structure in networks. **Nature Physics**, Nature Publishing Group, v. 8, n. 1, p. 25–31, 2012. Citado na página 5.
- NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical Review E**, American Physical Society, v. 69, n. 2, p. 026113, 2004. Citado 11 vezes nas páginas 16, 3, 14, 31 e 32.
- NEWMAN, M. E. J. Modularity and community structure in networks. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 103, n. 23, p. 8577–8582, 2006. Citado 3 vezes nas páginas 22 e 23.
- NICKEL, M.; MURPHY, K.; TRESP, V.; GABRILOVICH, E. A review of relational machine learning for knowledge graphs. **arXiv preprint arXiv:1503.00759**, p. 1–34, 2015. Citado na página 44.
- NIELSEN, R. D.; MASANZ, J.; OGREN, P.; WARD, W.; MARTIN, J. H.; SAVOVA, G.; PALMER, M. An architecture for complex clinical question answering. In: **ACM International Health Informatics Symposium**. [S.l.: s.n.], 2010. p. 395–399. Citado na página 55.
- NIELSEN, T. D.; JENSEN, F. V. **Bayesian networks and decision graphs**. [S.l.]: Springer Science & Business Media, 2009. Citado na página 4.
- NIGAM, K.; MCCALLUM, A. K.; THRUN, S.; MITCHELL, T. Text classification from labeled and unlabeled documents using EM. **Machine learning**, Kluwer Academic Publishers, v. 39, n. 2-3, p. 103–134, 2000. Citado na página 40.

OSHIRO, T. M.; PEREZ, P. S.; BARANAUSKAS, J. A. How many trees in a random forest? In: **Machine Learning and Data Mining in Pattern Recognition**. [S.l.]: Springer, 2012. v. 7376, p. 154–168. Citado na página 93.

OSMAN, A. H.; SALIM, N.; BINWAHLAN, M. S.; ALTEEB, R.; ABUOBIEDA, A. An improved plagiarism detection scheme based on semantic role labeling. **Applied Soft Computing**, Elsevier, v. 12, n. 5, p. 1493–1502, 2012. Citado 5 vezes nas páginas 55 e 60.

OZAKI, K.; SHIMBO, M.; KOMACHI, M.; MATSUMOTO, Y. Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data. In: **ACL Conference on Computational Natural Language Learning**. [S.l.: s.n.], 2011. p. 154–162. Citado 2 vezes nas páginas 8 e 42.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. **The PageRank citation ranking: bringing order to the web**. [S.l.], 1999. Citado na página 3.

PALLA, G.; DERÉNYI, I.; FARKAS, I.; VICSEK, T. Uncovering the overlapping community structure of complex networks in nature and society. **Nature**, Nature Publishing Group, v. 435, n. 7043, p. 814–818, 2005. Citado na página 5.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The proposition bank: An annotated corpus of semantic roles. **Computational Linguistics**, MIT Press, v. 31, p. 71–106, 2005. Citado na página 56.

PALMER, M.; GILDEA, D.; XUE, N. **Semantic Role Labeling**. [S.l.]: Morgan & Claypool Publishers, 2010. Citado 4 vezes nas páginas 9, 56 e 59.

PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic spreading in scale-free networks. **Physical Review Letters**, American Physical Society, v. 86, n. 14, p. 3200, 2001. Citado na página 14.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, JMLR.org, v. 12, p. 2825–2830, 2011. Citado na página 92.

PRADHAN, S.; HACIOGLU, K.; KRUGLER, V.; WARD, W.; MARTIN, J. H.; JURAFSKY, D. Support vector learning for semantic argument classification. **Machine Learning**, Kluwer Academic Publishers, v. 60, n. 1-3, p. 11–39, 2005. Citado na página 60.

PRADHAN, S. S.; WARD, W.; MARTIN, J. H. Towards robust semantic role labeling. **Computational Linguistics**, MIT Press, v. 34, n. 2, p. 289–310, 2008. Citado 16 vezes nas páginas 60, 61, 62, 126 e 132.

- PUNYAKANOK, V.; ROTH, D.; YIH, W.-t. The necessity of syntactic parsing for semantic role labeling. In: **International Joint Conference on Artificial Intelligence**. [S.l.: s.n.], 2005. p. 1117–1123. Citado na página 55.
- QUILES, M. G.; ZHAO, L.; ALONSO, R. L.; ROMERO, R. A. F. Particle competition for complex network community detection. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing, v. 18, n. 3, p. 033107, 2008. Citado na página 5.
- QUINLAN, J. Induction of decision trees. **Machine Learning**, Kluwer Academic Publishers, v. 1, p. 81–106, 1986. Citado 5 vezes nas páginas 4 e 36.
- ROSA, J. L. G.; ADAN-COELLO, J. M. Biologically plausible connectionist prediction of natural language thematic relations. **Journal of Universal Computer Science**, Graz University of Technology, v. 16, n. 21, p. 3245–3277, 2010. Citado 7 vezes nas páginas 56 e 62.
- ROWEIS, S. T.; SAUL, L. K. Nonlinear dimensionality reduction by locally linear embedding. **Science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2323–2326, 2000. Citado na página 3.
- SANTOS, C. N.; MILIDIÚ, R. L. Semantic role labeling. In: **Entropy Guided Transformation Learning: Algorithms and Applications**. [S.l.]: Springer, 2012. cap. 8, p. 59–69. Citado na página 60.
- SCHAEFFER, S. E. Graph clustering. **Computer Science Review**, Elsevier, v. 1, n. 1, p. 27–64, 2007. Citado 9 vezes nas páginas 34 e 43.
- SCOTT, J. **Social network analysis**. [S.l.]: Sage, 2012. Citado na página 2.
- SEGAL, E.; WANG, H.; KOLLER, D. Discovering molecular pathways from protein interaction and gene expression data. **Bioinformatics**, Oxford University Press, v. 19, n. suppl 1, p. i264–i272, 2003. Citado na página 44.
- SEN, P.; NAMATA, G.; BILGIC, M.; GETOOR, L.; GALLIGHER, B.; ELIASSI-RAD, T. Collective classification in network data. **AI magazine**, Association for the Advancement of Artificial Intelligence, v. 29, n. 3, p. 93, 2008. Citado na página 44.
- SEQUEIRA, J.; GONÇALVES, T.; QUARESMA, P. Semantic role labeling for portuguese – a preliminary approach. In: **Computational Processing of the Portuguese Language**. [S.l.]: Springer, 2012. v. 7243, p. 193–203. Citado na página 61.
- SIEGEL, M.; ENGEL, A. K.; DONNER, T. H. Cortical network dynamics of perceptual decision-making in the human brain. **Frontiers in Human Neuroscience**, Frontiers, v. 5, p. 21, 2011. Citado na página 3.

SILVA, T. C.; ZHAO, L. Network-based high level data classification. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 23, n. 6, p. 954–970, 2012. Citado 17 vezes nas páginas 5, 10, 43, 44, 45, 49, 50, 65, 68, 77, 87 e 122.

SILVA, T. C.; ZHAO, L. Network-based stochastic semisupervised learning. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 23, n. 3, p. 451–466, 2012. Citado 15 vezes nas páginas 5 e 43.

SILVA, T. C.; ZHAO, L. Stochastic competitive learning in complex networks. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 23, n. 3, p. 385–398, 2012. Citado na página 43.

SILVA, T. C.; ZHAO, L. Uncovering overlapping cluster structures via stochastic competitive learning. **Information Sciences**, Elsevier, v. 247, p. 40–61, 2013. Citado na página 5.

SILVA, T. C.; ZHAO, L. High-level pattern-based classification via tourist walks in networks. **Information Sciences**, Elsevier, v. 294, p. 109–126, 2015. Citado na página 43.

SILVA, T. C.; ZHAO, L. **Machine Learning in Complex Networks**. [S.l.]: Springer, 2016. Citado na página 2.

SINDHWANI, V.; NIYOGI, P.; BELKIN, M. Beyond the point cloud: from transductive to semi-supervised learning. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2005. p. 824–831. Citado na página 44.

SOUSA, C. A. R. de; REZENDE, S. O.; BATISTA, G. E. Influence of graph construction on semi-supervised learning. In: **Machine Learning and Knowledge Discovery in Databases**. [S.l.]: Springer, 2013. p. 160–175. Citado na página 8.

SPORNS, O. Network analysis, complexity, and brain function. **Complexity**, Wiley Online Library, v. 8, n. 1, p. 56–60, 2002. Citado na página 2.

STROGATZ, S. H. Exploring complex networks. **Nature**, Nature Publishing Group, v. 410, n. 6825, p. 268–276, 2001. Citado na página 14.

SURDEANU, M.; HARABAGIU, S.; WILLIAMS, J.; AARSETH, P. Using predicate-argument structures for information extraction. In: **Annual Meeting of the Association for Computational Linguistics**. [S.l.]: Association for Computational Linguistics, 2003. p. 8–15. Citado na página 62.

TABRIZI, S. A.; SHAKERY, A.; ASADPOUR, M.; ABBASI, M.; TAVALLAIE, M. A. Personalized pagerank clustering: A graph clustering algorithm based on random walks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 392, n. 22, p. 5772–5785, 2013. Citado 6 vezes nas páginas 3 e 81.

- TAULÉ, M. A. M. M.; RECASENS, M. Ancora: Multilevel annotated corpora for catalan and spanish. In: **International Conference on Language Resources and Evaluation**. [S.l.: s.n.], 2008. p. 96–101. Citado na página 56.
- TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **Science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000. Citado na página 3.
- TITOV, I.; KLEMENTIEV, A. Semi-supervised semantic role labeling: Approaching from an unsupervised perspective. In: **International Conference on Computational Linguistics**. [S.l.: s.n.], 2012. p. 2635–2652. Citado na página 62.
- VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer, 2013. Citado 4 vezes nas páginas 4 e 37.
- VAPNIK, V. N.; VAPNIK, V. **Statistical learning theory**. [S.l.]: Wiley, 1998. Citado na página 40.
- WANG, F.; ZHANG, C. Label propagation through linear neighborhoods. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 20, n. 1, p. 55–67, 2008. Citado na página 43.
- WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. **Nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado 11 vezes nas páginas 16, 14, 20, 25, 27 e 28.
- WU, D.; FUNG, P. Semantic roles for SMT: a hybrid two-pass model. In: **ACL Annual Conference of the North American Chapter of the Association for Computational Linguistics**. [S.l.: s.n.], 2009. p. 13–16. Citado na página 55.
- XUE, N. Labeling chinese predicates with semantic roles. **Computational Linguistics**, MIT Press, v. 34, n. 2, p. 225–255, 2008. Citado na página 56.
- XUE, N.; PALMER, M. Calibrating features for semantic role labeling. In: **ACL Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2004. p. 88–94. Citado na página 59.
- YAN, S.; XU, D.; ZHANG, B.; ZHANG, H.-J.; YANG, Q.; LIN, S. Graph embedding and extensions: A general framework for dimensionality reduction. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 29, n. 1, p. 40–51, 2007. Citado 21 vezes nas páginas 16, 5, 34, 43, 51 e 52.
- YANG, J.; LESKOVEC, J. Defining and evaluating network communities based on ground-truth. **Knowledge and Information Systems**, Springer, v. 42, n. 1, p. 181–213, 2015. Citado na página 3.

ZACHARY, W. W. An information flow model for conflict and fission in small groups. **Journal of Anthropological Research**, JSTOR, v. 33, n. 4, p. 452–473, 1977. Citado 3 vezes nas páginas 15 e 18.

ZHOU, D.; BOUSQUET, O.; LAL, T. N.; WESTON, J.; SCHÖLKOPF, B. Learning with local and global consistency. In: . [S.l.: s.n.], 2004. p. 321–328. Citado 3 vezes nas páginas 53, 54 e 131.

ZHU, X. **Semi-Supervised Learning Literature Survey**. [S.l.], 2008. Citado 19 vezes nas páginas 5, 34, 41 e 43.

ZHU, X.; GHAHRAMANI, Z.; LAFFERTY, J. Semi-supervised learning using gaussian fields and harmonic functions. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2003. p. 912–919. Citado 3 vezes nas páginas 40, 53 e 131.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. **Synthesis lectures on artificial intelligence and machine learning**, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–130, 2009. Citado na página 40.

ZHU, X.; LAFFERTY, J.; ROSENFELD, R. **Semi-supervised learning with graphs**. [S.l.]: Carnegie Mellon University, 2005. Citado 8 vezes nas páginas 3, 8 e 41.