Uso de Redes Complexas na Classificação Relacional

Robson Carlos da Motta

Uso de Redes Complexas na Classificação Relacional

Robson Carlos da Motta

Orientador: ***Prof. Dr. Alneu de Andrade Lopes***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional.

USP – São Carlos
Maio/2009

Dedicatória

Aos meus pais,
Alice e Francisco.

Agradecimentos

Aos meus pais, Francisco e Alice, pelo amor incondicional, pela compreensão, pelos sorrisos, e principalmente pelos momentos singelos que me mostraram verdadeiras lições de vida.

Aos meus irmãos, Vitor e Vinícius, pela troca constante de experiências nas quais todos crescemos juntos.

Ao professor Alneu de Andrade Lopes, pela dedicação e paciência com que me conduziu durante todo o tempo, e pela confiança e amizade sem as quais não seria possível o desenvolvimento deste trabalho com tamanha satisfação.

Aos amigos e professores do ICMC, pelo companheirismo e pelos bons momentos de convivência, dentro e fora do laboratório: André Maletzke, André Rossi, Bruno Magalhães, Carlos Ferrero, Edson Matsubara, Gustavo Batista, Ígor Braga, Leonardo Almeida, Márcio Basgalupp, Maria Carolina Monard, Maria Cristina Oliveira, Merley Conrado, Rafael Giusti, Renato Silva, Ronaldo Prati, Solange Rezende, Victor Laguna e Zhao Liang.

Aos meus amigos de Mogi Mirim, com os quais cresci, que são tantos e mal posso me lembrar de todos.

Aos meus amigos da República Chico Lopes, que convivem comigo diariamente, formando nossa família em São Carlos. E a todos meus amigos de São Carlos que sempre estiveram presentes.

Às funcionárias do setor de Pós-Graduação do ICMC/USP, Ana Paula, Beth, Laura e Lívia, pelos excelentes serviços prestados à comunidade acadêmica dessa unidade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro essencial para a realização deste trabalho.

E a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

Resumo

A vasta quantidade de informações disponível sobre qualquer área de conhecimento torna cada vez mais difícil selecionar e analisar informações específicas e relevantes sobre determinado assunto. Com isso, faz-se necessário o aprimoramento de técnicas automáticas para recuperação, análise e extração de conhecimento em conjuntos de dados, destacando-se dessa forma as pesquisas em Aprendizado de Máquina e em Mineração de Dados. Em aprendizado de máquina e em mineração, a grande maioria das técnicas utiliza-se de uma representação proposicional dos dados, que considera apenas características individuais dos objetos descritos em uma tabela atributo-valor. Porém, existem aplicações nas quais além da descrição dos objetos também estão disponíveis informações sobre relações existentes entre eles. Esses domínios podem ser representados via grafos, nos quais vértices representam objetos e arestas relações entre objetos, possibilitando a aplicação de técnicas relacionais aos dados. Conceitos de Redes Complexas (RC) podem ser utilizados neste contexto. RC é um campo de pesquisa recente e ativo, que estuda o comportamento de diversos sistemas reais, modelados via grafos. Entretanto, ainda há poucos trabalhos que utilizam Redes Complexas em aprendizado de máquina ou mineração de dados. Este projeto apresenta uma proposta de utilização do formalismo de redes complexas e grafos para descoberta de padrões no contexto de aprendizado supervisionado. O formalismo de grafos permite representar as relações entre objetos e características particulares do domínio, permitindo agregar informações estruturais das relações à descoberta de conhecimento. Especificamente, neste trabalho desenvolve-se uma representação relacional baseada em grafos construídos a partir de relações de similaridade entre objetos. Baseado nesta representação são propostas abordagens de classificação relacional. Também é proposto um modelo de rede denominado K-Associados. Propriedades da rede K-Associados foram investigadas. Os resultados experimentais demonstram um grande potencial para classificação utilizando os algoritmos de classificação e de formação de redes propostos.

Abstract

The vast amount of information available on any area of knowledge makes selecting and analyzing information on a specific topic increasingly difficult. Therefore, it is necessary the improvement of techniques for automatic information retrieval, analysis, and knowledge extraction from data sets. In this scenario, especial attention must be addressed for Machine Learning and Data Mining researches. In machine learning and data mining, most of the techniques uses a propositional representation, which considers only the characteristics of the objects described into an attribute-value table. However, there are domains where, in addition to the description of the objects, it is also available information about relationship between them. Such domains can be represented by graphs where vertices represent objects and edges relationship between objects, enabling the application of techniques for relational data. Concepts of complex networks (CN) can be useful in this context. CN is a recent and active research field, which studies the behavior of many real systems modeled by graphs. However, there is little work in machine learning or data mining applying CN concepts. This project presents a proposal to use the formalism of complex networks and graphs to discover patterns in the context of supervised learning. The formalism of graphs can represent relationships between objects and characteristics of the domain, allowing adding structural knowledge embedded in a graph into the data mining process. Specifically, this work develops a relational representation based on graphs constructed taking into consideration the similarity between objects. Based on this representation, relational classification approaches are proposed. It is also proposed a network referred to K-Associate Network. Properties of the K-Associate Network were investigated. The experimental results show great potential for the proposed classification and network construction algorithms.

Esta dissertação foi preparada com o formatador de textos LaTeX. Foi utilizado um estilo (*style*) desenvolvido por Ronaldo Cristiano Prati. O sistema de citações de referências bibliográficas utiliza o padrão *Chicago* do sistema BibTeX.

Algumas palavras utilizadas neste trabalho não foram traduzidas da língua inglesa para a portuguesa por serem amplamente conhecidas e difundidas na comunidade acadêmica.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

RC	Redes Complexas
CN	Complex Network
IDC	*International Data Corporation*
KDD	Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases*)
DM	Mineração de Dados (*Data Mining*)
ML	Aprendizado de Máquina (*Machine Learning*)
VDM	Mineração Visual de Dados (*Visual Data Mining*)
DER	Diagrama Entidade-Relacionamento
k-NN	*k-Nearest Neighnors*
Hc	Classificação de Hipertexto (*Hypertext classification*)
Lbc	Classificação baseada em *links* (*Link-based classification*)
wvRN	Classificador relacional baseado nos vizinhos com votação pesada (*Weighted-Vote Relational Neighbor Classifier*)
cdRN	Classificador relacional baseado na distribuição de classe dos vizinhos (*Class-Distribution Relational Neighbor Classifier*)
nBC	Classificador Bayesiano baseado apenas na rede (*Network-Only Bayes Classifier*)
nLB	Classificador baseado apenas nas conexões da rede (*Network-Only Link-Based Classifier*)

RH Rede hierárquica

RHD Rede hierárquica determinística

RHP Rede hierárquica probabilística

RHD(g) Rede hierárquica determinística construída com grau médio de entrada igual a g.

RHP(g) Rede hierárquica probabilística construída com grau médio de entrada igual a g.

cbRH Classificador baseado na rede hierárquica

kA Rede K-Associados

kA(k) Rede K-Associados construída com valor de k de entrada.

kAO Rede K-Associados Ótima

CBR-ILP-IR Base de textos sobre *Case Based Reasoning, Inductive Logic Programming e Information Retrieval*

CS Base de textos sobre Ciência da Computação (*Computer Science*)

Capítulo

1

Introdução

1.1 Contexto e motivação

O aumento do volume e do ritmo de crescimento na geração de novas informações, fez com que a capacidade mundial de armazenamento não fosse suficiente para armazenar, no universo digital, todo conteúdo produzido. Segundo a *International Data Corporation* - IDC (Gantz et al., 2008), em 2008, a quantidade de informação atingiu 287 hexabytes, e a capacidade de armazenamento digital suportou apenas 264 hexabytes. A estimativa da IDC é que em 2011 o volume de informações chegue a 1800 hexabytes, configurando um crescimento de 60% ao ano.

Essa vasta quantidade de informações disponível, torna cada vez mais difícil a exploração de informações específicas e relevantes sobre determinado assunto, pois, junto com informações úteis, ocorre uma grande quantidade de material de menor importância. Com isso, surge a necessidade de se aprimorar técnicas automáticas para exploração e análise de grandes conjuntos de dados, destacado-se as pesquisas na área de Inteligência Artificial conhecidas como Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD) (Fayyad et al., 1996).

A principal tarefa do KDD é a identificação de padrões, conhecida como Mineração de Dados (*Data Mining* - DM), que por sua vez, utiliza-se principalmente de técnicas de Aprendizado de Máquina (*Machine Learning* - ML) (Mitchell, 1997).

A maior parte dos algoritmos utilizados em aprendizado de máquina utiliza como entrada uma representação proposicional dos dados, geralmente em uma tabela atributo-valor, a qual está limitada a descrever apenas as características individuais dos objetos. Porém, a representação dos dados de forma relacional também pode ser utilizada nesse

contexto. Segundo Raedt (2008), a representação relacional apresenta uma descrição mais rica dos dados, representando, além das informações dos objetos, também a relação existente entre eles.

Um conjunto de objetos com relações entre si consiste de uma rede, que pode ser identificada em diversas situações como, por exemplo, em sistemas químicos, orgânicos e sociais (Newman, 2003).

Redes podem ser representadas por grafos, que são definidos como estruturas compostas de um conjunto de vértices e um conjunto de arestas que ligam pares de vértices segundo alguma espécie de relação. Além da representação de redes, o formalismo de grafos também é eficiente para visualização, especialmente de redes de pequeno porte. Porém, no caso de redes com milhares ou milhões de vértices, a visualização torna-se insuficiente para análise dos sistemas que representam.

No caso de grandes redes representadas por grafos, é necessário utilizar métodos estatísticos para o entendimento de características e comportamentos que se estabelecem nos sistemas modelados. As redes que representam tais sistemas são conhecidas como Redes Complexas (Newman, 2003).

O uso de propriedades de Redes Complexas permite agregar informações relacionais às informações individuais dos objetos, possibilitando a utilização de ambas as informações em um processo de mineração de dados.

Além disso, processos de visualização podem ser usados em conjunto com técnicas de mineração para auxiliar o usuário a compreender, isto é, construir rapidamente um modelo mental de um conjunto de dados, auxiliando na extração de conhecimento em um contexto de Mineração Visual de Dados (*Visual Data Mining* - VDM) (Oliveira & Levkowitz, 2003). Neste trabalho, ferramentas de visualização foram utilizadas para auxiliar na análise e compreensão das redes e do comportamento dos objetos, colaborando para a definição dos critérios dos algoritmos e para exploração dos conjuntos de dados.

1.2 Identificação do problema

Aprendizado de Máquina utiliza técnicas computacionais para obtenção automática de conhecimento, aprendendo a partir de experiências prévias na resolução de problemas (Mitchell, 1997). Uma das estratégias mais utilizadas para realização do aprendizado é a indução, que consiste em adquirir conceitos a partir de inferências indutivas sobre os fatos observados ou fornecidos (exemplos).

Uma das principais áreas do aprendizado indutivo é o aprendizado supervisionado, o qual utiliza exemplos rotulados para a construção de modelos que são utilizados para predizer os rótulos (as classes) de novos exemplos, em geral, utilizando conjuntos de dados na representação proposicional.

Porém, é possível usar uma representação mais expressiva agregando informações das

relações entre os objetos. Tal representação possibilita o uso de algoritmos de aprendizado que consideram, além das características individuais dos exemplos, as relações existentes entre eles.

É possível também, contendo ambas representações, ampliar as possibilidades de exploração dos classificadores, utilizando, por exemplo, classificação que utilizam mais de uma visão dos dados, uma dada pelas características dos objetos e outra pelas relações existentes entre eles.

Além disso, a construção de grafos possibilita a aplicação de propriedades de redes complexas para identificação de padrões e comportamentos nos conjuntos de dados, em um processo de mineração de dados.

1.3 Objetivos e metodologia

O objetivo geral deste trabalho é usar os formalismos de redes complexas e grafos para visualização dos dados e descoberta de padrões no contexto de aprendizado supervisionado. Especificamente, pretende-se abordar tarefas de classificação baseada em grafos[1].

Sendo assim, os objetivos específicos são descritos a seguir.

- Elaborar uma técnica para construção de uma representação relacional a partir de tabelas atributo-valor, realizando uma modelagem em grafos nos quais os objetos são os vértices e as arestas representam similaridade entre objetos.
- Aplicar as estatísticas definidas para Redes Complexas nos grafos construídos a partir das tabelas atributo-valor.
- Propor algoritmos de classificação baseados nos grafos gerados, e efetuar a comparação com algoritmos que utilizam a representação proposicional e outros algoritmos baseados em grafos.

Os experimentos foram realizados utilizando 18 conjuntos de dados, sendo dez conjuntos proposicionais, representados por tabelas atributo-valor, quatro conjuntos proposicionais de documentos textuais, e quatro conjuntos relacionais, representados por grafos. Para visualização dos grafos foi utilizada a API Prefuse[2] com diversas adaptações, possibilitando a avaliação visual dos grafos construídos, auxiliando no entendimento e definição de parâmetros dos algoritmos.

Para avaliação dos grafos construídos a partir de tabelas atributo-valor, utilizou-se medidas de pureza do vértice comparada com a pureza dos exemplos mais similares no espaço real dos objetos, avaliando se a estrutura de relacionamento do grafo manteve a relação de similaridade entre os dados no espaço original.

[1] No decorrer deste trabalho grafos e redes são usados como sinônimos

[2] http://prefuse.org/

Para avaliação dos classificadores relacionais propostos, quando aplicados aos conjuntos de dados proposicionais, efetuou-se a comparação com os classificadores Naive Bayes e *k-Nearest Neighbors* (k-NN). E para avaliação dos classificadores, quando aplicados aos conjuntos de dados relacionais, efetuou-se a comparação com o Classificador Bayesiano Baseado Apenas na Rede (*Network-Only Bayes Classifier* - nBC).

1.4 Contribuições

Neste trabalho, foi desenvolvida uma técnica para construção de grafos a partir de qualquer conjunto de dados no qual seja possível definir uma medida de similaridade entre os objetos.

Tal técnica constrói uma representação relacional dos dados via grafos nos quais os objetos são os vértices e as arestas ligam pares de vértices que possuem uma alta similaridade entre si. Com isso, é possível a aplicação de técnicas voltadas a grafos, como, por exemplo, utilização de propriedades de redes complexas para análise do conjunto de dados e o uso de classificadores baseados em grafos.

Para tarefa de classificação em um contexto de aprendizado supervisionado, foram propostos dois classificadores baseados em grafo, para os quais foram utilizadas as propriedades de redes complexas a fim de identificar uma forma eficiente de classificar os objetos não rotulados.

Também foi proposto um modelo de redes definido como Rede K-Associados e quatro algoritmos de classificação aplicados baseado nesse modelo de rede. As redes k-associados são formadas respeitando a distribuição dos objetos no espaço, ligando objetos próximos (similares) que possuem a mesma classe.

O modelo de rede k-associados e suas propriedades foram desenvolvidas em conjunto com os pesquisadores Alneu de Andrade Lopes, João Bertini e Liang Zhao, todos do mesmo grupo de pesquisa do qual participo.

Além das contribuições citadas, as técnicas estudadas foram implementadas em um sistema denominado *ComplexNet*, contemplando procedimentos de preparação dos dados, cálculo de propriedades de redes complexas e de vértices, técnicas de classificação proposicionais e relacionais, e métodos de visualização dos dados. Tal sistema vem sendo constantemente incrementado e utilizado por outros alunos do grupo de pesquisa.

Ressaltamos também, que as redes construídas possibilitam uma visualização alternativa dos conjuntos de dados, viabilizando o uso de técnicas de visualização em um processo de mineração visual de dados modelados em grafos.

1.5 Organização da monografia

O restante deste trabalho está organizada da seguinte forma: no Capítulo 2 é apresentado o embasamento teórico, nele são apresentados os conceitos de aprendizado de máquina, representação de dados, redes complexas e classificadores baseados em redes. No Capítulo 3 é descrita a proposta desta investigação relativa ao uso de Redes Complexas na tarefa de classificação, bem como a descrição dos algoritmos utilizados. No Capítulo 4 são especificados os procedimentos para avaliação experimental e os resultados alcançados são apresentados e discutidos. E por último, no Capítulo 5, são apresentadas as conclusões e os trabalhos futuros.

Capítulo

2

Aprendizado de Máquina Relacional e Redes Complexas

Nos últimos anos, foi desenvolvida uma grande variedade de técnicas de aprendizado de máquina, possuindo um limitador na perspectiva da representação do conhecimento utilizado. A maioria dessas técnicas trabalha apenas com representação proposicional, com dados representados em tabelas atributo-valor, não apropriada para domínios envolvendo relações entre seus objetos (Raedt, 2008).

Dessa forma, a representação da relação entre objetos com a utilização da modelagem em grafos pode, juntamente com o formalismo de redes complexas, apresentar um ganho substancial nas tarefas de aprendizado de máquina.

2.1 Aprendizado de Máquina

Como já comentado, aprendizado de máquina utiliza técnicas computacionais para obtenção automática de conhecimento a partir de exemplos. Para viabilizar o uso das tarefas de aprendizado é preciso representar as observações (conhecidas como exemplos ou instâncias) de um determinado domínio em termos computacionais. Em geral, o exemplo é descrito por um vetor de valores (um para cada atributo que descreve o exemplo) e eventualmente pelo rótulo da classe associada. Uma classe pode ser, por exemplo, em um conjunto de artigos científicos, a área de pesquisa referente ao artigo, e os exemplos podem ou não ser rotulados.

O aprendizado indutivo pode ser dividido em três principais abordagens, que se diferenciam pela forma com que utilizam a informação do rótulo dos dados. São elas o aprendi-

zado supervisionado, aprendizado não-supervisionado e aprendizado semi-supervisionado.

O aprendizado supervisionado é utilizado quando há um número expressivo de exemplos rotulados. O seu objetivo é induzir conceitos a partir destes exemplos, predizendo a classe de novos exemplos (exemplos de teste) tendo em vista os exemplos conhecidos rotulados (exemplos de treino). Para classes que possuem valores discretos o problema é conhecido como *classificação* e classes que possuem valores contínuos o problema é conhecido como *regressão*.

Quando os exemplos não possuem rótulos (ou o usuário não necessariamente deseja utilizá-los) é então aplicado o aprendizado não-supervisionado. Nesse, busca-se adquirir conhecimento baseado em regularidades encontradas nos dados. Em geral, é necessária uma análise posterior dos padrões e regularidades encontrados. Duas das principais técnicas de aprendizado não-supervisionado são regras de associação (Agrawal et al., 1996) e *clustering* (Jain et al., 1999). Regras de associação são utilizadas para identificar elementos que ocorrem em comum dentro de seu conjunto de dados, e *clustering* efetua agrupamentos de objetos segundo algum critério de semelhança.

O aprendizado semi-supervisionado é utilizado quando existem poucos exemplos rotulados. Assim como no aprendizado supervisionado o interesse também é induzir conceitos a partir dos exemplos, porém, é necessária uma atenção especial devido a baixa quantidade de exemplos rotulados. Um dos principais algoritmos para aprendizado semi-supervisionado é o *Co-training* (Blum & Mitchell, 1998), no qual são utilizadas duas visões independentes dos dados para induzir dois classificadores diferentes (utilizando um ou dois algoritmos de aprendizado supervisionado). Com isso um novo exemplo classificado por um dos classificadores, pode ser incorporado ao conjunto de exemplos rotulados para treinar o outro classificador. O processo é iterativo.

De forma semelhante ao *Co-training*, buscando melhorar o resultado na precisão dos classificadores, mas também para aprendizado supervisionado, técnicas de combinações de classificadores são utilizadas, chamadas comitê de classificadores (Zheng, 1998). Comitê de classificadores utiliza um conjunto de classificadores para predizer o rótulo de um novo exemplo, computando, por exemplo, a quantidade de vezes que cada rótulo foi escolhido ou a soma da probabilidade de cada rótulo (no caso de classificadores probabilísticos), classificando o exemplo com o rótulo majoritário, escolhido por votação, ou de maior probabilidade.

A escolha do sistema de aprendizado ideal depende principalmente do conjunto de dados a ser analisado. A seguir são descritas as principais representações dos dados, que tem papel fundamental na otimização do aprendizado.

2.2 Representação dos dados

A escolha da representação correta dos dados para modelar o problema é uma tarefa fundamental em inteligência artificial. De Raedt (2008) propõe uma hierarquia das principais representações, incluindo booleana, atributo-valor, multi-instância e relacional. De Raedt (2008) situa a representação em grafo entre a atributo-valor e a relacional, em termos de expressividade. A seguir essas representações são apresentadas, tendo como base o livro *Logical and Relational Learning* de Luc De Raedt (2008).

Na representação booleana, também conhecida como *item-sets*, há um conjunto de variáveis para as quais um exemplo possui os valores *verdadeiro* ou *falso*. Uma das mais populares tarefas de mineração de dados envolvendo dados booleanos é a análise de compras em um supermercado. Assumindo que temos um conjunto de produtos $I = \{salsicha$, $cerveja$, $vinho$, $mostarda\}$, um cliente pode comprar somente os produtos $salsicha$, $cerveja$ e $mostarda$, correspondendo ao *item-set* $\{salsicha$, $cerveja$, $mostarda\}$.

Outra linguagem de descrição bastante simples e muito popular é a tabela atributo-valor. Um exemplo tradicional devido a Quinlan (1986) é apresentado na Tabela 2.1, com informações sobre situações climáticas nas quais a classe representa se o tempo está bom (positivo) ou ruim (negativo) para praticar tênis. Nesta tabela, cada linha (ou tupla) corresponde a um exemplo e cada coluna a um atributo. Na representação atributo-valor um exemplo possui um único valor para cada atributo, definindo a representação como *tupla única* e *tabela única*.

Aparência	Temperatura	Umidade	Vento	Classe
sol	quente	alta	não	negativo
sol	quente	alta	sim	negativo
nublado	quente	alta	não	positivo
chuva	média	alta	não	positivo
chuva	fria	normal	não	positivo
chuva	fria	normal	sim	negativo
nublado	fria	normal	sim	positivo
sol	média	alta	não	negativo
sol	fria	normal	não	positivo
chuva	média	normal	não	positivo
sol	média	normal	sim	positivo
nublado	média	alta	sim	positivo
nublado	quente	normal	não	positivo
chuva	média	alta	sim	negativo

Tabela 2.1: Conjunto de dados *praticar tênis*, adaptado de Quinlan (1986).

Uma extensão da representação atributo-valor é a possibilidade de um exemplo possuir mais de um valor para um mesmo atributo, chamada representação multi-instância. Esta representação é definida como *multi-tuplas*, mas continua como *tabela única*.

Um exemplo desta representação é mostrado na Tabela 2.2. Suponha uma doença em que algumas famílias são portadores (positivo) e outras não (negativo), porém, nem todas as pessoas de uma família de portadores são necessariamente portadores, mas todos de uma família de não portadores são não portadores. É realizada uma série de testes genéticos em diversas famílias nas quais já se sabe se é uma família de portadores ou não, mas estes testes não são suficientes para saber qual pessoa possui a doença. Portanto, há um conjunto de grupos de tuplas (famílias nas quais cada tupla é uma pessoa), com cada grupo sendo rotulado em positivo ou negativo. Estas informações serão utilizadas para tentar identificar, por exemplo, qual tupla existente em uma nova família examinada indica que a família é portadora, e qual tupla existente indica que a família é não portadora.

Exemplo	Gene1	Gene2	Gene3	Gene4	Classe
exemplo 1	aa	aa	aa	bb	negativo
	aa	aa	aa	aa	
exemplo 2	bb	aa	aa	bb	positivo
	ab	bb	aa	bb	
	ab	ab	bb	bb	
	ab	ab	bb	aa	
exemplo 3	bb	ab	bb	aa	negativo
	aa	bb	aa	bb	
	aa	ab	bb	bb	
exemplo 4	ab	bb	bb	bb	positivo
	aa	bb	bb	aa	
	bb	bb	aa	aa	
	bb	aa	bb	bb	

Tabela 2.2: Conjunto de dados de testes genéticos por família, adaptado de Raedt (2008).

Seguindo o exemplo da Tabela 2.2 verifica-se que famílias que possuem as tuplas *(aa, aa, aa, bb)* e *(aa, aa, aa, aa)* são famílias não portadores, e as famílias que possuem uma tupla com valores *ab* para o *Gene1* e *bb* para o *Gene2* indicam famílias de portadores. Observa-se que, se consideradas as tuplas individualmente, essas informações não poderiam ser obtidas, pois não seriam todos os valores da mesma classe que iriam satisfazer a afirmação.

Da representação multi-instância para a representação relacional há apenas a diferença que esta última trabalha com *múltiplas tabelas* ou *relações entre os exemplos*. Muito importante devido a grande quantidade de informação armazenada em banco de dados relacionais.

Na Figura 2.1 é apresentado um diagrama Entidade-Relacionamento (DER) de banco de dados de publicações científicas, demonstrando as relações existentes entre os dados, destacando que um mesmo artigo pode citar e pode ser citado por muitos artigos, e um autor pode participar de muitos artigos assim como um artigo ter muitos autores. Na

Tabela 2.3 são representadas as tabelas com informações dos artigos e dos autores, e uma tabela de relação, representando co-autoria.

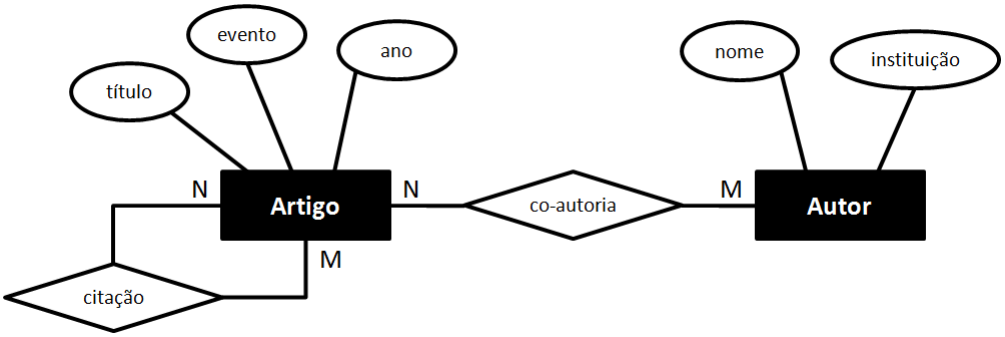


Figura 2.1: Exemplo de diagrama Entidade-Relacionamento de um de banco de dados de publicações científicas.

título do artigo	evento	ano
Finding community structure in very large networks	*Physical Review E*	2004
Community structure in social and biological networks	*PNAS*	2002
The Power of Choice in Network Growth	*European Journal of Physics B*	2007
Explosive Percolation in Random Networks	*Science*	2009

(a) publicações

autor	instituição
Aaron Clauset	*University of New Mexico*
Mark Newman	*University of Michigan*
Cristopher Moore	*University of New Mexico*
Michelle Girvan	*University of Michigan*
Raissa D'Souza	*University of California*
Paul Krapivsky	*Boston University*
Dimitris Achlioptas	*University of California*
Joel Spencer	*New York University*

(b) autores

título do artigo	autor
Finding community structure in very large networks	Aaron Clauset
Finding community structure in very large networks	Mark Newman
Finding community structure in very large networks	Cristopher Moore
Community structure in social and biological networks	Michelle Girvan
Community structure in social and biological networks	Mark Newman
The Power of Choice in Network Growth	Raissa D'Souza
The Power of Choice in Network Growth	Paul Krapivsky
The Power of Choice in Network Growth	Cristopher Moore
Explosive Percolation in Random Networks	Dimitris Achlioptas
Explosive Percolation in Random Networks	Raissa D'Souza
Explosive Percolation in Random Networks	Joel Spencer

(c) co-autoria

Tabela 2.3: Tabelas exemplificando um conjunto de dados de publicações científicas.

Um caso particular de representação relacional são os grafos, o qual tem recebido grande atenção em tarefas de mineração de dados. Washio et al. (2004) descrevem como

motivação para o uso de grafo o fato de serem mais expressivos que representações proposicionais e potencialmente mais eficientes que técnicas de mineração e aprendizado relacional.

Considerando o conjunto de dados apresentado na Tabela 2.3, é possível realizar a modelagem em grafo das relações de co-autoria (Figura 2.2, com os vértices representando os autores e as arestas participações conjuntas em algum artigo.

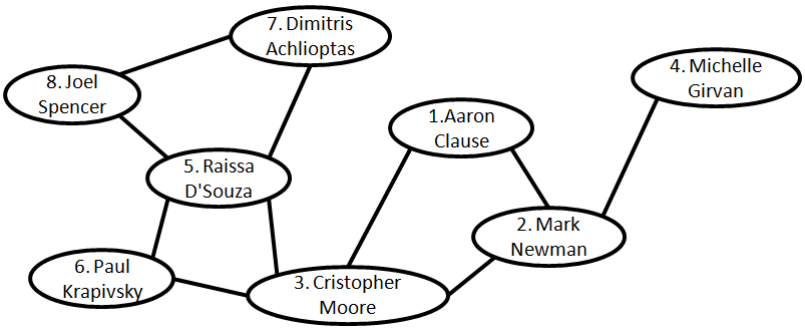


Figura 2.2: Exemplo de uma rede de co-autoria em artigos científicos representada em grafo, no qual os vértices são os autores e as arestas ligam autores que trabalharam juntos em um artigo.

Outros exemplos de domínios relacionais possíveis de serem modelados por grafos, são as páginas *webs*, podendo ser representadas com as páginas sendo os vértices e os *hyperlinks* as arestas, consumidores e produtos em lojas que realizam vendas pela Internet, e redes de proteínas em biologia computacional.

Em termos de sua representação computacional, os grafos geralmente são armazenados em listas ou matrizes de adjacência. No primeiro caso são armazenados apenas os pares de vértices (v_i, v_j) que possuem ligações, de forma que é gerada uma lista L com tamanho igual ao número de vértices, com cada item i de L sendo outra lista que contém todos os vértices adjacentes ao vértice v_i, porém, no caso de uma rede direcionada, somente possui os vértices em que o vértice v_i possui uma ligação direcionada a eles, ou seja, ligações nas quais o vértice v_i é a origem.

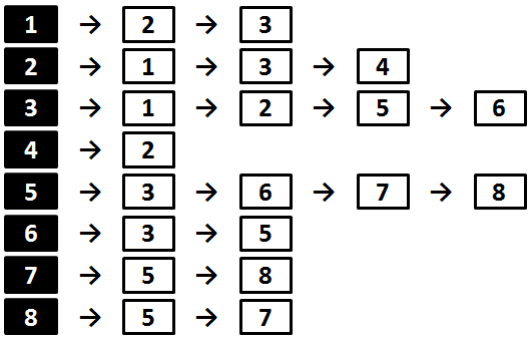
Já no caso de matriz de adjacência, é criada uma matriz A com número de linhas e colunas igual ao número de vértices e, no caso de uma rede não direcionada, se dois vértices v_i e v_j estão ligados as entradas a_{ij} e a_{ji} na matriz são iguais a 1, caso contrário são iguais a 0. No caso de uma rede direcionada, se o vértice v_i possui uma ligação direcionada ao vértice v_j então somente a entrada a_{ij} é igual a 1, caso contrário é igual a 0.

Ambas representações computacionais são apresentadas na Figura 2.3. Qualquer informação adicional do grafo, como por exemplo, o peso das arestas ou características individuais dos vértices, deve possuir uma estrutura própria ou ser adaptada à estrutura existente. Para grafos com arestas ponderadas, comumente os pesos das arestas substituem os valores binários na matriz de adjacência, com o 0 representando a ausência de

ligação.

	1	2	3	4	5	6	7	8
1	0	1	1	0	0	0	0	0
2	1	0	1	1	0	0	0	0
3	1	1	0	0	1	1	0	0
4	0	1	0	0	0	0	0	0
5	0	0	1	0	0	1	1	1
6	0	0	1	0	1	0	0	0
7	0	0	0	0	1	0	0	1
8	0	0	0	0	1	0	1	0

(a)



(b)

Figura 2.3: Representação computacional para o grafo da Figura 2.2 em forma de (a) matriz adjacência, com a primeira linha e primeira coluna indicando o vértice, e (b) lista adjacência.

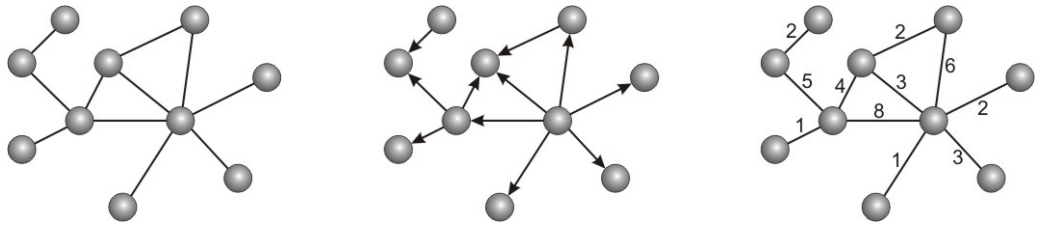
2.3 Redes Complexas

A pesquisa em Redes Complexas é naturalmente multidisciplinar e engloba conceitos de teoria dos grafos, estatística e sistemas complexos para apoiar a caracterização, análise e modelagem dos mais variados fenômenos. De fato, qualquer fenômeno formado por muitas partes que interagem entre si pode ser representado por redes.

Uma rede pode ser definida como um conjunto de itens conectados por relações existentes entre eles, podendo ser representada por grafos, nos quais itens são os vértices e suas conexões são as arestas. Formalmente, uma rede $R = (V, E)$ contém um conjunto de N vértices, $V = \{v_1, v_2, ..., v_N\}$, e um conjunto de M arestas, $E = \{e_1, e_2, ..., e_M\}$.

Tanto os vértices quanto as arestas podem carregar um peso, que quantifica a relação, podendo ser, por exemplo, a quantidade de vezes que dois autores trabalharam juntos, no caso de uma rede de co-autoria. Quando possui peso, além da rede ser formada pelos conjuntos V e E, a rede possui ainda um conjunto $W = \{w_1, w_2, ..., w_M\}$ representando o peso de cada aresta, e a rede passa a ser $R = (V, E, W)$. As conexões podem também ser direcionadas, e a rede ser cíclica ou acíclica de acordo com ela possuir caminhos fechados ou não. Um caminho entre dois vértices é uma sequência de vértices e arestas que ligam um vértice ao outro. Caminho fechado é um caminho no qual o vértice origem e destino é o mesmo, e caminho geodésico é o menor caminho possível entre um par de vértices, sendo o tamanho do caminho a quantidade de arestas existentes entre eles. A Figura 2.4 apresenta três exemplos de rede.

Segundo Newman (2003), a representação de sistemas em redes propicia a exploração visual dos dados, a qual, em geral, é suficiente para analisar as informações. Porém com



(a) (b) (c)

Figura 2.4: Três diferentes redes: (a) rede não direcionada sem peso, (b) rede direcionada sem peso, e (c) rede não direcionada com peso.

o crescimento da capacidade computacional, e o processamento de grandes quantidades de informações, tornam-se mais complexas ou impraticáveis as análises visuais desses sistemas modelados via rede, pois podem chegar a milhões de vértices. Com isso, surgiu a necessidade do desenvolvimento de diversos métodos para análise das propriedades de redes. Como exemplo de redes complexas pode-se citar a Internet, apresentada na Figura 2.5 (gerada pelo *The Opte Project*[1]), na qual os vértices são roteadores e as arestas são ligações entre roteadores.

A seguir são descritos os modelos e detalhadas algumas propriedades de redes complexas.

2.3.1 Modelos de Redes Complexas

Sistemas podem ser categorizados, em um contexto de redes complexas, a partir de suas características estruturais e dinâmicas, provendo uma maior compreensão de seu comportamento.

Uma das primeiras tentativas de se construir um modelo para grandes redes foi realizada pelos matemáticos Paul Erdös e Alfred Rényi, baseando-se em ligações aleatórias entre vértices, que ficou conhecido como grafos aleatórios de Erdös e Rényi (Erdös & Rényi, 1959, 1960, 1961). Grafos aleatórios são aqueles construídos iniciando-se com um conjunto de N vértices completamente desconectados e a cada passo dois vértices são escolhidos aleatoriamente e conectados com uma probabilidade p, sendo cada par de vértice considerado apenas uma vez. Quando N é grande e p é mantido constante para todos vértices, a distribuição de conectividade tende à distribuição de Poisson (Figura 2.6).

Porém, verificou-se que na maioria das redes reais, a presença de caminhos fechados formando triângulos é muito maior do que nas redes aleatórias com o mesmo número de vértices e arestas (Watts & Strogatz, 1998), sendo o primeiro indício de que redes reais não são completamente aleatórias, mas que são estocásticas, possuindo uma determinada lei de formação. Com isso, Watts e Strogatz sugeriram um novo modelo chamado modelo mundo pequeno (*small world*) de Watts-Strogatz, apresentando o efeito mundo pequeno,

[1] http://www.opte.org/

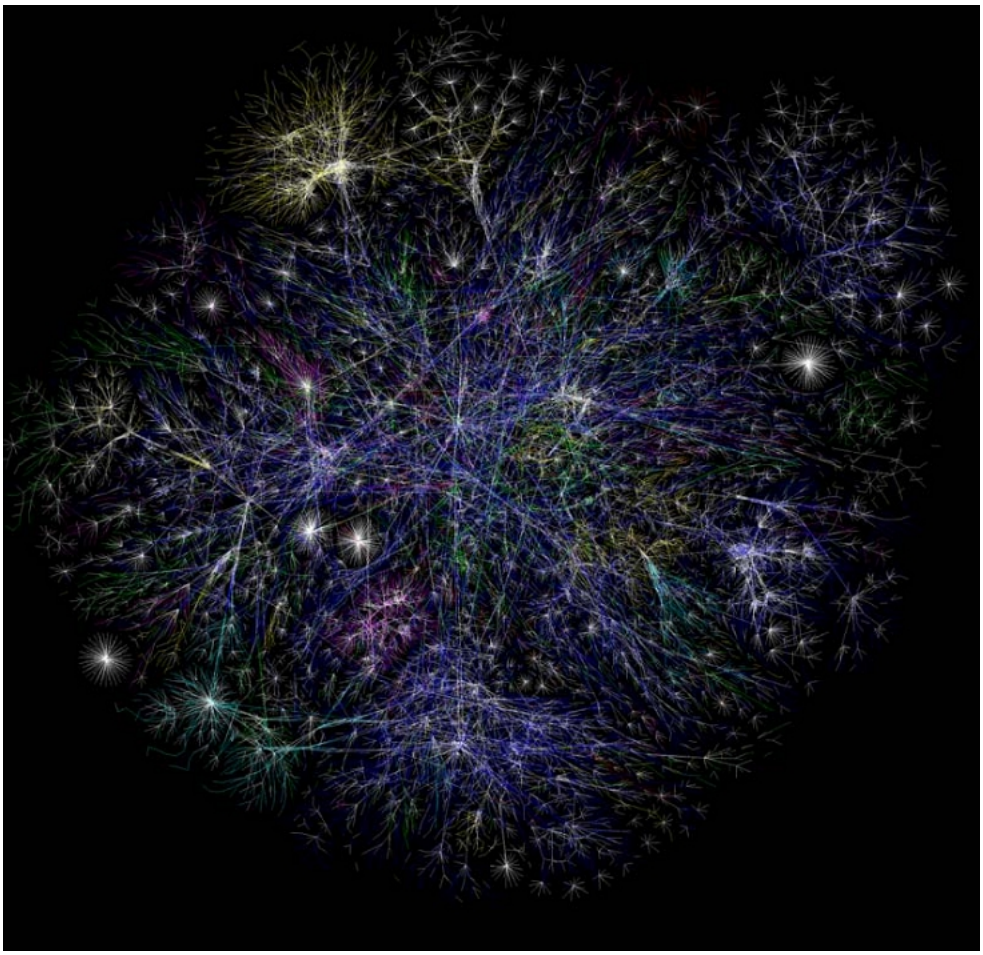
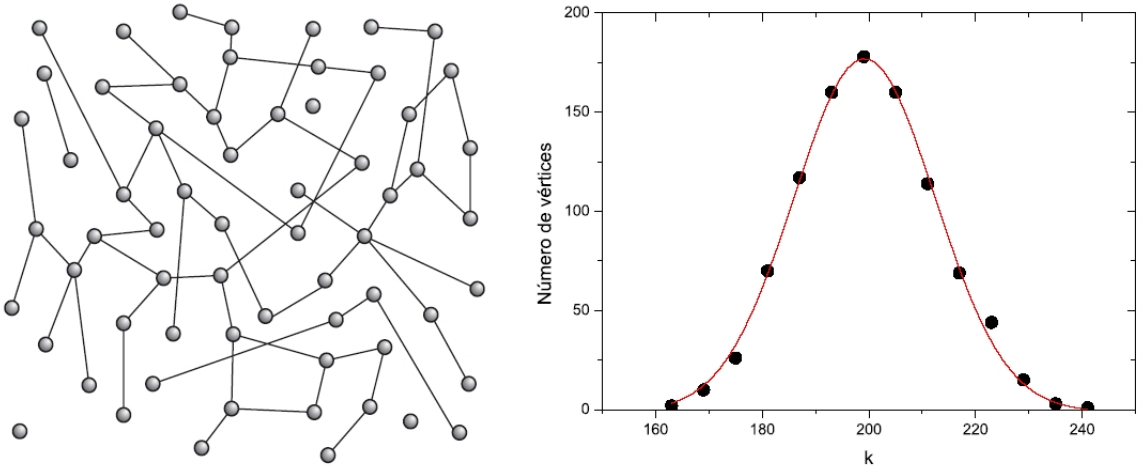


Figura 2.5: Rede de Internet gerada em 15 de janeiro de 2005 pelo The Opte Project (http://www.opte.org). As cores indicam os domínios: (i) net, ca, us (azul), (ii) com, org (verde), (iii) mil, gov, edu (vermelho), (iv) jp, cn, tw, au (amarelo), (v) de, uk, it, pl, fr (rosa escuro), (vi) br, kr, nl (azul claro) e (vii) desconhecido (branco)

no qual vértices podem ser alcançados de outros vértices por um caminho com um pequeno número de arestas, e a grande presença de caminhos fechados formando triângulos. Sua formação se inicia com um rede regular de N vértices ligados aos k vizinhos mais próximos em cada direção (totalizando $2k$ conexões iniciais por vértice), em seguida as conexões são aleatoriamente reconectadas com uma probabilidade fixa p. Quando a probabilidade p for igual a 0 a rede é completamente regular, quando p for igual a 1 a rede é completamente aleatória, portanto o modelo está entre as duas situações. A Figura 2.7 apresenta um exemplo de formação da rede e a Figura 2.8 apresenta uma rede mundo pequeno com sua distribuição do grau.

Ao analisar a rede mundial de computadores (*World Wide Web*), os pesquisadores Albert-László Barabási e Reka Albert verificaram a presença do fenômeno mundo pequeno, mas verificaram também que a distribuição de conexões não é aleatória. Na *web* e em diversas redes reais a distribuição de conexões tem um decaimento seguindo uma lei de potência, tal distribuição é chamada livre de escala (Barabási & Albert, 1999). Com



(a) (b)

Figura 2.6: (a) Um exemplo de um grafo aleatório de Erdös e Rényi. (b) A distribuição da conectividade para uma rede com 10.000 vértices, usando uma probabilidade $p = 0{,}2$. Cada ponto no gráfico é a média sobre 10 redes. Figura de Costa et al. (2007).

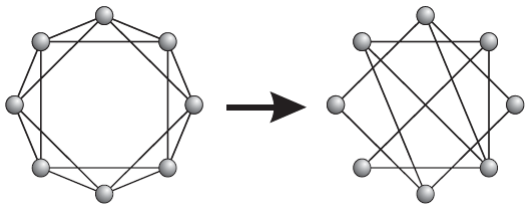
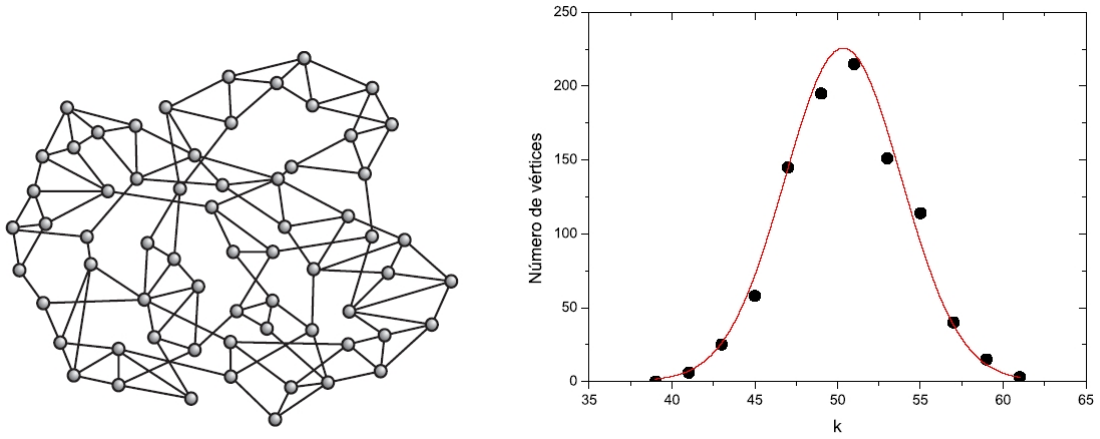


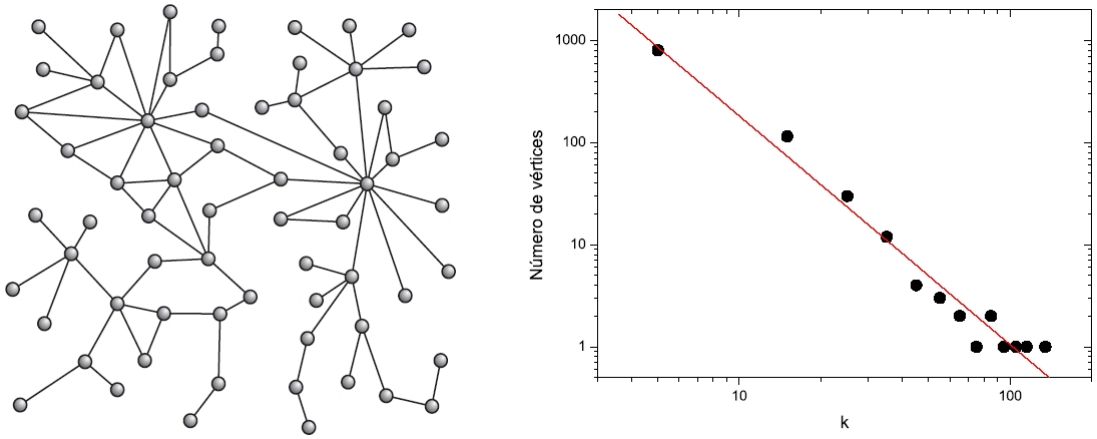
Figura 2.7: Exemplificação de construção de uma rede mundo pequeno de Watts e Strogatz, as quais são construídas a partir de uma rede regular, reconectando as arestas com uma probabilidade p. Figura de Costa et al. (2007).



(a) (b)

Figura 2.8: (a) Um exemplo de uma rede mundo pequeno formada por 64 vértices. Nota-se a presença de um elevado número de caminhos fechados de ordem três. (b) A distribuição da conectividade para uma rede mundo pequeno formada por 1.000 vértices, $k = 25$ e $p = 0{,}3$. Cada ponto é uma média sobre 10 redes. Figura de Costa et al. (2007).

isso reforça-se ainda mais a idéia de que o universo aleatório de Erdös e Rényi tende a não estar presente na natureza, pois o modelo de Watts e Strogatz ainda mantem um caráter aleatório. Já o modelo de Barabási e Albert descarta a aleatoriedade e mostra que há leis que regem a estrutura das redes naturais (Figura 2.9). Barabási & Albert (1999) propuseram um modelo de formação da rede livre de escala que se inicia com uma pequena quantidade N_0 de vértices, e a cada passo é inserido um vértice com g ($g \leq N_0$) arestas que se conectam aos vértices já presentes, dando preferência aos vértices mais conectados.



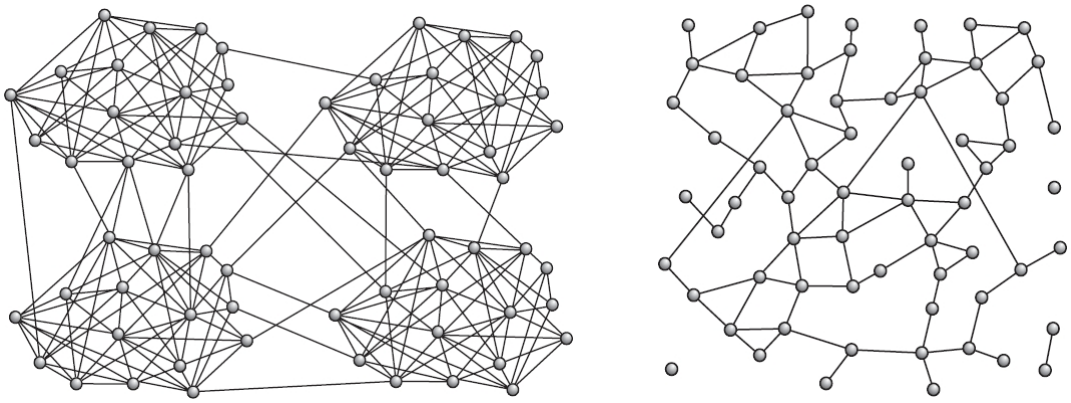
(a) (b)

Figura 2.9: (a) Exemplo de uma rede gerada pelo modelo livre de escala de Barabási e Albert. (b) Distribuição das conexões para uma rede livre de escala formada por 10.000 vértices considerando m = 5. Cada ponto é uma média sobre 10 redes e os eixos estão em escala logarítmica. Figura de Costa et al. (2007).

Algumas redes reais, como redes sociais e biológicas, apresentam grupos de vértices com muitas conexões entre si e com poucas ligações a vértices de outros grupos, este é o modelo de rede com estrutura de comunidade ou redes modulares. Um modelo para formação de redes com esta propriedade é proposto por Girvan & Newman (2002), no qual, inicialmente, um conjunto de N vértices é classificado em c comunidades, em seguida dois vértices são selecionados e conectados com uma probabilidade p_{in} ou p_{out}, de acordo com a comunidade inicialmente definida para cada vértice, caso os vértices estejam na mesma comunidade utiliza-se p_{in}, e quando os vértices estão em comunidades distintas utiliza-se p_{out}. Quando $p_{out} \ll p_{in}$ as comunidades são facilmente identificas, e o contrário ocorre quando $p_{out} \approx p_{in}$. A Figura 2.10 apresenta um modelo de rede com estrutura de comunidade.

Nos modelos de Redes Complexas vistos até agora a posição dos vértices não possui um significado particular, sendo um espaço abstrato. Porém, há redes nas quais a posição dos vértices tem importância em sua estrutura, como, por exemplo, redes de rodovias ou até mesmo a Internet, nas quais as posições das cidades e dos roteadores são importantes, pois têm relações com entidades físicas. Essas redes são chamadas redes geográficas.

Para modelar essas redes, Waxman (1988) propôs um modelo no qual distribui-se N vértices aleatoriamente em um espaço bidimensional e as ligações são inseridas com uma probabilidade que decai com a distância Euclidiana entre eles (Figura 2.10). Este modelo gera distribuição de conexões similiar ao modelo aleatório de Erdös e Rényi.



(a) (b)

Figura 2.10: (a) Exemplo de uma rede com estrutura de comunidades formada por 64 vértices com 4 comunidades. (b) Exemplo de rede geográfica formada por 64 vértices. Figura de Costa et al. (2007).

2.3.2 Propriedades de Redes Complexas

As propriedades de redes complexas são utilizadas para a extração de informação e identificação de comportamentos, também auxiliando na definição do modelo da rede. A seguir são detalhadas algumas propriedades individuais e globais relativas a redes complexas, como sugeridas por Liu et al. (2005). As propriedades individuais são aquelas relacionadas com cada vértice particular, muitas vezes representam relevância ou centralidade do vértice na rede; e as propriedades globais procuram descrever as características da rede como um todo, como por exemplo, o diâmetro da rede, a distribuição do grau e a média do menor caminho, colaborando também na identificação de seu modelo e seu comportamento.

Porém, antes de apresentarmos as propriedades é necessário compreender alguns conceitos de grafos relacionados com a conectividade dos vértices. O grau (g) de um vértice representa o número de ligações que o vértice possui. Após gerar uma matriz de adjacência A, o grau g_i de um vértice v_i é obtido pela Equação 2.1. Em uma rede direcionada o grau é o número de arestas em que o vértice é a origem. Já para as redes nas quais as conexões possuem um peso, o grau do vértice é substituído pela força (s), computada de forma semelhante, porém utilizando-se do peso das ligações (Equação 2.2).

$$g_i = \sum_{j=1}^{N} A_{ij} \tag{2.1}$$

$$s_i = \sum_{j=1}^{N} w_{ij} \tag{2.2}$$

Considerando o grau de todos os vértices é possível obter o grau máximo e mínimo da rede e outras duas propriedades que são a distribuição do grau e a média do grau.

A distribuição do grau (também chamada distribuição de conectividade) é fundamental para a definição do modelo de uma rede complexa. Para gerá-la é preciso obter o grau de cada vértice e criar um gráfico de frequência de vértices que possuem cada valor de grau, com o grau variando de zero ao grau máximo, exemplos podem ser vistos nas Figuras 2.6, 2.8 e 2.9. Sendo assim, é possível usá-la para comparar com os modelos de redes já citados e identificar seu modelo.

A média do grau $\langle g \rangle$ é obtida a partir dos graus de cada vértice da rede, descrita pela Equação 2.3. Com isso, é possível identificar vértices na rede, com alto grau em relação a média, que são conhecidos como *hubs*. Estes vértices, além de seu destaque individual, têm importante papel na formação da estrutura das redes complexas, representando importantes conexões.

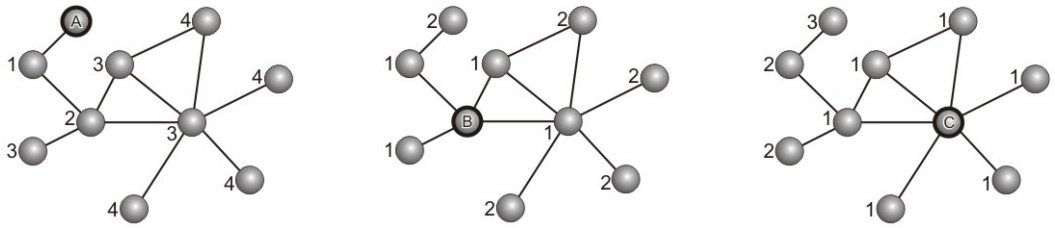
$$\langle g \rangle = \frac{1}{N} \sum_{i=1}^{N} g_i \tag{2.3}$$

Entretanto, um vértice pode possuir um alto grau, mas ser parte de um grupo isolado, sendo que embora localmente ele seja bem conectado, globalmente ele pode não ser. Para isso há a medida chamada grau de proximidade (*closeness*) (Wasserman & Faust, 1994), a qual explora a relação de um vértice com todos os vértices restantes na rede. O grau de proximidade de um vértice v_i, expresso por c_i, é o inverso da média dos caminhos geodésicos para todos os outros vértices da rede, como mostrado na Equação 2.4, sendo N o número de vértices que possuem um caminho possível para o vértice v_i, não contendo o próprio vértice. Quanto maior o valor do grau de proximidade, como o próprio nome diz, menor a distância, em média, do vértice v_i para os outros vértices da rede, determinando vértices centrais na rede.

$$c_i = \frac{N}{\sum_{j=1}^{N} tamanho\_caminho\_geodesico_{ij}} \tag{2.4}$$

Na Figura 2.11 estão ilustrados todos os caminhos geodésico considerando a origem nos vértices A, B e C, gerando assim o grau de proximidade para os três vértices, no qual o vértice A obteve 0,32; o vértice B, 0,64; e o vértice C, 0,69.

Outra medida que indica centralidade de vértices na rede é o grau de intermediação (*betweenness*) (Freeman, 1977), para sua obtenção também se utiliza o caminho geodésico. O grau de intermediação b_i de um vértice v_i é o número de caminhos geodésicos que possuem o vértice v_i no percurso, para isso é preciso gerar todos os caminhos geodésicos



(a) (b) (c)

Figura 2.11: Geração de todos caminhos geodésicos considerando a origem em três vértices, para obtenção do grau de proximidade.

possíveis. Caso entre um par de vértices quaisquer v_a e v_b haja mais de um caminho geodésico então o grau de intermediação do vértice v_i deve ser incrementado em 1 se ele participar de todos os caminhos geodésicos possíveis entre v_a e v_b, ou deverá ser incrementado apenas de uma fração do número de caminhos geodésicos em que participa. A Equação 2.5 descreve o grau de intermediação para o vértice v_i, em um conjunto V de vértices, sendo $qtd\_caminhos\_geodesicos_{ab}$ o número de caminhos geodésicos entre o vértice v_a e o vértice v_b e $qtd\_caminhos\_geodesicos_{ab}(i)$ a quantidade destes caminhos que passam pelo vértice v_i. Vértices com altos valores de grau de intermediação podem representar vértices de ligação entre grupos de vértices com fortes conexões internas.

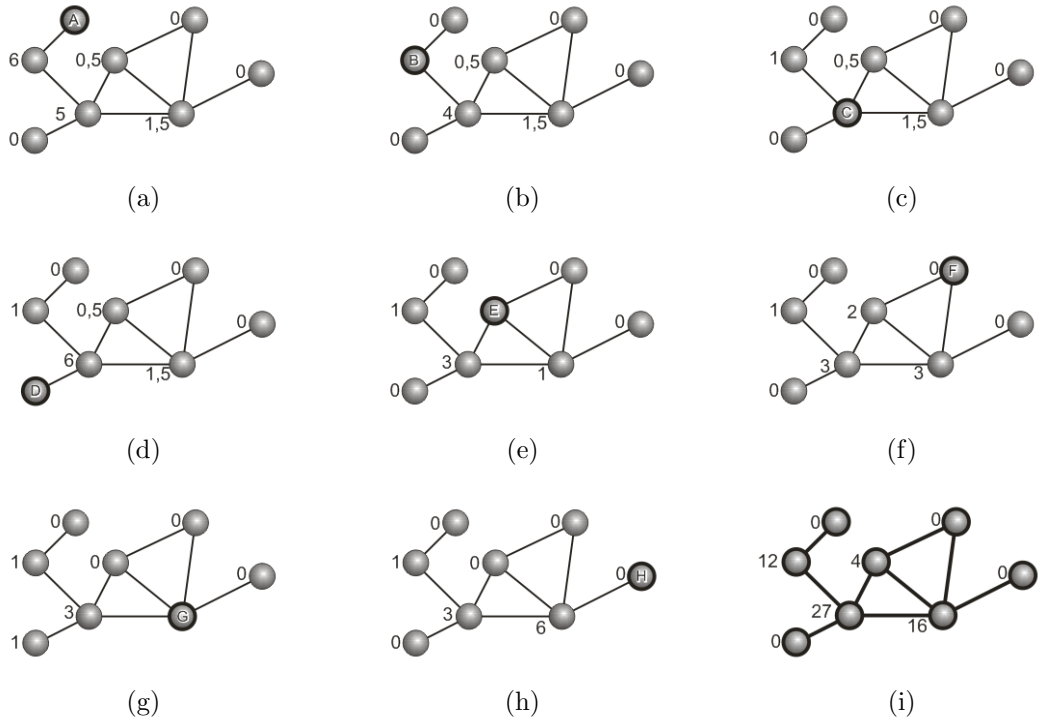
$$b_i = \sum_{a \neq b \neq i \in V} \frac{qtd\_caminhos\_geodesicos_{ab}(i)}{qtd\_caminhos\_geodesicos_{ab}} \tag{2.5}$$

Um ilustração da geração do grau de intermediação pode ser observada na Figura 2.12, sendo que primeiramente foi obtido o grau de intermediação considerando caminhos com origem em um único vértice, somente para exemplificar, utilizando-se 5 dos 10 vértices da rede. E, por fim, é apresentada a soma de todos os graus de intermediação considerando a rede inteira.

Tanto o grau de proximidade quanto o grau de intermediação são baseados nos caminhos geodésicos da rede, sendo assim é necessário que a rede seja conexa, isto é, que haja um caminho válido entre todos pares de vértices na rede, ou que os componentes sejam analisados individualmente. Um componente é um subconjunto conexo do grafo, isso é, sempre há um caminho entre quaisquer dois vértices deste subconjunto, e não há caminho de qualquer vértice pertencente ao subconjunto para um vértice não pertencente. Devido a possibilidade de haver vários componentes, é muito usual a aplicação das propriedades somente nos componentes maiores, na Figura 2.13 há um exemplo simples de componentes.

Outras duas propriedades que podem ser consideradas para redes conexas (ou individualmente para cada componente de uma rede não conexa) são o coeficiente de agrupamento (Barrat & Weigt, 2000) e o afunilamento (*funneling*) (Newman, 2001).

O coeficiente de agrupamento de um vértice v_i expressa a probabilidade de dois vértices

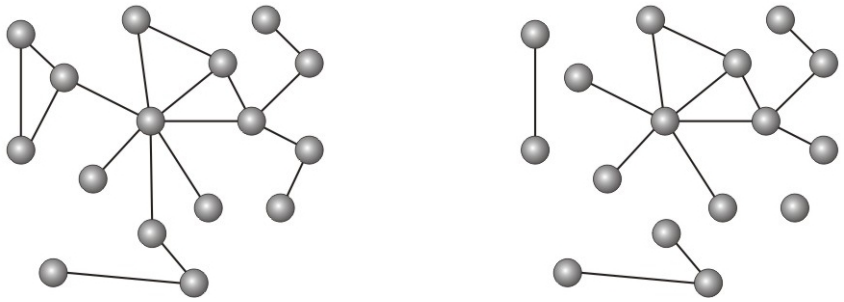


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figura 2.12: As imagens (a),(b),(c),(d),(e),(f),(g) e (h) apresentam o grau de intermediação dos vértices considerando todos caminhos geodésicos com origem nos vértices A, B, C, D, E, F, G e H, respectivamente. O grau de intermediação final de cada vértice, obtido ao efetuar a soma do grau de intermediação considerando todos caminhos geodésicos da rede, é apresentado na imagem (i).



(a) (b)

Figura 2.13: Exemplo de (a) rede conexa e (b) de rede não conexa.

estarem conectados dado que são adjacentes a v_i. Devido as redes reais apresentarem uma alta ocorrência de caminhos fechados de ordem três, formando subgrafos de três vértices totalmente conectados, essas redes têm alto coeficiente de agrupamento. Para se obter a fração de tais subgrafos de um único vértice v_i, mede-se a razão entre o número de arestas existentes entre os vizinhos do vértice v_i, denotado por e_i, e o número máximo possível de arestas entre esses vizinhos, dado por $g_i(g_i-1)/2$. Portanto, o coeficiente de agrupamento de um vértice, utilizando-se matriz de adjacência na qual a_{ij} contém o valor 1 se houver ligação entre o vértice v_i e o vértice v_j e 0 caso contrário, é calculado pela Equação 2.6,

proposta por Watts & Strogatz (1998).

$$cc_i = \frac{2e_i}{g_i(g_i - 1)} = \frac{\sum_{j=1}^{N} \sum_{m=1}^{N} a_{ij} a_{jm} a_{mi}}{g_i(g_i - 1)} \tag{2.6}$$

De maneira similar o coeficiente de agrupamento para uma rede com peso é calculado pela Equação 2.7, proposta por Barrat et al. (2004).

$$cc_i^w = \frac{1}{s_i(g_i - 1)} \sum_{j,m} \frac{w_{ij} + w_{im}}{2} a_{ij} a_{im} a_{jm} \tag{2.7}$$

Como propriedade global é preciso calcular a média do coeficiente de agrupamento entre todos os vértices da rede, na Equação 2.8 é descrito o cálculo para redes sem peso e na Equação 2.9 o cálculo para redes com peso.

$$\langle cc \rangle = \frac{1}{N} \sum_{i=1}^{N} cc_i \tag{2.8}$$

$$\langle cc^w \rangle = \frac{1}{N} \sum_{i=1}^{N} cc_i^w \tag{2.9}$$

Já a propriedade afunilamento descreve redes que possuem poucos vértices com um alto valor de grau de intermediação e a maioria com um baixo valor, formando uma espécie de afunilamento devido a muitos caminhos geodésicos terem em comum alguns poucos vértices.

Ainda relacionado a caminhos na rede, há o menor caminho médio, que indica a média de todos caminhos mínimos da rede. A média dos caminhos mínimos (l) é calculada gerando-se uma matriz de distâncias D, no qual os elementos d_{ij} contém o menor caminho entre os vértices v_i e v_j, de acordo com a Equação 2.10. Considerando o caminho mínimo entre todos pares de vértices, o diâmetro da rede é dado pelo tamanho do maior caminho mínimo. Tanto a média do menor caminho como o diâmetro devem ser obtidos utilizando redes conexas.

$$l = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{2.10}$$

Outra propriedade importante é a existência de comunidades na rede. Newman (2003) define estrutura de comunidades como uma propriedade de redes que possuem grupos de vértices nos quais as conexões são densamente distribuídas internamente e esparsamente distribuídas entre vértices de grupos distintos, ou seja, os elementos de cada grupo são fortemente conectados entre si e fracamente conectados com os elementos de outros grupos (ver Figura 2.14).

Uma importante tarefa para explorar estruturas de comunidades na rede é a definição

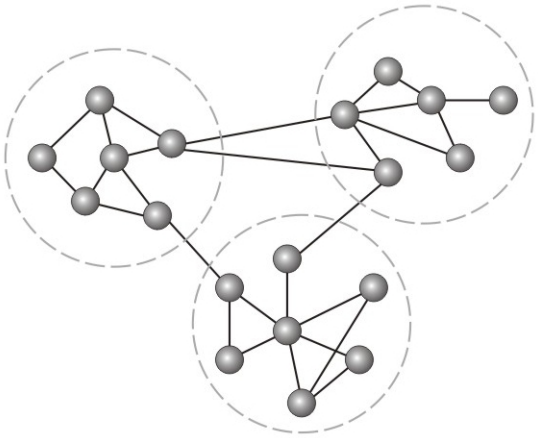


Figura 2.14: Exemplo de rede possuindo três comunidades.

do método para efetuar a divisão de comunidades. Rodrigues (2007) analisou os métodos de divisão de comunidades e sintetizou da seguinte forma: (i) espectrais, que são baseados na análise dos autovetores de matrizes derivadas na rede (Newman, 2006); (ii) divisivos, que efetuam a remoção iterativa das conexões entre as comunidades até a obtenção da maior modularidade possível (Girvan & Newman, 2002; Radicchi et al., 2004); (iii) aglomerativos, que são baseados na ideia de similaridade entre vértices da mesma comunidade (Newman, 2004a; Clauset et al., 2004); (iv) maximização de modularidade, que busca a melhor divisão de comunidade quando o maior valor de modularidade é obtido (Duch & Arenas, 2005); e (v) métodos locais, que determinam comunidades localmente, não considerando as informações globais da rede (Clauset, 2005; Bagrow & Bollt, 2005).

A seguir, dois métodos são descritos, um método divisivo propostos por Girvan & Newman (2002) e um método aglomerativo proposto por Newman (2004b).

Seguindo a ideia de que há poucas conexões entre os grupos fortemente conectados internamente, Girvan & Newman (2002) propuseram um algoritmo para detecção de comunidades baseado em remoção de arestas. Os autores utilizam para as arestas o mesmo conceito de grau de intermediação criado para os vértices, removendo sempre a de maior grau de intermediação, como ilustrado na Figura 2.15. Porém, com a remoção de uma aresta, todos os graus de intermediação devem ser recalculados, e isso torna o algoritmo muito custoso, com complexidade $O(N^2M)$, sendo M o número de arestas e N o número de vértices. Outro problema nesse método é determinar a quantidade ideal de comunidades, número este que obrigatoriamente deve ser dado como entrada.

Buscando resolver o problema de identificar a quantidade ideal de comunidades, Newman (2004a) criou uma medida para se determinar a qualidade de uma divisão particular da rede, chamada medida de modularidade, tipicamente representada por Q. Para uma rede dividida em c comunidades, Q é calculada por uma matriz simétrica E de c linhas e c colunas, na qual os elementos ao longo da diagonal principal, e_{ii}, fornecem a fração das conexões entre os vértices na mesma comunidade, e os elementos e_{ij}, com $i \neq j$, re-

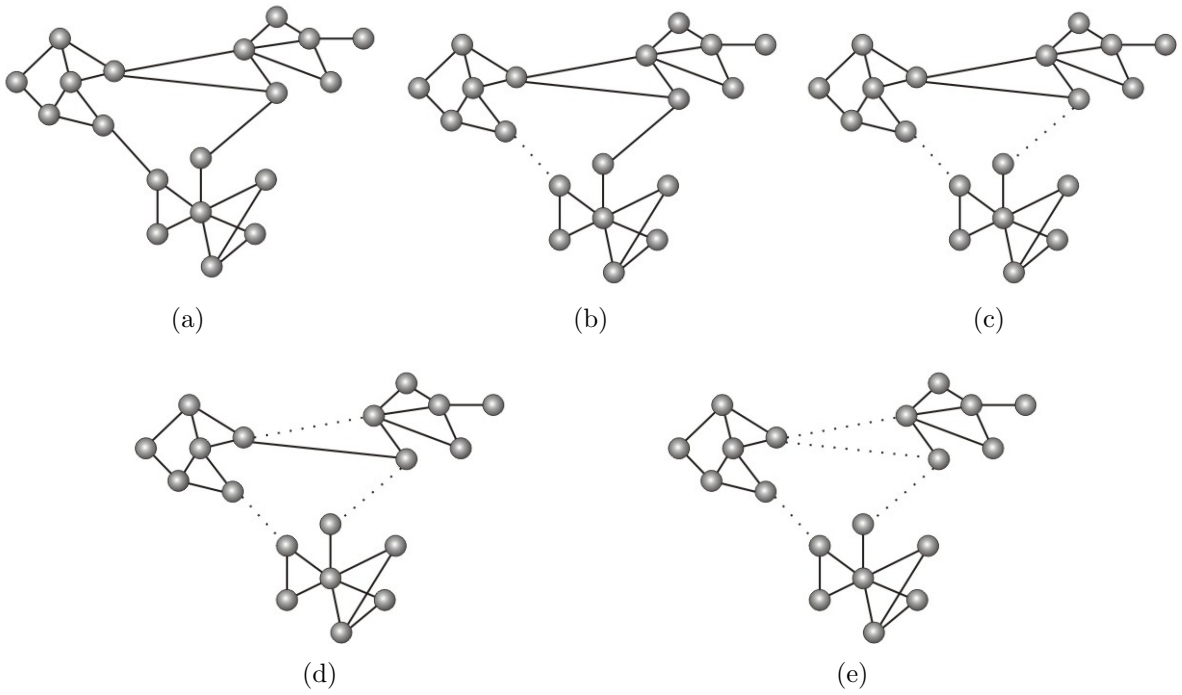


Figura 2.15: Divisão da rede em três comunidades pela remoção de arestas com maior grau de intermediação, seguindo o método de Girvan & Newman (2002).

presentam a fração de conexões entre as comunidades i e j. A modularidade Q é obtida pela Equação 2.11, quando $Q = 1$ a rede é formada por módulos desconectados e valores altos de Q são redes com estrutura modular bem definida.

$$Q = \sum_{i} [e_{ii} - (\sum_{j} e_{ij})^2] \tag{2.11}$$

Utilizando-se dessa medida de modularidade, Newman (2004b) propôs um método aglomerativo, no qual em uma rede de N vértices inicia-se sem nenhuma das conexões, representando N comunidades, e a cada iteração são escolhidos dois componentes c_i e c_j (que possuam uma conexão na rede real) cujo agrupamento forneça o maior acréscimo (ou menor decréscimo) no valor da modularidade, o agrupamento consiste na inserção de todas arestas existentes entre os dois componentes. A Equação 2.12 é utilizada para identificar o par de componentes c_i e c_j que deverá ser agrupado, sendo o par que maximizar ΔQ_{ij}. Em cada passo o algoritmo, no pior caso, tem complexidade $O(M+N)$, e devido a quantidade máxima de iterações ser N-1, pois iterativamente agrupa-se dois componentes, o algoritmo possui complexidade $O((M + N)N)$.

$$\Delta Q_{ij} = 2(e_{ij} - \sum_{j} e_{ij} \sum_{i} e_{ji}) \tag{2.12}$$

A Figura 2.16 ilustra quatro etapas do método de identificação de comunidades proposto por Newman (2004b), contendo o estado inicial, sem nenhuma aresta, dois estados

intermediários (iterações 11 e 16), com o segundo sendo o máximo valor de Q (melhor divisão de comunidades), e o estado final (iteração 18), formando um único componente. Obviamente o método seria interrompido no máximo valor de Q.

A divisão da rede que possuir o maior valor de modularidade é a melhor divisão encontrada segundo este critério. Além de sugerir a quantidade de comunidades para a rede, outra vantagem do algoritmo é seu melhor tempo de processamento em relação aos anteriores, mas mesmo assim, Wakita & Tsurumi (2007), afirmam que o algoritmo possui um limite prático suportando redes contendo até 500 mil vértices, afirmando que esta limitação é devido a junção das comunidades ocorrer de maneira desequilibrada, e sugerem três heurísticas que tentam balancear o tamanho das comunidades enquanto são criadas, afirmando melhorar o desempenho do algoritmo. Danon et al. (2006) também observa que a Equação 2.12 tem uma limitação quando o tamanho das comunidades não é homogêneo, e para eliminar tal efeito sugere a Equação 2.13 que faz com que ΔQ_{ij} seja normalizado pelo número de conexões dentro da comunidade c_i, minimizando o tempo computacional para $O(Nlog^2N)$.

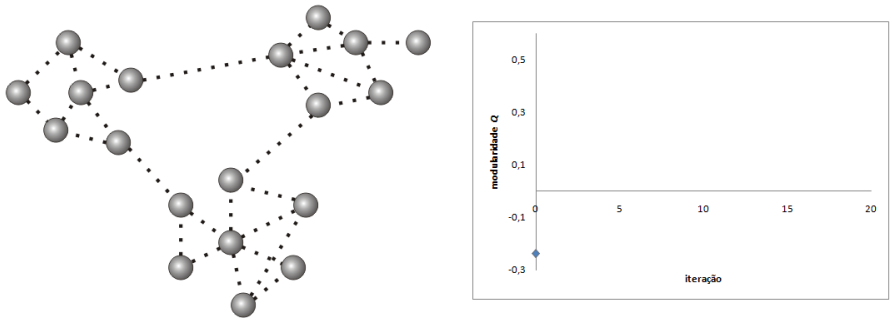
$$\Delta\widehat{Q}_{ij} = \frac{\Delta Q_{ij}}{\sum_i e_{ij}} \tag{2.13}$$

Há também trabalhos relacionados a identificação de comunidades em redes utilizando técnicas de *clustering* aglomerativo hierárquico (Balakrishnan & Deo, 2006), *clustering* hierárquico e algoritmo *K-means* (Gustafsson et al., 2006), e outras técnicas que não utilizam de propriedades de redes complexas, que fogem do foco deste trabalho. Mesmo assim, tanto os próprios autores como outros que analisaram a eficiência destas técnicas comparando com as de redes complexas já informadas anteriormente (Zhang et al., 2006), verificaram que não há diferença significante entre as técnicas. Zhang et al. (2006) concluíram que os métodos de redes complexas obtiveram melhores resultados nas redes analisadas por ele, por serem mais consistente com a realidade.

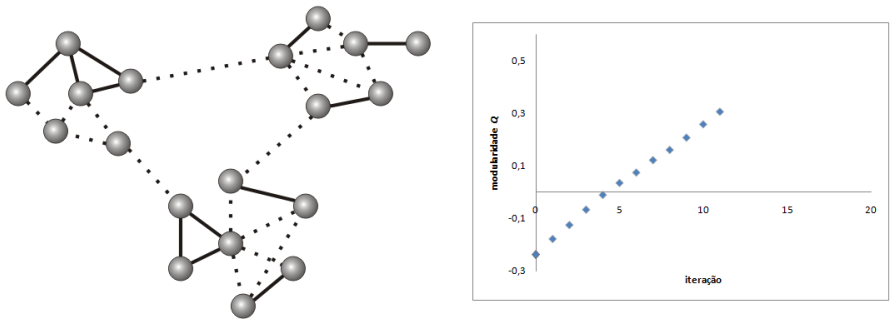
2.4 Classificação relacional

Como já comentado, conjuntos de dados podem possuir uma variedade de representações, influenciando diretamente as técnicas de aprendizado de máquina aplicáveis. As principais técnicas são baseadas na representação atributo-valor, que caracterizam individualmente os objetos. Porém, técnicas que trabalham com conjuntos de dados relacionais têm sido estudadas, assim como, técnicas que utilizam as duas representações dos dados, ou seja, técnicas que utilizam tanto a informação individual dos objetos quanto das relações entre eles.

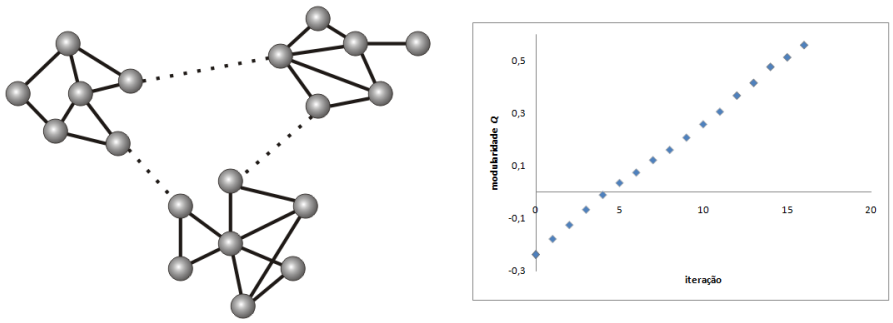
Em se tratando de dados modelados em grafos, é possível que haja duas situações distintas, uma na qual o grafo contém somente exemplos rotulados e outra na qual o grafo



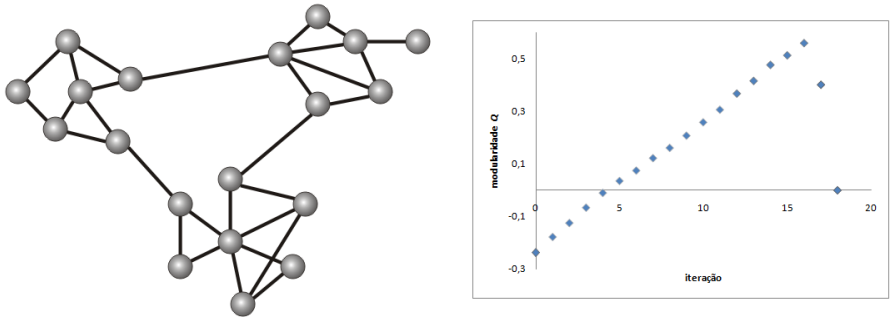
(a)



(b)



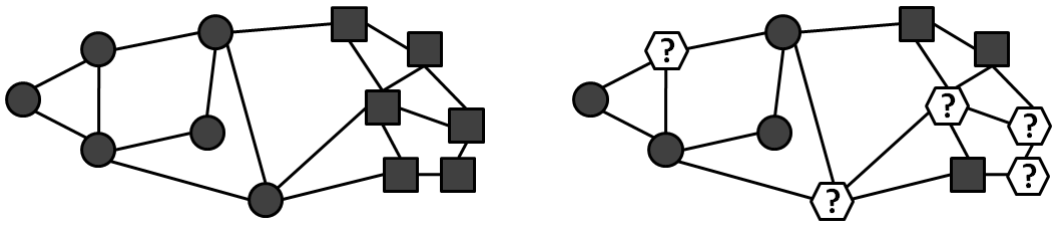
(c)



(d)

Figura 2.16: Ilustração de quatro etapas do método de Newman (2004b) para identificação de comunidades. Em (a) nenhuma aresta está inserida, nesse caso Q = -0,239; em (b) alguns componentes já estão agrupados, Q = 0,308; em (c) é a melhor divisão de comunidades, Q = 0,56; e em (d) todos componentes estão agrupado, Q = 0.

contém tanto exemplos rotulados quanto não rotulados. A Figura 2.17 apresenta ambas as situações.



(a) (b)

Figura 2.17: (a) Exemplo de um grafo que possui somente exemplos rotulados e (b) exemplo de um grafo contendo exemplos rotulados e não rotulados. As formas geométricas representam os rótulos dos exemplos, e a interrogação representa exemplos não rotulados.

Para casos com o grafo formado apenas por exemplos rotulados, as estratégias de classificação baseadas em grafos realizam a rotulação de novos exemplos podendo considerar todos exemplos presentes no grafo e as relações existentes.

No caso do grafo conter também os exemplos não rotulados, é necessário optar por considerar somente os exemplos rotulados existentes no grafo ou utilizar uma estratégia para considerar também os exemplos não rotulados. Considerar somente os exemplos rotulados durante a classificação pode ser problemático em conjuntos de dados com poucos exemplos rotulados. Por exemplo, se o método de classificação utilizado considerar somente os adjacentes de um exemplo não rotulado para classificá-lo, pode prejudicar a precisão se muitos adjacentes também forem exemplos não rotulados (tal situação pode ser observada na imagem (b) da Figura 2.17).

Comumente, técnicas de classificação relacional baseada em grafos formados por exemplos rotulados e por exemplos não rotulados, utilizam técnicas de inferência coletiva para induzir valores dos rótulos destes exemplos, estimando a classe de cada exemplo não rotulado ou sua distribuição de probabilidade, possibilitando serem considerados durante um procedimento de classificação relacional. A distribuição de probabilidade de um exemplo é a probabilidade dele pertencer a cada classe, utilizada principalmente por métodos probabilísticos de classificação.

A seguir são apresentados três métodos de inferência coletiva, a Amostragem de Gibbs (*Gibbs Sampling - GS*) (Geman & Geman, 1984), a Relaxação de Rótulos (*Relaxation Labeling - RL*) (Chakrabarti et al., 1998) e a Classificação Iterativa (*Iterative Classification - IC*) (Lu & Getoor, 2003). E, em seguida, também são apresentados seis classificadores relacionais, sendo dois classificadores que utilizam as informações individuais dos objetos em conjunto com as informações relacionais, a Classificação de Hipertexto - *Hypertext classification* (Chakrabarti et al., 1998) e a Classificação baseada em *links* - *Link-based classification* (Lu & Getoor, 2003), e quatro classificadores contidos na plataforma de classificadores denominada *NetKit-SRL* (Macskassy & Provost, 2007), que consideram apenas as informações relacionais, o Classificador relacional baseado nos vizinhos com

votação pesada - *Weighted-Vote Relational Neighbor Classifier*, o Classificador relacional baseado na distribuição de classe dos vizinhos - *Class-Distribution Relational Neighbor Classifier*, o Classificador Bayesiano baseado apenas na rede - *Network-Only Bayes Classifier* e o Classificador baseado apenas nas conexões da rede - *Network-Only Link-Based Classifier*.

2.4.1 Inferência Coletiva

Inferência coletiva significa inferir simultaneamente valores inter-relacionados, podendo ser aplicada a dados modelados em redes, estimando a classe de cada exemplo não rotulado ou sua distribuição de probabilidade. Basicamente, técnicas de inferência coletiva possuem três etapas principais: (i) um modelo de classificação local, que nada mais é do que um classificador que utiliza as informações individuais dos exemplos, (ii) um processo iterativo de atualização da distribuição dos exemplos, e (iii) um modelo de classificação relacional, o classificador baseado em grafo.

Portanto, inicia-se o procedimento com a utilização de um modelo de classificação local, para estimar uma distribuição de probabilidade inicial para cada exemplo não rotulado, seguido de um processo iterativo no qual se utiliza um modelo de classificação relacional, para atualização da distribuição de probabilidade de cada exemplo, sendo interrompido, em geral, quando os valores se estabilizam.

A utilização de inferência coletiva no processo de classificação reduz o erro dos resultados devido a exploração das dependências relacionais nos dados (Jensen et al., 2004). A principal vantagem de se estimar as classes por inferência coletiva é que, com isso, não é necessário descartar os exemplos não rotulados durante a classificação, obtendo uma distribuição de probabilidade (ou uma classe estimada) para todos exemplos, sendo que, para os exemplos rotulados, quando necessária sua distribuição de probabilidade, a probabilidade dele pertencer a sua classe é *1* e a probabilidade de pertencer a qualquer outra classe é *0*.

A seguir são apresentados três métodos de inferência coletiva.

Amostragem de Gibbs

O algoritmo de amostragem de Gibbs (Geman & Geman, 1984) implementado por Macskassy & Provost (2007) possui 5 principais etapas:

1. Estima-se uma distribuição de probabilidade de cada exemplo não rotulado utilizando um modelo de classificação local M_L. A classe inicial de cada exemplo é, então, obtida por uma amostra considerando sua distribuição de probabilidade, ou seja, a definição da classe inicial do exemplo segue um esquema de roleta, priorizando as classes com maiores valores de probabilidade.

2. Gera-se uma ordenação aleatória O desses exemplos não rotulados.

3. Para cada elemento x_i de O utiliza-se um modelo de classificação relacional M_R para se estimar a classe de x_i, também por amostragem a partir da distribuição de probabilidade gerada pelo classificador M_R, sendo que para obter a nova classe do exemplo x_i são utilizadas sempre as classes mais recentemente obtidas, incluindo as "novas" classes de x_1,...x_{i-1}.

4. Repete-se o passo anterior até que se estabilize a distribuição de probabilidade de cada exemplo para a ordenação aleatória O.

5. Repete-se o processo iterativo (passos 2, 3 e 4) uma quantidade de vezes que faça com que a ordenação aleatória O dada em cada repetição não influencie no resultado, obtendo a quantidade de vezes que cada classe foi definida para cada exemplo ao final de cada iteração, normalizando essa contagem para se obter uma distribuição de probabilidade final de cada exemplo.

Macskassy & Provost (2007) define uma repetição de 200 vezes no item 4 e 2000 vezes no item 5, assumindo que esses valores são utilizados comumente e são suficientes para se estabilizar a estimativa da probabilidade de cada exemplo.

Relaxação de Rótulos

O método de relaxação de rótulos proposto por Chakrabarti et al. (1998) é semelhante ao algoritmo de amostragem de Gibbs, porém considera, em cada iteração, a probabilidade do exemplo pertencer a cada classe, não atribuindo uma classe ao exemplo. Nesse método, a cada iteração t é alterada a distribuição de probabilidade de cada exemplo não rotulado, considerando a distribuição de probabilidade $t-1$ dos adjacentes desse exemplo. Note que nesse caso considera-se a distribuição de probabilidade da iteração anterior $t-1$, desconsiderando as distribuições de probabilidade já atualizadas na iteração t. As principais etapas desse algoritmo são:

1. Utilizando um modelo de classificação local M_L obtém-se a distribuição de probabilidade inicial $d_0(x_i)$ para cada exemplo x_i não rotulado.

2. Para cada elemento x_i do conjunto de exemplos não rotulados aplica-se o modelo de classificação relacional M_R para se obter sua nova distribuição de probabilidade $d_t(x_i)$, utilizando para isso as distribuições de probabilidade d_{t-1} dos vizinhos de x_i na rede.

3. Repete-se o passo anterior até que os valores sejam estabilizados.

Observando que em determinadas situações a relaxação de rótulos não converge para uma situação estável, mas oscila entre dois ou mais estados, Macskassy & Provost (2007) adaptaram o método fazendo com que em cada iteração t a estimativa obtida para o exemplo x_i na iteração $t-1$ tenha mais peso e que a nova distribuição de probabilidade obtida na iteração t tenha menos peso.

Para isso, é necessário uma alteração na forma de atribuição de uma nova distribuição de probabilidade a um exemplo. No processo de relaxação de rótulo, na iteração t temos a distribuição de probabilidade $d_t(x_i)$ para o exemplo x_i, e na iteração $t+1$ temos a distribuição de probabilidade de $d_{t+1}(x_i)$, que é baseada nas distribuições d_t dos vizinhos de x_i. Para se estimar a nova distribuição de probabilidade da iteração $t+1$ (agora definida como $nd_{t+1}(x_i)$), com peso, é necessário considerar, além da distribuição $d_{t+1}(x_i)$ obtida pelas distribuições d_t dos vizinhos, também a distribuição $nd_t(x_i)$ da iteração t, de acordo com a Equação 2.14.

$$nd_{k+1}(x_i) = \beta_{k+1}.d_{k+1}(x_i) + (1-\beta_{k+1}).nd_k(x_i) \tag{2.14}$$

Sendo que β_0 é iniciado entre 0 e 1 e $\beta_{t+1} = \beta_t.\alpha$ com α sendo uma constante de decaimento. Em seus estudos de casos os autores utilizaram β_0 com o valor 1 e α com o valor 0,99, afirmando que experimentos mostraram que a técnica é robusta para altos valores de α. Além disso, Macskassy & Provost (2007) utilizaram para a quantidade de iterações o valor fixo de 99.

Classificação Iterativa

Proposto por Lu & Getoor (2003), o método classificação iterativa não gera probabilidade, mas estima uma determinada classe para todos exemplos não rotulados. As etapas do método são:

1. Utilizando um modelo de classificação local M_L obtém-se uma classe para cada exemplo não rotulado.

2. Gera-se uma ordenação O dos exemplos não rotulados pela quantidade de diferentes classes existentes em seus adjacentes, uma vez que há maior confiança na classificação dos exemplos com menor diversidade de classes nos adjacentes. Para cada exemplo não rotulado em O aplica-se o modelo de classificação relacional M_R, obtém-se sua distribuição de probabilidade, e se atribui a classe de maior probabilidade para o exemplo. Caso todos adjacentes ainda não possuam classe atribuída então esse exemplo continua não rotulado.

3. Repete-se o passo anterior até que nenhum exemplo tenha sua classe alterada.

2.4.2 Classificadores relacionais

Com os métodos de inferência coletiva já conceitualizados, agora são apresentados os classificadores relacionais, os quais consideram duas formas de se obter a distribuição de probabilidade de cada exemplo. A primeira é utilizada apenas para obter uma distribuição de probabilidade inicial para os exemplos não rotulados, obtida por métodos de inferência coletiva, conhecida como distribuição de probabilidade conjunta. E a segunda considera a vizinhança do exemplo na rede e é obtida por métodos de classificação relacional baseada em grafos, conhecida como distribuição de probabilidade marginal.

A seguir são apresentados seis classificadores. Os dois primeiros classificadores utilizam as informações individuais e relacionais dos exemplos, e os quatro classificadores contidos na plataforma *NetKit-SRL* utilizam apenas as informações relacionais, desconsiderando as informações individuais dos objetos, caso exista.

Classificação de Hipertexto (Hypertext classification - Hc)

O classificador Hc, proposto por Chakrabarti et al. (1998), utiliza as informações individuais e relacionais dos exemplos para classificação de dados relacionais com conteúdo textual, como por exemplo, páginas *web* e documentos de patentes. A probabilidade de um exemplo x_i pertencer a classe c é descrita na Equação 2.15, obtida a partir de duas distribuições de probabilidade, a primeira é obtida por um classificador local que utiliza o conteúdo textual dos documentos, e a segunda utiliza um classificador relacional baseado nos exemplos adjacentes a x_i. Nesta equação, t_i representa o conteúdo textual estruturado de x_i e N_i os exemplos que estão em sua vizinhança na rede, a classe atribuída à x_i é a que maximiza essa probabilidade..

$$P(x_i = c|t_i, N_i) = P(x_i = c|t_i).P(x_i = c|N_i) \tag{2.15}$$

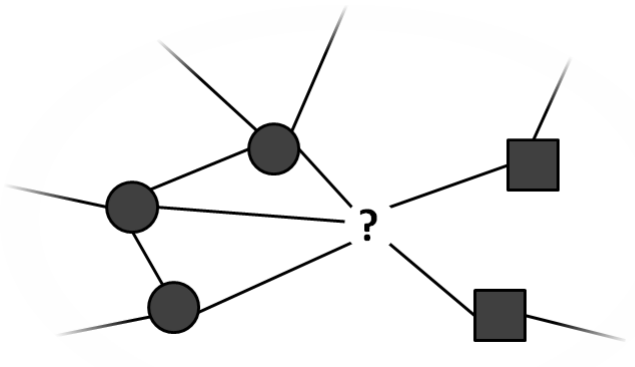
Para considerar uma distribuição de probabilidade também para os exemplos não rotulados, os autores utilizam relaxação de rótulos, utilizando o conteúdo textual para o modelo de classificação local ($P(x_i = c|t_i)$) e a vizinhança do vértice para o modelo de classificação relacional ($P(x_i = c|N_i)$), atribuindo como classe sempre a que maximizar a probabilidade. Chakrabarti et al. (1998) afirmam que este é o primeiro classificador a combinar a informação individual e relacional para classificação de exemplos textuais, e demonstra uma boa redução do erro quando comparado a um classificador baseado somente no texto, principalmente quando não há muitos exemplos rotulados.

Classificação baseada em links (Link-based classification - Lbc)

Proposto por Lu & Getoor (2003), Lbc também considera, além da estrutura relacional do conjunto de dados, os atributos dos objetos. A técnica utiliza um modelo de regressão

logística (Hosmer & Lemeshow, 1989) para realizar a classificação. Regressão logística é utilizada em casos binários para se estimar a probabilidade de uma variável pertencer a cada uma das duas classes. Para conjuntos de dados multi-classes os autores consideram, para cada classe, a probabilidade de pertencer e de não pertencer à classe.

Sendo assim, utiliza-se um grafo direcionado, no qual são observadas as características individuais (atributos) dos exemplos e suas ligações na rede. Portanto há dois vetores de informações que serão utilizados, o primeiro contendo os atributos dos exemplos (informações locais) e o segundo obtido de sua vizinhança (informações relacionais). Três modelos diferentes foram propostos para construção do vetor baseado na vizinhança da rede, ilustrados na Figura 2.18, todos utilizando um vetor com tamanho igual a quantidade de classes. O modelo *mode-link*, contém o valor 1 na posição da classe mais frequente dos vizinhos e 0 no restante do vetor, o modelo *count-link* faz uma contagem do número de adjacentes de cada classe, e o modelo *binary-link*, com cada posição do vetor possuindo 1 caso o exemplo tenha algum adjacente da classe ou 0 caso não tenha.



	●	■	▲
mode-link	1	0	0
count-link	3	2	0
binary-link	1	1	0

Figura 2.18: Exemplo dos modelos *mode-link*, *count-link* e *binary-link* para representação da vizinhança de um exemplo na rede.

Com esses vetores, os autores utilizam para classificação o modelo de inferência coletiva denominado classificação iterativa, proposto pelos próprios autores. Aplica-se então regressão logística tanto como modelo de classificação local como modelo de classificação relacional, sendo que no modelo local utiliza-se apenas o vetor de atributos e no modelo relacional utiliza-se os dois vetores, de atributos e de observações extraídas da vizinhança. Portanto, seguindo o método de classificação iterativa, a cada etapa computa-se a probabilidade de cada exemplo pertencer a cada classe, classificando com a classe de maior probabilidade, interrompendo o processo iterativo ao se estabilizar as classes atribuídas aos exemplos.

Experimentos foram realizados, utilizando conjuntos de dados de páginas *web* e de citação entre artigos científicos, comparando com outros métodos de representação das ligações em vetores, sendo que os resultados, em geral, foram melhores para os métodos *count-link* e *binary-link*, com uma leve vantagem, mas não estatisticamente significante, para o *count-link*. Também verificou-se os efeitos individuais do sentido das arestas,

considerando individualmente as que chegam ao exemplo e as que saem, observando que considerando as arestas que saem dos exemplos os resultados, em geral, foram melhores que considerando as arestas que chegam, mas os melhores resultados foram considerando ambas arestas.

Plataforma NetKit-SRL

Como já citado, os quatro classificadores descritos a seguir, propostos por Macskassy & Provost (2007), utilizam apenas as informações relacionais dos dados.

Para descrição dos classificadores são utilizadas três diferentes notações relacionadas a probabilidade de um exemplo x_i pertecer a classe c dada sua vizinhança N_i na rede:

- $P_{MR}(x_i = c|N_i)$: é a probabilidade utilizada pelo modelo de classificação relacional, obtendo uma distribuição de probabilidade para cada exemplo, porém, desconsiderando adjacentes não rotulados.

- $P_{IC}(x_i = c|N_i)$: é a probabilidade obtida após aplicar o modelo de classificação relacional, descrito anteriormente, em um processo de inferência coletiva, obtendo uma distribuição de probabilidade para os exemplos não rotulados. Para os exemplos rotulados a probabilidade dele pertencer a sua classe é *1* e a probabilidade de pertencer a qualquer outra classe é *0*.

- $P(x_i = c|N_i)$: é a probabilidade final utilizada para classificação, que é influenciada também pelas distribuições de probabilidade dos exemplos não rotulados, obtidas por técnicas de inferência coletiva.

Além disso, são utilizadas as notações P_{MR}, P_{IC} e P para denominar a distribuições de probabilidades obtidas, respectivamente, pela normalização de $P_{MR}(x_i = c|N_i)$, $P_{IC}(x_i = c|N_i)$ e $P(x_i = c|N_i)$ para cada classe c.

Devido a esses classificadores não considerarem as informações locais, os autores desconsideram o modelo de classificação local no processo de inferência coletiva, utilizando apenas o modelo de classificação relacional.

Classificador relacional baseado nos vizinhos com votação pesada (*Weighted-Vote Relational Neighbor Classifier* - wvRN): o classificador considera a vizinhança do vértice ponderada pelo peso das arestas. Na Equação 2.16 é descrita a probabilidade utilizada pelo modelo de classificação relacional, sendo w_{ij} o peso da ligação entre os exemplos x_i e x_j, e Z o somatório do peso de todas arestas incidentes em x_i, para normalização dos valores.

$$P_{MR}(x_i = c|N_i) = \frac{1}{Z} \sum_{(x_j \in N_i | classe(x_j) = c)} w_{ij} \tag{2.16}$$

Esse modelo de classificação relacional é então utilizado em um processo de inferência coletiva obtendo-se uma distribuição de probabilidades P_{IC} para todos exemplos, que é considerada na obtenção da probabilidade final utilizada pelo classificador (Equação 2.17).

$$P(x_i = c|N_i) = \frac{1}{Z} \sum_{(x_j \in N_i | classe(x_j) = c)} w_{ij}.P_{IC}(x_j = c|N_j) \tag{2.17}$$

Classificador relacional baseado na distribuição de classe dos vizinhos (*Class-Distribution Relational Neighbor Classifier* - cdRN): o classificador cdRN obtém a distribuição de probabilidade de um exemplo x_i utilizando a similaridade de dois vetores, um de classes VC obtido para cada exemplo e um de referências VR obtido para cada classe. O vetor de clases do exemplo x_i, denominado $VC(x_i)$, possui tamanho igual a quantidade de classes, e cada posição k do vetor contém o somatório do peso de todas arestas que ligam x_i a vértices da classe c_k. Na Equação 2.18 é descrita a obtenção do valor para a posição k do vetor.

$$VC(x_i)_k = \sum_{(x_j \in N_i | classe(x_j) = c_k)} wij \tag{2.18}$$

O vetor de referências VR da classe c, denominado $VR(c)$, é a normalização dos vetores de classes VC de todos exemplos que são da classe c, e é definido na Equação 2.19, sendo X_c o conjunto dos exemplos rotulados que são da classe c.

$$VR(c) = \frac{1}{|X_c|} \sum_{(x_i \in X_c)} VC(x_i) \tag{2.19}$$

Portanto, o modelo de classificação relacional considera os dois vetores e obtém a probabilidade de um exemplo x_i pertencer a uma classe c (Equação 2.20) de acordo com a similaridade entre $VC(x_i)$ e $VR(c)$ (os autores utilizam similaridade cosseno).

$$P_{MR}(x_i = c|N_i) = similaridade(VC(x_i), VR(c)) \tag{2.20}$$

Os vetores VC e VR são gerados apenas para os exemplos rotulados, sendo necessária a utilização desse modelo de classificação relacional (o qual utiliza a Equação 2.20) em uma técnica de inferência coletiva para considerar também os não rotulados durante a classificação. Com isso, se obtém a distribuição de probabilidades P_{IC} a ser utilizada para construção de um novo vetor de classes nVC (Equação 2.21) a ser utilizado na obtenção da probabilidade final (Equação 2.22).

$$nVC(x_i)_k = \sum_{(x_j \in N_i | classe(x_j) = c_k)} wij.P_{IC}(x_j = c_k|N_j) \tag{2.21}$$

$$P(x_i = c|N_i) = similaridade(nVC(x_i), VR(c)) \tag{2.22}$$

Classificador Bayesiano baseado apenas na rede (*Network-Only Bayes Classifier* - nBC): este classificador é uma adaptação do algoritmo *Hypertext classification* (Chakrabarti et al., 1998), desconsiderando as informações individuais dos exemplos. Nesse método, o modelo de classificação relacional se baseia na probabilidade dada pela Equação 2.23, mas utiliza a Equação 2.24, desconsiderando a divisão por $P(N_i)$ devido a essa probabilidade independer da classe c.

$$P_{MR}(x_i = c|N_i) = \frac{P(N_i|c).P(c)}{P(N_i)} \tag{2.23}$$

$$P_{MR}(x_i = c|N_i) = P(N_i|c).P(c) \tag{2.24}$$

Considerando que a probabilidade $P(c)$ é a proporção de exemplos da classe c no conjunto de exemplos rotulados e a probabilidade $P(N_i|c)$ é dada pela Equação 2.25, sendo a probabilidade de x_j ser de sua classe c_j (informação já conhecida, pois x_j é um exemplo rotulado) dado que a classe de x_i é c, e ponderado pelo peso w_{ij} da aresta entre x_i e x_j.

$$P(N_i|c) = \frac{1}{Z} \prod_{x_j \in N_i} P(classe(x_j) = c_j|classe(x_i) = c)^{w_{i,j}} \tag{2.25}$$

Para os exemplos x_j em N_i que não são rotulados é utilizada uma técnica de inferência coletiva, utilizando como modelo de classificação relacional um classificador que considera a probabilidade obtida pela Equação 2.24. Nesse caso, a distribuição de probabilidade final P será semelhante a P_{MR}, porém, considerando os exemplos rotulados e os não rotulados.

Classificador baseado apenas nas conexões da rede (*Network-Only Link-Based Classifier* - nLB): o método nLB é uma adaptação do algoritmo *Link-based classification* (Lu & Getoor, 2003). Assim como no algoritmo original, cria-se um vetor de características para cada exemplo rotulado baseado em sua vizinhança na rede, então utiliza-se regressão logística para criar um modelo de classificação relacional de acordo com esses vetores, obtendo a distribuição de probabilidade P_{MR}.

A diferença do nBL para o algoritmo *Link-based classification* é que o nBL não utiliza as informações individuais dos exemplos, somente considera as relações entre eles. Além disso, utiliza apenas uma estratégia para criação do vetor (enquanto o *Link-based classification* utiliza três estratégias diferentes), esta estratégia é semelhante ao *count-link*, o qual apresentou melhores resultados no trabalho de Lu & Getoor (2003), e constrói um vetor equivalente ao vetor de classes $VC(x_i)$ (definido na Equação 2.18), utilizado no classificador cdRN. A única alteração no algoritmo foi a normalização do vetor de classes $VC(x_i)$, pois, após experimentos iniciais, foi constatada uma melhor performance, em geral, segundo Macskassy & Provost (2007).

Da mesma forma que o classificador cdRN, a distribuição de probabilidade P_{IC} é obtida por um método de inferência coletiva utilizando o modelo de classificação relacional, e essa nova distribuição de probabilidade P_{IC} é utilizada para gerar o novo vetor de classes nVC (Equação 2.21) a ser utilizado na regressão logística, considerando também os exemplos não rotulados.

Macskassy & Provost (2007) utilizaram conjuntos de dados relacionais e verificaram que, para inferência coletiva, o método de relaxamento de rótulos é melhor quando existem poucos exemplos rotulados, porém todos se comportam bem quando existem muitos exemplos rotulados. Para classificação relacional, o classificador nLB apresenta melhores resultados quando existem muitos exemplos rotulados, e os classificadores wvRN e cdRN, que apresentam baixa variância, se mostram eficientes quando existem poucos exemplos rotulados. Em se tratando de combinações, o nLB com qualquer método de inferência coletiva se comporta melhor quando muitos exemplos são rotulados, e o wvRN e cdRN, ambos em conjunto com relaxação de rótulos, são melhores quando poucos exemplos são conhecidos.

2.4.3 Considerações finais

Observa-se que trabalhos que utilizam informações relacionais consideram conjuntos de dados que possuem inerentemente uma estrutura relacional, como, por exemplo, páginas *web* e documentos de patentes e artigos científicos contendo citações.

No capítulo a seguir é apresentada uma alternativa para representar relacionalmente conjuntos de dados proposicionais, por meio de grafos baseados na similaridade entre os objetos. Também é apresentada uma técnica de formação de redes, possível de ser aplicada somente aos exemplos rotulados ou a todo conjunto, formando grafos contendo ou não os exemplos não rotulados.

Devido a possibilidade de formação de grafos constituídos apenas por exemplos rotulados, também são apresentados dois classificadores relacionais baseados nesses grafos. Com isso, evita-se a necessidade do uso de inferência coletiva, já bastante explorada por Macskassy & Provost (2007).

Além disso, é apresentado o modelo de redes denominado K-Associados, sua formação, e classificadores específicos, que exploram características dessa rede.

Capítulo

3

Redes complexas em classificação relacional

Neste Capítulo são apresentadas as técnicas desenvolvidas durante este trabalho para uso de redes complexas em um processo de classificação relacional. É descrita uma técnica para construção de uma representação relacional baseada em similaridade a partir de dados em uma tabela atributo-valor, e algoritmos para classificação relacional.

3.1 Modelagem em redes e classificação

Nesta seção é descrita uma técnica para a construção de redes baseada na similaridade entre os exemplos, a qual denominamos Redes Hierárquicas, e também os classificadores baseados nessas redes.

A motivação principal da técnica de formação de redes baseadas em similaridade entre vértices decorreu da possibilidade de se aplicar técnicas de classificação relacional baseada em grafos mesmo para dados proposicionais representados no formato atributo-valor. A técnica descrita possibilita a construção de grafos formados apenas por exemplos rotulados, tornando desnecessário o uso de métodos de inferência coletiva.

A opção por não utilizar os modelos de redes vistos na Seção 2.3.1, é devido ao fato de tenderem a uma estrutura previamente determinada, como por exemplo, se utilizarmos o modelo livre de escala para construção das redes estaríamos "forçando" uma estrutura livre de escala. Da mesma forma, se simplesmente fossem inseridas arestas entre pares de vértices com similaridade acima de uma similaridade mínima, duas situações poderiam ocorrer, ou seria obtido um grafo com muitos componentes, para um alto limiar de similaridade mínima, ou um grafo com muitas arestas, para menores valores do limiar de similaridade mínima. Para esses casos, muitas propriedades de redes complexas precisariam ser anali-

sadas individualmente por componente ou a estrutura se tornaria densamente conectada, impactando de forma negativa na análise dos conjuntos de dados. Ou seja, procuramos construir uma rede baseada nas relações de similaridade entre os vértices, que possua um número de arestas controlado pelo usuário e permita a identificação de estruturas de comunidades (ou agrupamentos) na rede.

As redes hierárquicas buscam a conexão de elementos com alta similaridade, utilizando uma função de interconectividade, descrita a seguir, construindo um grafo com grupos de vértices similares sendo fortemente conectados entre si e fracamente conectados a outros grupos.

3.1.1 Construção das redes hierárquicas

As redes hierárquicas são grafos conexos não direcionados construídos baseados na similaridade entre os objetos, atingindo um grau médio próximo do desejado e priorizando a existência de arestas entre objetos com alta similaridade. Essas redes podem ser configuradas como probabilísticas ou determinísticas, se diferenciando apenas na forma como são selecionadas as arestas, descrita a seguir.

A construção da rede visa manter uma estrutura com maior número de ligações intra-comunidades e entre vértices mais similares e com menor número de ligações entre comunidades distintas. O processo de construção inicia-se com a rede contendo todos os vértices e nenhuma aresta, ou seja, cada vértice constituindo um único componente. Um processo aglomerativo hierárquico é iniciado conectando pares de vértices iterativamente, baseado em um limiar de similaridade mínima. Assumimos a medida de similaridade na faixa entre 0 e 1, com o valor 1 indicando exemplos altamente similares. Tal limiar é inicializado com um alto valor, mais especificamente, um valor que permita selecionar os 5% pares de vértices mais similares.

Uma estrutura de repetição externa (o segundo *Enquanto* no Algoritmo 1) identifica as arestas potenciais, isto é, todos pares de vértices com similaridade acima do atual limiar de similaridade mínima e pertencentes a componentes distintos. Após identificar as arestas potenciais é realizada a chamada de uma estrutura de repetição interna (Algoritmo 2), a qual seleciona os vértices ou componentes a serem agrupados. O limiar de similaridade mínima é atualizado a cada iteração externa, obtendo um limiar que permita adicionar os próximos 5% pares de vértices mais similares. O processo é interrompido ao gerar um único componente, formando uma rede conexa.

Porém, considerar todas as arestas potenciais não garante a construção de uma rede com o grau médio definido como entrada, é necessário um novo processo de seleção de arestas, as arestas pré-selecionadas. Portanto, para cada componente C_i, verifica-se a quantidade de arestas que o componente precisa para atingir o grau médio definido, e essa é a quantidade de arestas potenciais selecionadas para o conjunto de arestas pré-

selecionadas do componente C_i, considerando somente as arestas potenciais conectadas ao componente C_i.

A identificação das arestas pré-selecionadas de cada componente é realizada durante a estrutura de repetição interna, seguida da seleção do par de componentes a ser agrupado. Dois componentes são agrupados se (i) possuem um alto número de arestas pré-selecionadas entre vértices dos componentes, e (ii) há vértices altamente similares entre os componentes. A seleção dos dois componentes a serem agrupados é baseada em uma função de interconectividade, definida na Equação 3.1 e ilustrada na Figura 3.1. Os dois componentes C_i e C_j que maximizarem a equação de interconectividade serão agrupados inserindo-se definitivamente as arestas pré-selecionadas que ligam vértices entre C_i e C_j.

Na Equação 3.1, $\#C$ corresponde ao número de vértices no componente C, existe a aresta (x_i, x_j), e $similaridade(x_i, x_j)$ corresponde a similaridade entre os vértices $x_i \in C_i$ e $x_j \in C_j$.

$$interconectividade(C_i, C_j) = \frac{1}{\#C_i + \#C_j} \sum_{\substack{x_i \in C_i, x_j \in C_j, \\ \exists aresta(x_i, x_j)}} similaridade(x_i, x_j) \tag{3.1}$$

Na Figura 3.1 pode-se observar que se usássemos um critério apenas baseado em distância (ou similaridade), os componentes que seriam unidos seriam o C1 e C2. No critério adotado, além da similaridade entre os exemplos, também se consideram as conexões entre e intra-componentes após a inserção das arestas. A rede formada por este algoritmo, portanto, tende a ter uma estrutura de comunidade bem definida, i.é, grupos de vértices similares altamente conectados entre si e fracamente conectados a outros grupos.

Algoritmo 1 Construção da rede hierárquica baseada em similaridade

Entrada:
Conjunto de vértices: $V = v_1,...,v_n$
Grau médio: $grauMedio$
Matriz de similaridade entre exemplos: $similaridade$
Saída:
Rede gerada, sendo um conjunto de vértices e de arestas: (V,A)

Componentes $C \leftarrow V$
Arestas $A \leftarrow \emptyset$
$minSim \leftarrow$ similaridade para se obter os 5% pares de vértices mais similares
Enquanto ($\#C > 1$)
  Enquanto ($\exists$ par (x,y) | similaridade$(x,y) \geq minSim$, $x \in C_i$, $C_i \in C$, $y \in C\text{-}C_i$)
    Agrupamento dos componentes(C,A,$grauMedio$,$minSim$)
  $minSim \leftarrow$ similaridade para se acrescentar 5% pares de vértices mais similares

Retorna (V,A)

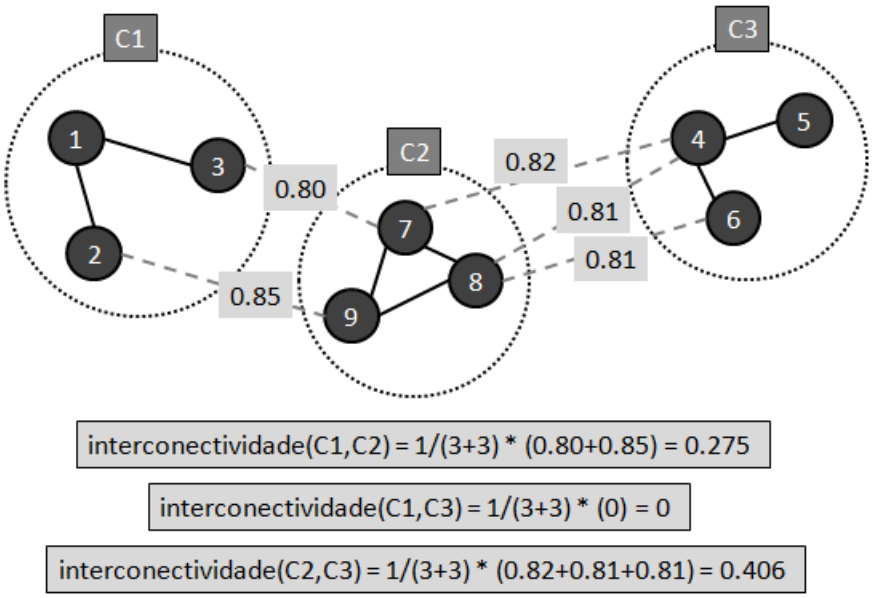


Figura 3.1: Exemplo da aplicação da equação de interconectividade para uma rede com três componentes e cinco novas arestas candidatas a serem inseridas, nesse caso seriam agrupados os componentes $C2$ e $C3$ inserindo as duas arestas candidatas existentes entre eles.

Na etapa do Algoritmo 2 referente a forma de obtenção dos pares de vértices, determinística ou probabilística, para as redes hierárquicas probabilísticas essa seleção é feita de forma aleatória, e para as determinísticas são selecionadas sempre as arestas que representam maior similaridade.

A complexidade do algoritmo é $O(N^2)$, sendo N o número de vértices, pois, no pior caso, serão verificadas todas as arestas existentes na lista de similaridades para construção da lista *arestas possíveis* de cada bloco.

3.1.2 Classificador

A seguir é apresentado o classificador proposto, que utiliza grafos modelados apenas com exemplos rotulados (conjunto de treino), sendo que os exemplos não rotulados (conjunto de teste) são considerados como um grupo de exemplos desconectados do grafo.

Classificador baseado na rede hierárquica - cbRH: O classificador baseado em rede hierárquica utiliza um conjunto de dados modelado em rede somente com seus exemplos rotulados, e para cada exemplo de teste é identificado o mais similar na rede e seus adjacentes para se obter a classe de maior probabilidade, ponderando pela similaridade entre os vértices (Figura 3.2). Na Equação 3.2 é descrita a probabilidade do exemplo x_i pertencer a classe c dado o conjunto X_j, que contém o exemplo x_j mais similar a x_i na rede e seus vizinhos diretos, sendo Z o somatório da similaridade entre x_i e os exemplos pertencentes a X_j.

Algoritmo 2 Agrupamento dos componentes

Entrada:
Conjunto de componentes: C
Conjunto de arestas: A
Grau médio: $grauMedio$
Limiar de similaridade: $minSim$
Saída:
Conjunto de componentes: C
Conjunto de arestas: A

arestasPreSelecionadas $\leftarrow \emptyset$
Para cada componente C_i de C
$qtdDeArestas \leftarrow$ ($grauMedio$ * $\#C_i$ / 2) - $\#A(C_i)$
Se (qtdDeArestas ≤ 0)
qtdDeArestas $\leftarrow 1$)
Para todo par(i,j) obtido determinística ou probabilisticamente | $i \in C_i$ e $j \in C$ - C_i
Se (similaridade(i,j) $\geq minSim$)
arestasPreSelecionadas(C_i,C_j)$\leftarrow$arestasPreSelecionadas(C_i,C_j) $\cup$ (i,j)
$qtdDeArestas--$
Se ($qtdDeArestas == 0$)
$break$
(C_a,C_b) $\leftarrow$ max(interconectividade(C_i,C_j)) %componentes selecionados
$C_a \leftarrow C_a \cup C_b$ %união dos componentes
$A(C_a) \leftarrow A(C_a) \cup A(C_b) \cup$ arestasPreSelecionadas(C_a,C_b) %união das arestas dos componentes unidos
$C \leftarrow C$ - C_b %remove componente que foi unido
$A \leftarrow A$ - $A(C_b)$ %remove de A as arestas do componente removido

Retorna (C,A)

$$P(x_i = c|X_j) = \frac{1}{Z} \sum_{(x_j \in X_j | classe(x_j)=c)} w_{ij} \tag{3.2}$$

Como não há exemplos não rotulados na rede, então não são utilizados métodos de inferência coletiva para obtenção da distribuição da probabilidade conjunta. A principal diferença deste classificador para os classificadores propostos por Macskassy & Provost (2007) é o fato de possibilitar a construção de um modelo com os exemplos rotulados, tornando desnecessária a inserção dos exemplos não rotulados no grafo.

3.2 Redes K-Associados

Outra proposta de formação de rede baseada na similaridade entre exemplos rotulados é o modelo de rede denominado K-Associados. As redes k-associados foram desenvolvidas

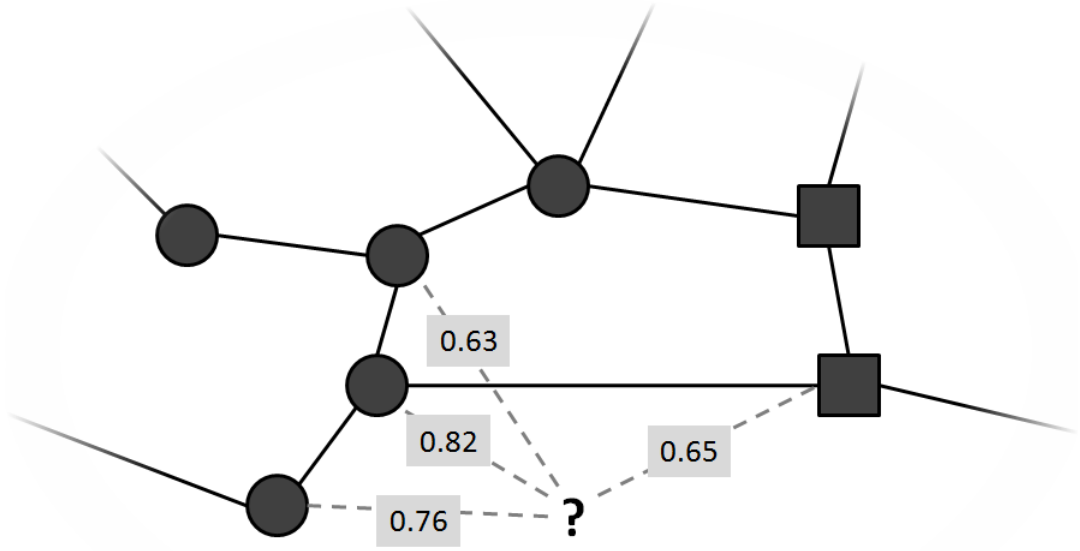


Figura 3.2: Exemplo de utilização do classificador mais similar e adjacentes, no qual o exemplo a ser classificado não está na rede. Considera-se o exemplo mais similar que está na rede e seus adjacentes para identificar a classe com maior probabilidade, utilizando a similaridade entre os exemplos.

em conjunto com os pesquisados Alneu de Andrade Lopes (orientador deste trabalho), João Bertini e Zhao Liang, e vêm demonstrando grande potencial para o processo de classificação (Lopes et al., 2009) de dados com ruídos e dados dinâmicos.

3.2.1 Construção das redes k-associados

As redes k-associados são grafos formados por um ou mais componentes completamente puros, ou seja, componentes formados por objetos que pertencem a mesma classe, mas na qual somente exemplos similares são conectados entre si. Portanto, essa rede gera um modelo relacional, que pode ser considerado como aprendizado supervisionado, pois utiliza-se de um conjunto de treino.

A construção da rede é relativamente simples, define-se um valor k de vizinhos e, para cada vértice v_i, obtêm-se um conjunto N_i contendo dos k vizinhos mais próximos, apenas os vértices que são da mesma classe de v_i. Portanto, é possível que haja até duas arestas entre um par de vértices e o número máximo de aresta em um componente C é de $k.|C|$. A Figura 3.3 ilustra a rede k-associado, demonstrando os componentes gerados para diferentes valores de k, e o Algoritmo 3 detalha a formação dessa rede k-associados.

Com isso, a quantidade de componentes da rede será no mínimo a quantidade de classes existentes nos dados, mas possivelmente uma mesma classe é dividida em mais de um componente. De acordo com o valor k definido na entrada dos dados e com o grau de cada vértice é possível extrair medidas de purezas dos vértices, dos componentes, e da rede como um todo, além de possibilitar também a aplicação de outras medidas de redes

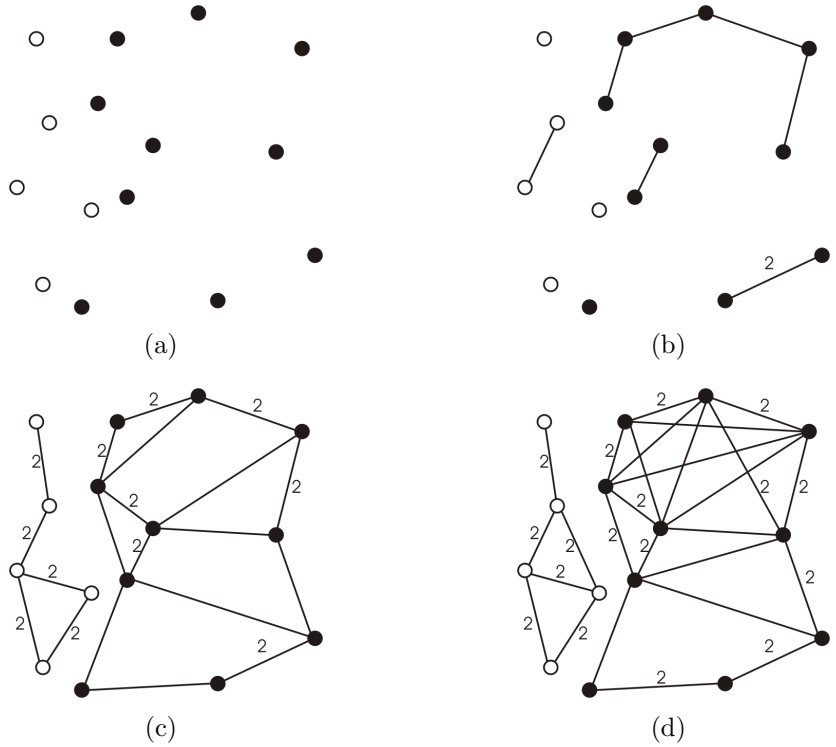


Figura 3.3: (a) distribuição do conjunto de dados, e (b), (c) e (d) correspondem a rede k-associados com k sendo 1, 3 e 5, respectivamente. Observe que as arestas podem representar mais de uma conexão, e as cores representam as duas classes presentes.

Algoritmo 3 Formação da rede k-associados

Entrada:
Conjunto de vértices: $V = v_1,...,v_n$
Conjunto de classes: $L = classe(v_1),...,classe(v_n)$
Matriz de similaridades: S
Número de vizinhos: k
Saída:
Rede gerada, sendo um conjunto de vértices e de arestas: (V,E)

Arestas $E \leftarrow \emptyset$
Para cada vértice v_i de V
$similaridades\_ordenadas \leftarrow$ ordenar_decrescente$(S(v_i))$
Para $j = 0$ até k
Se $(classe(v_i) = classe(similaridades\_ordenadas(j))$
$E \leftarrow E \cup aresta(v_i, similaridades\_ordenadas(j))$

Retorna (V,E)

complexas para extração de informações da rede.

Na Figura 3.4 é possível observar a construção de das redes k-associados com k igual a 3 para um conjunto de dados artificial, no qual os exemplos apresentam diferentes separações. A quantidade de componentes e de arestas internas nos componentes são

relacionadas a pureza.

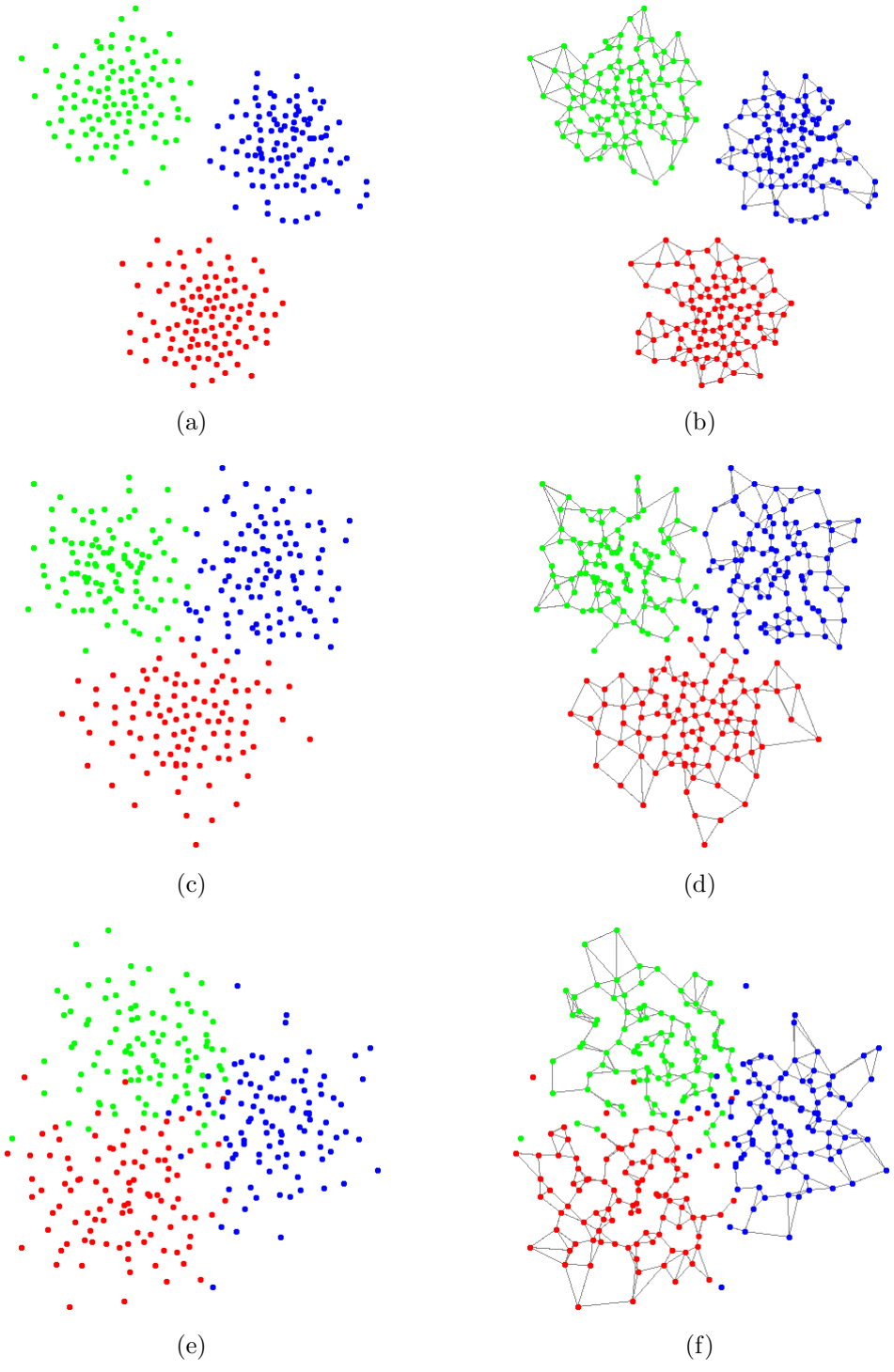


Figura 3.4: Conjunto de dados artificial, as figuras (a), (c) e (e) apresentam a distribuição dos dados em diferentes separações, e as figuras (b), (d) e (f) apresentam, respectivamente, as redes k-associados formadas com k igual a 3.

A complexidade do algoritmo é $O(N^2 \log N)$, sendo N o número de vértices, pois, para cada vértice é necessário ordenar sua lista de similaridade para os outros vértices.

3.2.2 Medida de pureza dos componentes da rede k-associados

O método de geração da rede k-associados permite a construção de um modelo utilizando o conjunto de treino e possibilita calcular a medida de pureza de cada componente gerado. Utilizando a topologia da rede para identificar o quão misturados estão os vértices de diferentes classes.

Portanto, a medida de pureza de um componente está relacionada a quantidade de arestas que o componente possui em relação a quantidade de arestas máxima que poderia possuir, considerando que a ausência de arestas em um componente indica que havia elementos de outra classe na k-vizinhança dele.

Sendo g_i o grau do vértice i, N o número total de vértices da rede e k a quantidade de vizinhos definido na geração da rede, então $g_i/2k$ corresponde a fração de ligações existentes entre o vértice i e os outros vértices em seu componente, variando entre 0 e 1, inclusive. Portanto, o total de arestas entre vértices do componente C é dado pela Equação 3.3.

$$|E_c| = \frac{1}{2}\sum_{i=1}^{N_c} g_i = \frac{N_c}{2}\sum_{i=1}^{N_c}\frac{g_i}{N_c} = \frac{N_c}{2}\langle G_c\rangle \tag{3.3}$$

Sendo N_c a quantidade de vértices do componente C, e G_c o grau médio no componente. O número máximo de arestas existentes no componente é kN_c, e a probabilidade de arestas entre elementos do componente C, ou seja, a medida de pureza, é dada pela Equação 3.4.

$$pureza(C) = \frac{\frac{N_c\langle G_c\rangle}{2}}{kN_c} = \frac{\langle G_c\rangle}{2k} \tag{3.4}$$

É possível demonstrar empiricamente o comportamento dessa equação, na Figura 3.5 representamos o valor de $pureza(C)$ (média de 10 execuções) para um componente analisado em cinco conjuntos de dados artificiais contendo 250 vértices. Esse conjuntos de dados, chamados de P90, P80, P70, P60, P50, foram criados usando uma distribuição normal com, respectivamente, 90, 80, 70, 60, e 50% de "pureza".

Com esses experimentos é possível verificar que $pureza(C)$ é uma boa aproximação da pureza do componente.

Para obtenção da pureza da rede toda, é considerada a Equação 3.5 ponderando a pureza de cada componente pelo número de exemplos existentes no componente.

$$P_r = \sum_{i=1}^{|C|}\frac{\#C_i}{\#R}P_{Ci} \tag{3.5}$$

Em geral, raramente uma rede construída com um único valor de k conterá os maiores valores de pureza para todos componentes se comparado com redes formadas por todos

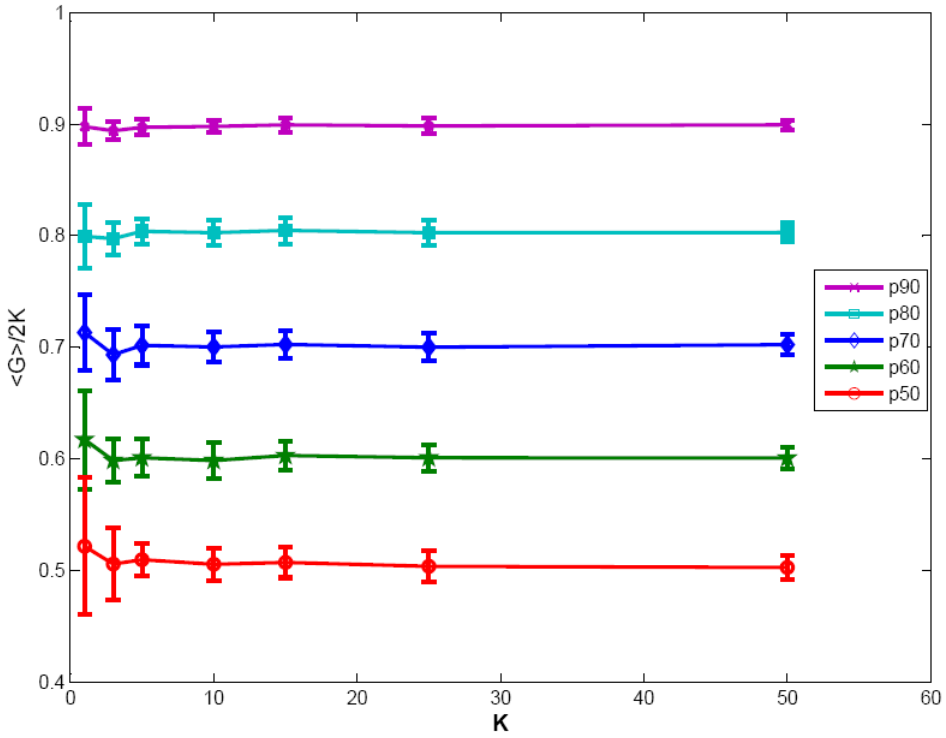


Figura 3.5: A média de $pureza(C)$ do componente analisado, existente em redes k-associados de conjuntos de dados com 90, 80, 70, 60, e 50% de pureza na região do componente.

valores possíveis de k. Dessa forma, a seguir é apresentada uma técnica para construção de uma rede k-associados ótima, a qual conterá componentes formados por diferentes valores de k, buscando altos valores de pureza.

3.2.3 Construção das redes k-associados ótima

Nas redes k-associados é possível que componentes com vértices em comum, mas formados com diferentes valores de k tenham purezas diferentes. Com isso, propomos a criação de uma rede que mantém um conjunto dos melhores componentes, em termo de pureza, que contenham todos os vértices da rede, chamada rede K-Associados Ótima.

Para construção da rede k-associados ótima, a ideia é variar o k iterativamente mantendo os melhores componentes encontrados, baseado na medida de pureza já descrita. O Algoritmo 4 descreve o processo de formação dessa rede. Iniciando com a construção de uma rede k-associados com $k = 1$, um processo iterativo é realizado incrementando k até atingir um k_{max}, selecionando novos componentes caso possuam uma pureza maior que o anterior. Observe que um componente gerado para um determinado valor de k irá conter um ou mais componentes gerados para $k - 1$, sendo assim, para selecionar o mais recente componente obtido é preciso que este tenha uma pureza maior que a pureza de algum dos componentes anteriores, os quais serão substituídos.

A Figura 3.6 exibe a formação de uma rede k-associados ótima para o conjunto de da-

Algoritmo 4 Redes k-associados Ótima

Entrada:

Conjunto de vértices: $V = v_1,...,v_n$

Conjunto de classes: $L = classe(v_1),...,classe(v_n)$

Matriz de similaridades: S

Número de iterações: k_{max}

Saída:

Rede gerada, sendo um conjunto de vértices e de arestas: (V,E)

$k \leftarrow 1$

$C_{otimo} \leftarrow k - associados(V, L, S, k)$

Para $k = 2$ até k_{max}

$C \leftarrow k - associados(V, L, S, k)$

Para cada componente C_i de C

identificar os componentes $\{Cj\}$ em C_{otimo} correspondentes a C_i

Se $pureza(C_i) > pureza$ de algum dos componentes em $\{C_j\}$

$C_{otimo} \leftarrow C_{otimo}$ - $\{C_j\} \cup C_i$

Retorna (V,E)

dos Zoo, obtido do UCI Repositório[1], com k_{max} igual a 50. Esse conjunto de dados possui 101 exemplos distribuídos em 7 classes, e devido a possuir 17 atributos a distribuição dos dados no plano aproxima os exemplos similares mas não garante que os mais próximos no plano são os mais próximos no espaço real. Nas figuras as cores dos vértices representam as classes, arestas em cinza são da rede k-associados e arestas em preto da k-associados ótima, os números indicam a pureza dos componentes para a rede k-associados, sendo que na figura não foram destacadas as arestas duplicadas e as purezas iguais a 1 para a rede $k = 1$.

Na Figura 3.7 é mostrada a rede final k-associados ótima com os respectivos valores de k considerado para cada componente, contendo também arestas duplicadas em destaque. É possível observar que conforme o valor de k aumenta componentes mais puros tendem a considerar maiores k, representando regiões mais puras no espaço dos dados.

A complexidade dessa técnica se mantém $O(N^2 \log N)$, quando $k_{max} \ll N$, porém, obviamente, na prática é mais demorado que o método de construção de redes k-associados.

3.2.4 Classificador baseado na rede k-associados

Os classificadores baseados nas redes k-associados são divididos em quatro algoritmos, diferenciados pela forma com que selecionam os vizinhos mais próximos, para obtenção da distribuição de probabilidade utilizando a medida de pureza já conceitualizada. A Equação 3.6 descreve a probabilidade do exemplo x_i pertencer a classe c dada uma vizinhança

[1] http://archive.ics.uci.edu/ml/

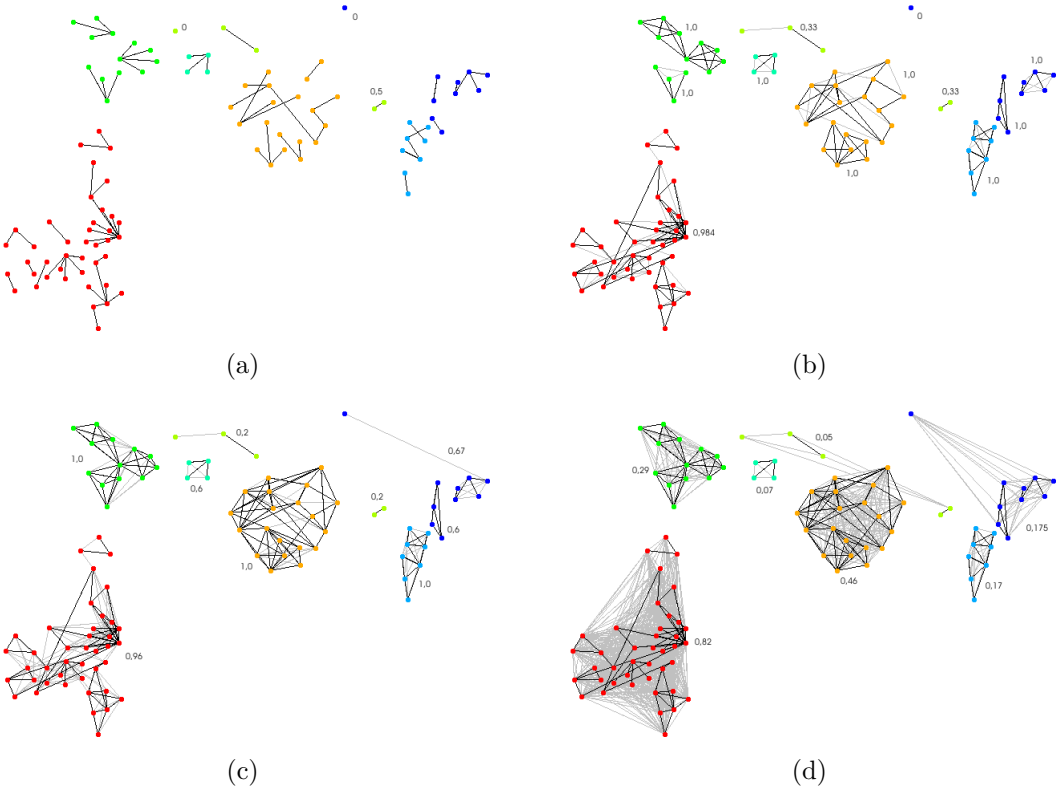


Figura 3.6: Formação das redes k-associados e k-associados ótima para o conjunto de dados Zoo, com valores para k igual a (a) 1, (b) 3, (c) 5 e (d) 50, e k_{max} igual a 50.

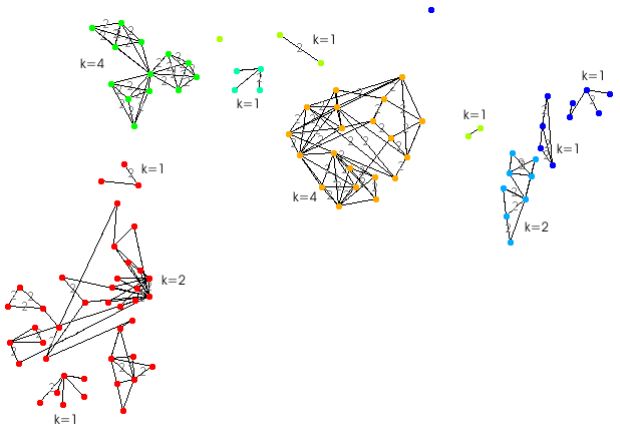


Figura 3.7: Rede final k-associados ótima para conjunto de dados Zoo, com o valor de k_{max} igual a 50.

N_i, com $pureza(C(x_j))$ sendo a pureza do componente em que o vértice x_j pertence e Z é utilizado para normalização dos valores, sendo a soma da probabilidade de x_i pertencer a cada classe c.

$$P(x_i = c|N_i) = \frac{1}{Z} \sum_{(x_j \in N_i | classe(x_j) = c)} w_{ij}.pureza(C(x_j)) \tag{3.6}$$

A seguir é descrito como cada algoritmo faz a seleção dos exemplos para obtenção da probabilidade dada pela Equação 3.6. A diferença entre eles é na forma com que é considerada a inserção do exemplo de teste na rede. A visão teste-rede considera os k exemplos de treino mais próximos (k sendo o valor usado na construção da rede), ou seja, para quais exemplos de treino o exemplo de teste teria uma aresta ligando. A visão rede-teste considera o oposto, os exemplos da rede que teriam o exemplo de teste como um dos k mais próximos, ou seja, o exemplo de teste receberia uma aresta desses exemplos de treino. Já para os classificadores baseados em comitê, ambas visões seriam consideradas.

Visão teste-rede: A visão teste-rede considera que para classificar um exemplo situado, por exemplo, entre dois componentes como representado na Figura 3.8, este exemplo se conectaria aos k vértices mais próximos.

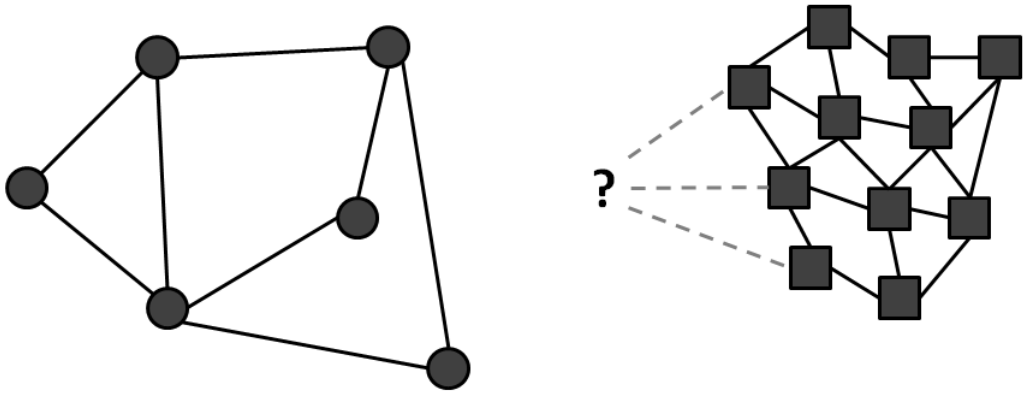


Figura 3.8: Exemplo de utilização do classificador k-associados visão teste-rede, para uma rede com k igual a 3 contendo dois componentes com pureza igual a 1. As arestas tracejadas indicam ligações aos 3 mais próximos exemplos de teste.

Visão rede-teste: O classificador k-associados visão rede-teste considera as ligações que existiriam entre um exemplo posicionado entre os dois componentes (Figura 3.9) se cada vértice da rede fosse conectado a seus k vizinhos considerando também o exemplo novo. Nesta situação apenas o componente à esquerda se conectaria ao vértice novo, pois o componente a direita têm todos os k-vizinhos de qualquer de seus vértices dentro do próprio componente.

Classificador baseado em comitê: Considerando as duas visões, teste-rede e rede-teste, com ambas realizando observações diferentes dos dados, é possível realizar a classificação das duas formas e considerar um comitê dos classificadores baseado em certeza, ou seja, somente classifica um exemplo de teste caso ambas visões concordarem.

Nesse método de classificação por comitê, não se garante a classificação de todos os exemplos de teste, mas espera-se que a precisão dos que forem classificados seja significativamente maior quando comparado com a classificação utilizando as visões individualmente.

Classificador baseado em comitê com maior probabilidade: Semelhante ao

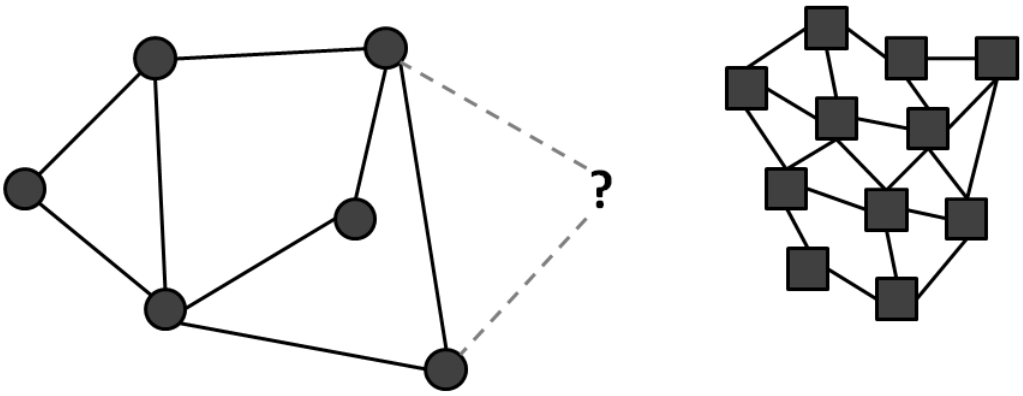


Figura 3.9: Exemplo de utilização do classificador k-associados visão rede-teste, para uma rede com k igual a 3 contendo dois componentes com pureza igual a 1. As arestas tracejadas indicam quais exemplos da rede teriam o exemplo de teste entre os 3 mais próximos se ele estivesse na rede.

classificador de comitê com certeza, o comitê por votação também utiliza ambas visões, teste-rede e rede-teste, porém, um exemplo de teste será classificado com a classe de maior probabilidade, ou seja, cada uma das visões irá gerar uma probabilidade do exemplo de teste pertencer a cada classe e a maior probabilidade define a classe do exemplo. Com isso, todos exemplos são classificados.

Capítulo

4

Avaliação Experimental

As avaliações foram efetuadas em conjuntos de dados proposicionais, apresentados em tabela atributo-valor, e em conjuntos de dados relacionais, representados por seus grafos. A avaliação foi dividida em duas etapas principais, a avaliação das técnicas de construção das redes hierárquicas e k-associados, e a avaliação dos classificadores.

4.1 Metodologia

4.1.1 Construção e avaliação das redes

Para construção das redes é necessário o cálculo da similaridade entre os exemplos e a definição dos valores para os parâmetros de entrada do algoritmo. Para os conjuntos numéricos a medida de similaridade usada foi o oposto da distância euclidiana normalizada, (0 - pouco similar e 1 - muito similar). Para os conjuntos de documentos textuais foi usada a medida do Cosseno e para os conjuntos relacionais a similaridade de Jaccard com profundidade de uma aresta (Choi & Krishnamoorthy, 2007).

Na formação das redes hierárquicas foram utilizados, como grau médio de entrada, os valores 1, 3, 5 e 15, pois em experimentos preliminares observou-se que a estrutura de comunidades já fica bem definida para valores próximos de 5. Para construção das redes k-associados foi utilizado como valor k de entrada os valores 1, 3, 5 e 15, e para o k_{max} das redes k-associados ótima foi utilizado o valor 15. Além da divisão do conjunto em 10 partes para validação cruzada 10-*fold*, o procedimento de avaliação das redes também foi realizado em três execuções, para que fosse obtida uma média mais estável dos resultados, importante principalmente para conjunto de dados com baixa quantidade de exemplos.

Para a avaliação das redes, buscou-se identificar se as redes construídas representaram de forma razoável a distribuição dos exemplos nos conjuntos de dados. Para isso, são realizadas comparações da pureza da vizinhança dos vértices nas redes geradas com a pureza dos exemplos mais similares nos conjuntos de dados proposicionais. O processo para cálculo dessas purezas é descrito a seguir. Para os conjuntos de dados relacionais, a comparação foi realizada entre a pureza da vizinhança nas redes geradas e a pureza da vizinhança nas redes do próprio conjunto de dados.

Para cada conjunto de dados, são necessárias três medidas de pureza, a primeira consiste na pureza do próprio conjunto, a segunda na pureza das redes hierárquicas e a terceira na pureza das redes k-associados construídas.

A pureza dos conjuntos de dados foi definida de tal forma a avaliar a pureza na vizinhança de um exemplo no espaço original e nas redes. As medidas são diferenciadas para conjuntos proposicionais e relacionais. No caso proposicional é considerada a média aritmética da pureza de cada exemplo x_i de treino, obtida pela proporção dos k vizinhos mais próximos de x_i que possuem a mesma classe do exemplo. O valor de k usado deve ser o mesmo que será usado como parâmetro de entrada para construção da rede (hierárquica ou k-associados)

No caso dos conjuntos relacionais também foi considerada a média aritmética da pureza de cada exemplo, porém, considerando a proporção de adjacentes que possui a mesma classe do exemplo.

A pureza das redes hierárquicas é idêntica à pureza definida para os conjuntos relacionais, baseado nos adjacentes de cada exemplo. E a pureza utilizada para as redes k-associados é a pureza já definida na Seção 3.2.2.

O processo de avaliação das redes construídas consistiu em três etapas principais, (i) a separação do conjunto em 10 partes para validação cruzada 10-*fold* (também utilizada durante a classificação), com cada *fold* contendo o conjunto de treino formado por nove partes e o de teste por uma parte, (ii) construção das redes hierárquicas e das redes k-associados utilizando os conjuntos de treino, (iii) utilização de medidas de pureza para avaliação de cada rede construída. Esse processo é esquematizado na Figura 4.1.

Caracterização das redes

Para a caracterização das redes foram computadas as propriedades globais de redes complexas apresentadas na revisão bibliográfica, gerando informações que auxiliam na caracterização e no entendimento do comportamento da rede.

Também foram computadas medidas relativas aos vértices como: o grau de todos vértices, grau mínimo e máximo, média do grau e sua distribuição. Em seguida, foram gerados o coeficiente de agrupamento médio da rede, a média do menor caminho, o diâmetro e a máxima modularidade Q da rede ao aplicar o método de identificação de comunidades proposto por Newman (2004b).

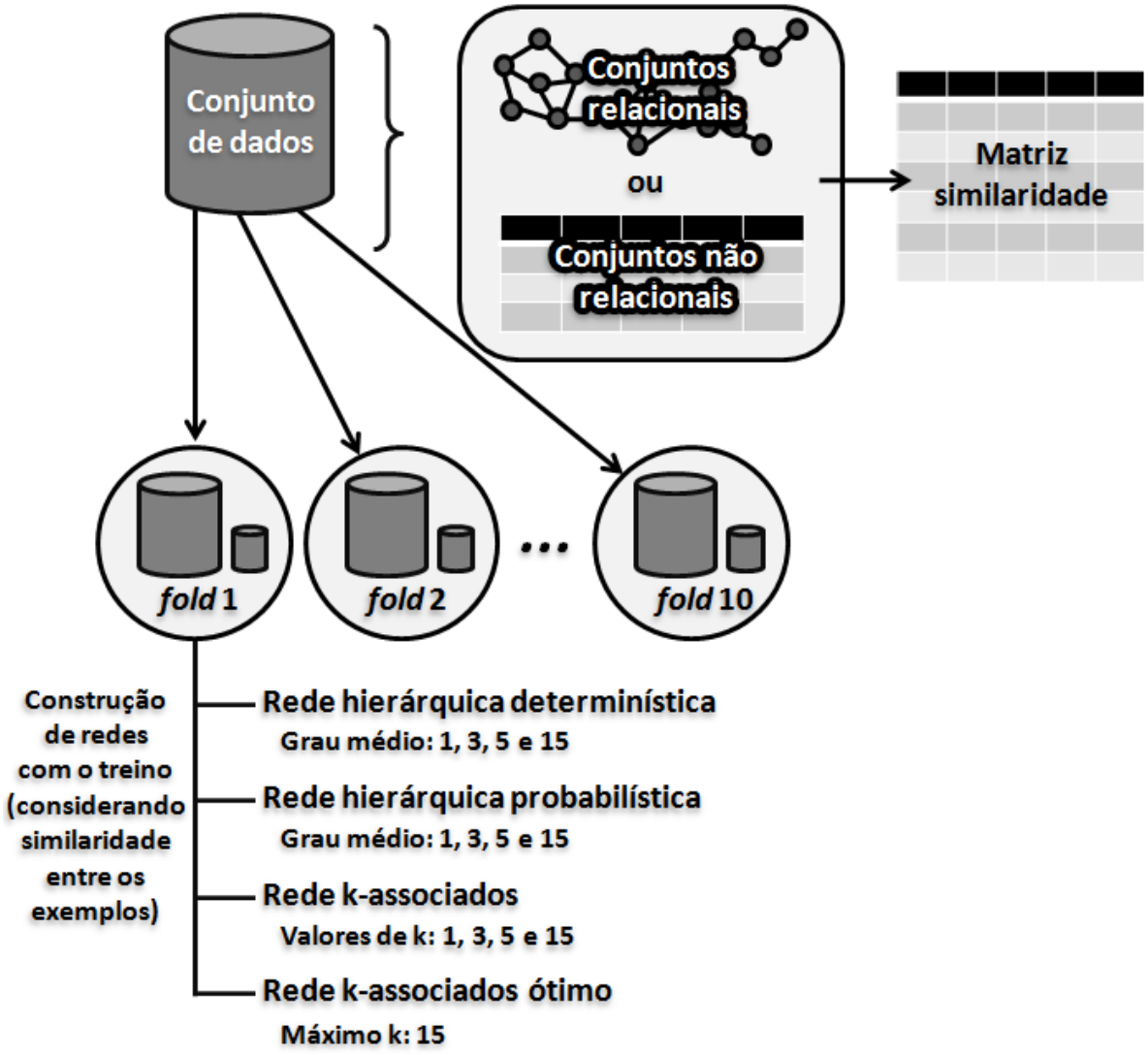


Figura 4.1: Esquema de divisão dos dados para construção e avaliação das redes hierárquicas.

Para análise visual das redes foi usada uma adaptação da API Prefuse em conjunto com Java2D, a qual busca efetuar a melhor distribuição possível dos vértices em um ambiente bidimensional, evitando sobreposição de arestas.

4.1.2 Avaliação dos classificadores baseados em grafos

Neste trabalho são propostos classificadores baseados nas redes hierárquicas e redes k-associados. Para as redes hierárquicas é proposto o classificador denominado *Classificador baseado nas redes hierárquicas* (cbRH) e para as redes k-associados são propostos quatro classificadores, o *Classificador visão teste-rede*, o *Classificador visão rede-teste*, o *Classificador baseado em comitê* e o *Classificador baseado em comitê com maior probabilidade*.

A avaliação dos classificadores é esquematizada na Figura 4.2. Sendo que para os

conjuntos de dados proposicionais os classificadores foram comparados com os classificadores *k-Nearest Neighbors* - k-NN (k = 1, 3, 5 e 15) e Naive Bayes. Para os conjuntos de dados relacionais, representados por seus grafos, os classificadores foram comparados com o *Classificador Bayesiano baseado apenas na rede* (nBC), em conjunto com o método de inferência coletiva relaxação de rótulos. A opção pelo classificador nBC com relaxação de rótulos é por ter apresentado melhores resultados dentre os classificadores contidos na plataforma *NetKit-SRL*.

O processo de avaliação dos classificadores é dividido nas seis partes seguintes:

1. Construção da matriz de similaridade entre os exemplos e divisão do conjunto de dados em 10 partes para validação cruzada 10 -*fold*, com cada *fold* contendo o conjunto de teste formado por uma parte e o conjunto de treino formado pelas outras nove partes.

2. Para os conjuntos relacionais é realizada a aplicação do classificador nBC com relaxação de rótulos para a rede principal do conjunto de dados, num processo de validação cruzada considerando os 10-*fold* gerados. E para os conjuntos de dados proposicionais é realizada a aplicação dos classificadores k-NN e Naive Bayes realizando validação cruzada 10-*fold*.

3. Construção das redes hierárquicas e k-associados considerando os exemplos do conjunto de treino (corresponde a fase de aprendizado).

4. Classificação do conjunto de teste de cada *fold* com o classificador baseado em redes hierárquicas (cbRH), e obtenção da média e variância.

5. Classificação do conjunto de teste de cada *fold* com os quatro classificadores k-associados e k-associados ótima, e obtenção da média e variância.

6. Avaliar os conjuntos de teste com o classificador nBC com relaxação de rótulos. Observa-se que este classificador considera o grafo com exemplos rotulados e não rotulados. Portanto, para se obter uma média da precisão do classificador adotamos o critério de considerar esse grafo em duas partes para treino e teste, com 90% e 10% dos vértices, respectivamente. A precisão é obtida classificando os 10% de teste, desconsiderando eventuais rótulos que possam haver nesses vértices. Para obtenção da média e variância esse processo é repetido 10 vezes em um esquema de validação cruzada.

Como existem procedimentos aleatórios na divisão dos conjuntos e construção das redes, para se determinar o padrão de comportamento dos classificadores com maior precisão foram consideradas as médias de três execuções da validação cruzada.

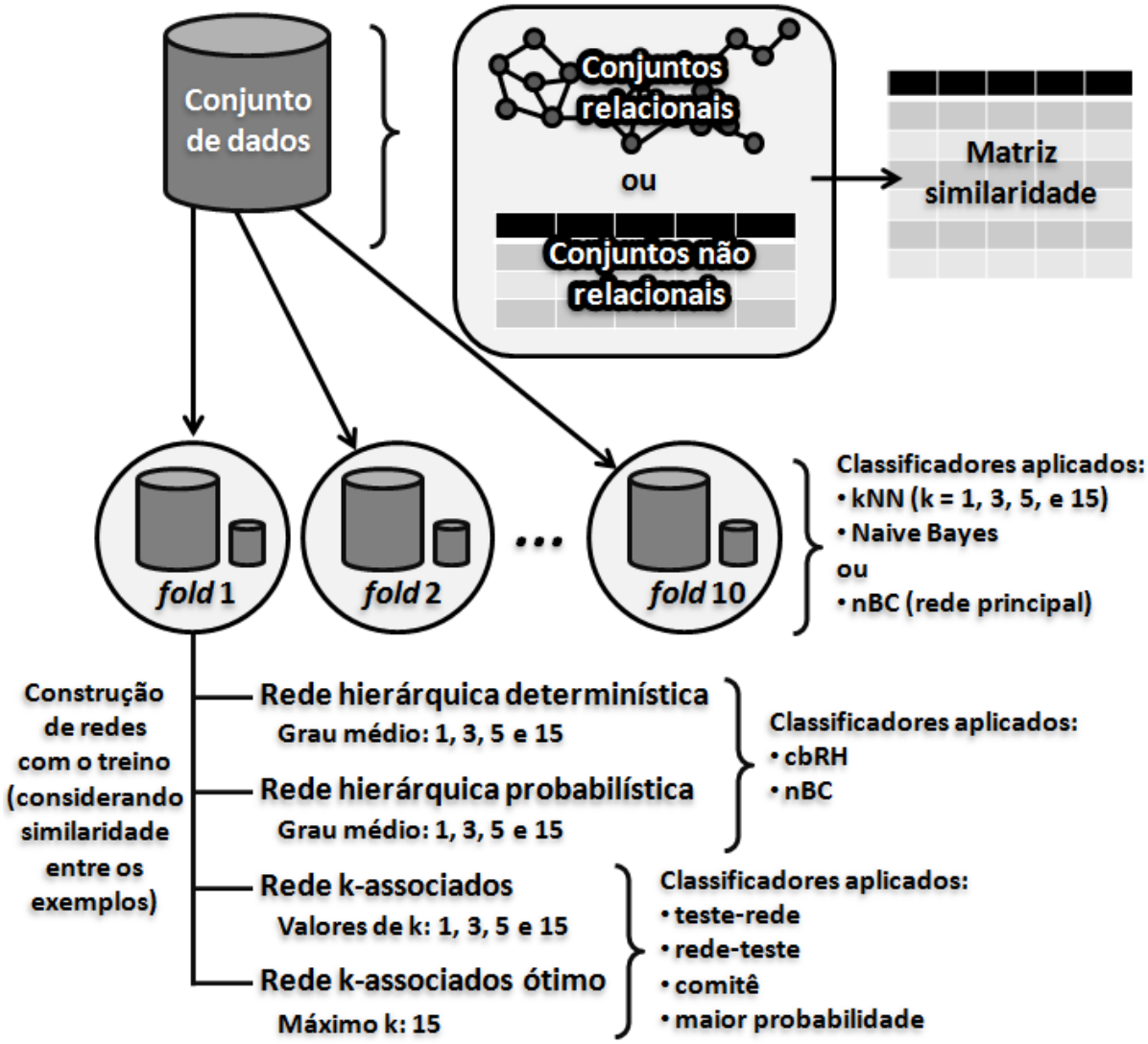


Figura 4.2: Esquema de divisão dos dados para construção das redes e avaliação dos classificadores.

Para avaliação dos classificadores é utilizado o teste estatístico não-paramétrico de Kruskal-Wallis, aplicando o pós-teste de múltiplas comparações de Dunn. Também são realizados testes estatísticos para avaliação dos classificadores, assim verificaremos se é viável o uso de classificadores relacionais em conjuntos de dados proposicionais, e também avaliando os classificadores relacionais.

4.2 Resultados

4.2.1 Conjuntos de dados

Foram selecionados 18 conjuntos de dados para avaliação das técnicas descritas, sendo 10 conjuntos de dados numéricos, 4 conjuntos de dados relacionais e 4 conjuntos de dados textuais. A opção por conjuntos de dados numéricos, relacionais e textuais permite uma

avaliação ampla do potencial das técnicas.

Conjuntos atributo-valor numéricos

Os conjuntos numéricos utilizados foram conjuntos comumente utilizados em trabalhos de classificação, obtidos do UCI Repositório[1], e são eles:

- *Balance*: Contém características e resultados de experimentos psicológicos.
- *Ecoli*: Informações para classificação de sítios de localização de proteínas.
- *Glass*: Informações extraídas de análises de amostras de vidro.
- *Ionosphere*: Informações para classificação de elétrons livres na ionosfera.
- *Iris*: Informações de medições de comprimento e largura de sépalas e pétalas da planta Íris.
- *Sonar - Connectionist Bench (Sonar, Mines vs. Rocks)*: Informações para classificação de sinais sonares.
- *Wdbc - Breast Cancer Wisconsin (Diagnostic)*: Informações para classificação de câncer de mama.
- *Wine*: Informações químicas de vinhos cultivados numa mesma região da Itália.
- *Yeast*: Informações para classificação de sítios de localização de proteínas.
- *Zoo*: Informações para identificação da classe de animais.

Por simplicidade utilizamos apenas conjuntos sem valores ausentes e apenas com dados numéricos, sendo que todos passaram por um processo de normalização dos valores dos atributos. E, obviamente, todos os conjuntos possuem classes discretas por se tratar de classificação.

Na Tabela 4.1 detalhes dos conjuntos de dados são descritos, tais como o nome do conjunto de dados, informações sobre a quantidade de exemplos, de atributos e de classes, e o erro majoritário.

As classes de cada conjunto de dados são apresentadas a seguir, com suas respectivas proporções:

- *Balance*: *balanced* (7,8%), *left* (46,1%) e *right* (46,1%).
- *Ecoli*: cp (42,6%), im (22,9%), pp (15,5%), imU (10,4%), om (5,9%), omL (1,5%), imL (0,6%) e imS (0,6%).

[1] http://archive.ics.uci.edu/ml/

Tabela 4.1: Conjuntos de dados numéricos

Conjunto de dados	#Exemplos	#Atributos	#Classes	Erro majoritário
Balance	625	4	3	53.92
Ecoli	336	8	8	57.44
Glass	214	10	7	64.49
Ionosphere	351	34	2	35.89
Iris	150	4	3	33.33
Sonar	208	60	2	46.63
Wdbc	569	32	2	37.26
Wine	178	13	3	60.11
Yeast	1484	8	10	68.73
Zoo	101	17	7	59.41

- *Glass*: *building windows float processed* (32,7%), *building windows non float processed* (7,9%), *vehicle windows float processed* (35,5%), *vehicle windows non float processed* (0,0%), *containers* (6,1%), *tableware* (4,2%) e *headlamps* (13,6%).
- *Ionosphere*: *good* (64,1%) e *bad* (35,9%).
- *Iris*: *Iris Setosa* (33,3%), *Iris Versicolour* (33,3%) e *Iris Virginica* (33,3%).
- *Sonar*: *rock* (46,7%) and *mine* (53,3%).
- *Wdbc*: *malignant* (37,3%) e *belign* (62,7%).
- *Wine*: *type 1* (33,2%), *type 2* (39,9%) e *type 3* (26,9%).
- *Yeast*: CYT (31,2%), NUC (28,9%), MIT (16,4%), ME3 (11,0%), ME2 (3,4%), ME1 (3,0%), EXC (2,5%), VAC (2,0%), POX (1,3%) e ERL (0,3%).
- *Zoo*: *class 1* (40,6%), *class 2* (19,8%), *class 3* (5%), *class 4* (12,9%), *class 5* (3,9%), *class 6* (7,9%) e *class 7* (9,9%).

Conjuntos atributo-valor textuais

Os conjuntos textuais foram obtidos de diferentes fontes, sendo todos textos científicos. O conjunto *CBR-ILP-IR* faz parte do domínio de Inteligência Artificial, o *CS* do domínio de Ciência da Computação, o *Chemistry* do domínio de Química e o *Physics* do domínio de Física.

Na Tabela 4.2 são apresentados os conjuntos com suas respectivas classes e proporção de documentos.

Na Tabela 4.3, são apresentadas a quantidade de *stems*(palavras reduzidas ao seu radical) obtidas para cada conjunto de dados e a quantidade após aplicar o corte de Luhn (Luhn, 1958). Observa-se uma diferença significativa no número de *stems* do conjunto *CBR-ILP-IR*, isso ocorreu devido a esse conjunto ser constituído apenas do resumo dos artigos.

Tabela 4.2: Conjuntos de dados textuais

Conjunto de dados	#Docs.	Classes	Erro majoritário
CBR-ILP-IR	574	*Case Based Reasoning* (48,1%) *Inductive Logic Programming* (20,7%) *Information Retrieval* (31,2%)	51,9%
CS	398	*Computer Hardware* (22,1%) *Humam-Computer Interaction* (23,1%) *Artificial Intelligence* (23,1%) *Security & Criptology* (31,7%)	68,3%
Chemistry	372	*Analytical Chemistry* (22,3%) *Inorganic Chemistry* (25%) *Organic Chemistry* (26,1%) *Plumer Science* (26,6%)	73,4%
Physics	383	*Biophysics* (24,8%) *Geophysics* (25,3%) *Mechanics* (28,5%) *Quantum Physics* (21,4%)	71,5%

Tabela 4.3: Quantidade de *stems* dos conjuntos de dados

Conjunto de dados	#Documentos	#*Stems* Gerados	#*Stems* Final (Luhn)
CBR-ILP-IR	574	4101	1634
CS	398	23295	6937
Chemistry	372	28194	9067
Physics	383	22195	8638

Conjuntos relacionais

Os conjuntos relacionais são descritos a seguir, obtidos diretamente no formato de grafo. Os conjuntos de dados *books*, *football* e *blogs* foram obtidos em um repositório disponível no site do pesquisador Mark Newman[2], e o conjunto *industry-yh* é disponibilizado no site da plataforma NetKit[3]:

- *books*: Rede de livros sobre política nos EUA publicados no período da eleição presidencial de 2004 e vendidos pela *Amazon.com*. As arestas indicam livros vendidos simultaneamente para um mesmo comprador, e as classes são liberal, neutro ou conservador . Conjunto de dados compilado por Valdis Krebs[4].

- *football*: Rede de jogos de futebol americano entre equipes colegiais da divisão I durante a temporada de 2000, com as equipes sendo divididas em 12 conferências. Os vértices são as equipes e as arestas ligam equipes que se enfrentaram. Conjunto compilado por Girvan & Newman (2002).

- *blogs*: Rede de citação entre blogs políticos no EUA, com os dados divididos em conservadores e liberais, registrados em 2005 (Adamic & Glance, 2005).

[2]http://www-personal.umich.edu/ mejn/netdata/
[3]http://netkit-srl.sourceforge.net/
[4]http://www.orgnet.com/

- *industry-yh*: Rede de industrias relacionadas por co-ocorrência em notícias obtidas na web entre 4/1/1999 e 8/4/1999, com as industrias divididas em 12 setores que representam as classes. Conjunto de dados compilado por Fawcett & Provost (1999).

As redes foram consideradas sem peso, não direcionadas e sem arestas duplicadas entre pares de vértices. Para extrair a similaridade entre os exemplos utilizou-se a medida Jaccard, utilizada em grafos por Choi & Krishnamoorthy (2007). Na Tabela 4.4 são apresentados detalhes de cada conjunto.

Tabela 4.4: Conjuntos de dados relacionais

Conjunto de dados	#Vértices	#Arestas	#Classes	Erro majoritário
books	105	441	Conservador (46,7%) Liberal (40,9%) Neutro (12,4%)	53,3%
football	115	613	*0* (7,8%) 1 (7,0%) 2 (9,6%) 3 (10,4%) 4 (8,7%) 5 (4,3%) 6 (11,3%) 7 (7,0%) 8 (8,7%) 9 (10,4%) 10 (6,1%) 11 (8,7%)	88,7%
blogs	1222	16714	Liberal (48,0%) Conservador (52,0%)	48,0%
industry-yh	1798	14146	*BasicMaterials* (5,8%) *CapitalGoods* (4,6%) *Conglomerates* (0,8%) *ConsumerCyclical* (5,5%) *ConsumerNonCyclical* (3,3%) *Energy* (3,9%) *Financial* (9,5%) *Healthcare* (10,0%) *Services* (24,7%) *Technology* (28,1%) *Transportation* (2,1%) *Utilities* (1,7%)	71,9%

4.2.2 Resultados obtidos

A seguir serão descritos os resultados obtidos na avaliação das redes, suas propriedades e os classificadores propostos. Para isso, utilizaremos a seguinte nomenclatura para simplificar a exibição dos resultados:

- RH: rede hierárquica.

- RHD e RHP: rede hierárquica determinística e probabilística.
- RHD(g) e RHP(g): rede hierárquica determinística e probabilística construída com grau médio de entrada igual a g.
- kA e kAO: rede k-Associados e k-Associados ótima.
- kA(k): rede k-Associados construída com valor de k de entrada.

Pureza das redes

Como já mencionado, as redes são avaliadas a partir de medidas de pureza, compreendidas como a pureza dos dados ou da rede original. A Tabela A.1 contém a pureza média dos conjuntos de dados, das redes hierárquicas, das redes k-Associados e das redes k-Associados ótima, todas considerando a pureza média das redes construídas com os treinos gerados pelas três execuções do particionamento 10-*fold* já descrito.

A coluna *Rede Principal* indica a pureza da rede, somente para os conjuntos relacionais, e as colunas *Conj. Original* - *min* e *max* indicam a pureza para os conjuntos proposicionais. Devido a baixa variação dos resultados, em geral, independente dos valores de entrada definidos (grau médio para redes hierárquicas, k para redes k-Associados e número de vizinhos para o kNN), optou-se por exibir apenas o menor e o maior valor para cada caso. Sendo assim, não estão exibidas as purezas das redes k-Associados com valores de k igual a 1, 3, 5 e 15, mas apenas a menor e a maior pureza dentre essas quatro. Também optamos por ocultar o desvio padrão devido ao maior desvio ser 0,038, e considera-se que a pureza máxima é 1 e a mínima 0. Os resultados completos são apresentados no Apêndice A.

Na maioria dos casos a pureza máxima ocorreu na rede formada com número de vizinhos e grau médio de entrada igual a 15, e a pureza mínima com número de vizinhos e grau médio de entrada igual a 1. Além disso, destacamos a diferença significativa entre os métodos de obtenção de pureza do conjunto original e utilizando as RH, pois no caso do conjunto original consideramos, para cada exemplo, sempre a mesma quantidade de vizinhos mais próximos, e no caso da RH, mesmo que a média do número de adjacentes de cada exemplo seja muito próxima do valor de k, há uma discrepância muito grande do grau de cada vértice, o que poderá ser observado a seguir nas propriedades extraídas das redes, com o grau médio, grau mínimo e grau máximo de cada rede.

Observa-se também, que a pureza dos conjuntos relacionais não estão próximas do conjunto original, possivelmente a similaridade de Jaccard usada não é uma boa medida para se utilizar na construção das redes. Na Figura 4.3 são apresentadas quatro redes, (a) e (b) do conjunto relacional *Books*, sendo (a) sua rede original e (b) a RHD com grau médio de entrada 5, (c) e (d) as RHD dos conjuntos *Iris* e *Chemistry*, respectivamente, com grau médio de entrada também 5. Observando a figura nota-se que realmente a rede

Tabela 4.5: Pureza das redes

Conjunto de dados	Rede Principal	Conj. Original		RH		kA		kAO
		min	max	min	max	min	max	
Balance	–	0,776	0,785	0,739	0,779	0,774	0,785	0,871
Ecoli	–	0,767	0,81	0,679	0,799	0,763	0,808	0,790
Glass	–	0,757	0,911	0,773	0,866	0,75	0,91	0,865
Ionosphere	–	0,942	0,989	0,904	0,98	0,937	0,989	0,955
Iris	–	0,929	0,953	0,906	0,934	0,922	0,953	0,950
Sonar	–	0,639	0,875	0,688	0,807	0,629	0,868	0,757
Wdbc	–	0,933	0,953	0,923	0,948	0,931	0,952	0,948
Wine	–	0,916	0,949	0,901	0,924	0,915	0,951	0,946
Yeast	–	0,467	0,53	0,439	0,5	0,464	0,526	0,489
Zoo	–	0,796	0,967	0,847	0,896	0,779	0,962	0,973
Books	0,819	–	–	0,668	0,7	0,686	0,728	0,714
Football	0,636	–	–	0,082	0,099	0,079	0,099	0,087
Blogs	0,904	–	–	0,632	0,654	0,622	0,656	0,640
Industry-yh	0,457	–	–	0,186	0,19	0,189	0,192	0,195
CBR-ILP-IR	–	0,929	0,979	0,835	0,956	0,925	0,977	0,962
Chemistry	–	0,853	0,965	0,811	0,933	0,844	0,963	0,919
CS	–	0,755	0,917	0,674	0,855	0,742	0,905	0,826
Physics	–	0,844	0,971	0,791	0,943	0,834	0,968	0,919

para o conjunto relacional não obteve uma separação dos exemplos tão boa quanto os conjuntos proposicionais.

Caracterização das redes

A aplicação de propriedades de redes complexas para caracterização das redes gera uma quantidade muito grande de informações, tentamos aqui apresentá-las de forma que possibilitem a identificação de padrões e extração de algum conhecimento útil. Sendo assim, optamos pela apresentação dos resultados considerando o menor e o maior conjunto de dados para cada tipo de conjunto, numérico, textual e relacional, inserindo os resultados completos no Apêndice B.

Primeiramente, são apresentadas as informações referentes ao grau dos vértices, contendo o grau mínimo, máximo e médio da rede na Tabela 4.6. Nas tabelas mostradas aqui, as linhas indicadas com as letras D e P são referentes as redes hierárquicas determinísticas e probabilísticas, respectivamente.

As redes construídas a partir dos conjuntos de dados com maior quantidade de exemplos alcançaram o grau médio muito próximo do grau médio dado como entrada, com as redes hierárquicas determinísticas se aproximando ainda mais do que as probabilísticas. Os conjuntos menores constroem uma rede conexa rapidamente, antes que a quantidade de aresta necessária para atingir o grau médio seja inserida.

A Tabela 4.7 contém informações relacionadas a caminhos nas redes. As médias dos menores caminhos obtidas são valores baixos até mesmo para redes com os maiores conjuntos de dados. Como exemplo, é possível observar o conjunto de dados relacional *Industry-yh*, nas redes hierárquicas determinística e probabilística com grau médio igual a 15,0 e

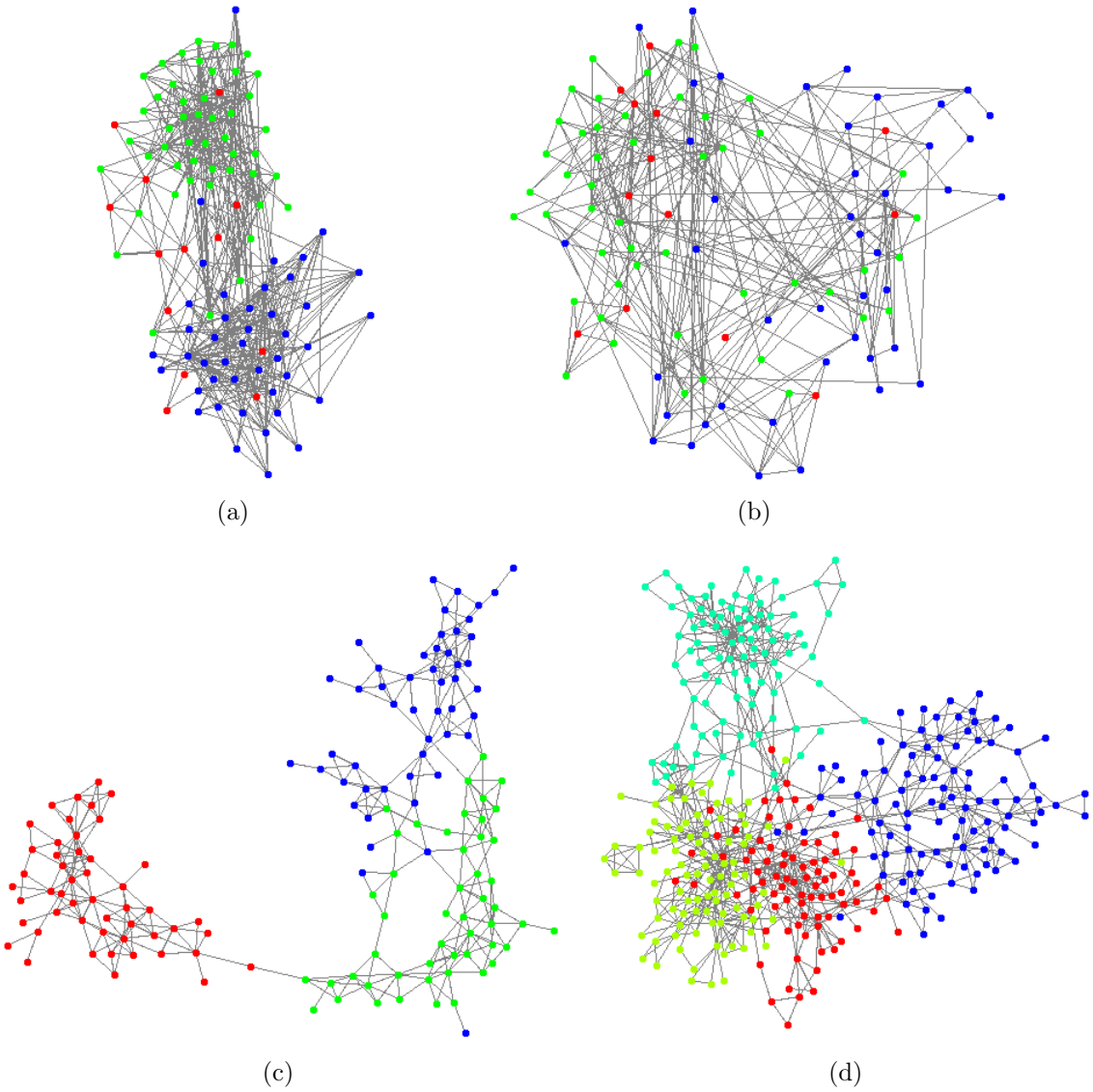


Figura 4.3: Redes com as cores representando as classes dos exemplos. (a) e (b) apresentam as redes do conjunto de dados *Books*, sendo respectivamente a rede original do conjunto e a RHD com grau médio 5, (c) apresenta a RHD com grau médio 5 para o conjunto *Iris* e (d) a RHD também com grau médio 5 para o conjunto *Chemistry*.

13,33, respectivamente, as médias do menor caminho foram 3,39 e 3,45, muito próximas da rede original, que é 3,41 com grau médio igual a 15,74.

A seguir são apresentados os resultados do coeficiente de agrupamento e da modularidade Q das redes, na Tabela 4.8. São atingidos altos valores mesmo com grau médio de entrada 3 e 5, não necessitando de uma alta quantidade de arestas inseridas. Porém, os valores de coeficiente de agrupamento e da modularidade Q obtidos não tem grande importância em casos nos quais a pureza da rede é baixa, como os conjuntos relacionais, pois haveria muitos triângulos e boas estruturas de comunidades, mas ligando exemplos de diferentes classes.

Por fim, são apresentados os gráficos de distribuição do grau para todos os conjuntos de

Tabela 4.6: Grau mínimo, máximo e médio das redes

Conjunto de dados	(grau médio | maior grau | menor grau) **Rede Conj.**		**RH(1)**	**RH(3)**	**RH(5)**	**RH(15)**
Yeast		D	2,24|10|1	3,00|24|1	5,00|61|1	14,70|287|1
		P	2,07|6|1	3,00|10|1	5,00|42|1	13,95|139|1
Zoo		D	2,97|13|1	3,17|12|1	4,75|12|1	6,71|18|1
		P	2,46|7|1	3,01|7|1	4,71|12|1	7,45|22|1
Books	8,40|25|2	D	2,50|6|1	3,01|7|1	4,59|9|1	5,64|14|1
		P	2,46|6|1	2,95|6|1	4,46|10|1	6,74|17|1
Industry-yh	15,74|250|1	D	2,20|10|1	3,01|36|1	5,00|82|1	15,00|173|1
		P	2,04|7|1	3,01|13|1	5,01|23|1	13,33|189|1
CBR-ILP-IR		D	2,31|9|1	2,99|17|1	4,98|31|1	14,44|60|1
		P	2,09|6|1	3,00|9|1	4,86|17|1	13,07|55|1
Chemistry		D	2,32|9|1	3,02|11|1	4,90|30|1	12,49|58|1
		P	2,16|7|1	2,99|14|1	4,80|24|1	12,17|71|1

Tabela 4.7: Média do menor caminho e diâmetro das redes

Conjunto de dados	(média do menor caminho | diâmetro) **Rede Conj.**		**RH(1)**	**RH(3)**	**RH(5)**	**RH(15)**
Yeast		D	8,38|19	6,46|16	4,84|11	3,28|8
		P	9,38|22	6,63|15	4,85|11	3,24|7
Zoo		D	4,32|10	3,98|8	3,29|8	2,89|7
		P	4,77|11	4,10|9	3,26|8	2,72|6
Books	2,76|5	D	5,20|12	4,75|12	3,50|9	3,17|8
		P	5,40|13	4,46|10	3,60|8	2,92|7
Industry-yh	3,41|9	D	9,08|23	6,93|17	5,10|12	3,39|8
		P	9,79|22	6,74|16	4,98|12	3,45|7
CBR-ILP-IR		D	7,15|19	5,72|13	4,22|9	2,76|5
		P	8,62|24	5,80|14	4,27|10	2,83|6
Chemistry		D	8,30|20	5,97|14	4,40|10	3,14|8
		P	8,70|24	6,28|16	4,34|11	2,88|7

Tabela 4.8: Coeficiente de agrupamento e modularidade Q das redes

Conjunto de dados	(coeficiente de agrupamento | modularidade Q) **Rede Conj.**		**RH(1)**	**RH(3)**	**RH(5)**	**RH(15)**
Yeast		D	0,09|0,94	0,24|0,88	0,40|0,81	0,46|0,64
		P	0,01|0,93	0,06|0,77	0,10|0,63	0,19|0,54
Zoo		D	0,35|0,79	0,40|0,78	0,51|0,76	0,54|0,67
		P	0,11|0,76	0,24|0,76	0,38|0,70	0,47|0,60
Books	0,49|0,51	D	0,17|0,79	0,34|0,77	0,43|0,69	0,48|0,67
		P	0,12|0,76	0,18|0,74	0,37|0,68	0,46|0,63
Industry-yh	0,17|0,28	D	0,10|0,95	0,21|0,86	0,39|0,74	0,42|0,50
		P	0,00|0,94	0,03|0,71	0,07|0,55	0,20|0,46
CBR-ILP-IR		D	0,10|0,90	0,18|0,84	0,34|0,74	0,33|0,59
		P	0,01|0,89	0,05|0,75	0,10|0,66	0,18|0,53
Chemistry		D	0,11|0,88	0,24|0,82	0,38|0,73	0,43|0,61
		P	0,01|0,87	0,11|0,78	0,22|0,68	0,25|0,55

dados. E devido ao fato das redes hierárquicas determinísticas e probabilísticas geradas com o mesmo grau médio de entrada demonstrarem uma distribuição do grau muito semelhante, optamos por exibir apenas as curvas das redes hierárquicas determinísticas para evitar que o gráfico contenha muita informação e dificulte a leitura. Também visando

facilitar a leitura dos gráficos, todos eles foram exibidos com o grau máximo 50.

Os gráficos dos conjuntos numéricos são apresentados na Figura 4.4, dos conjuntos textuais na Figura 4.5, e dos conjuntos relacionais na Figura 4.6.

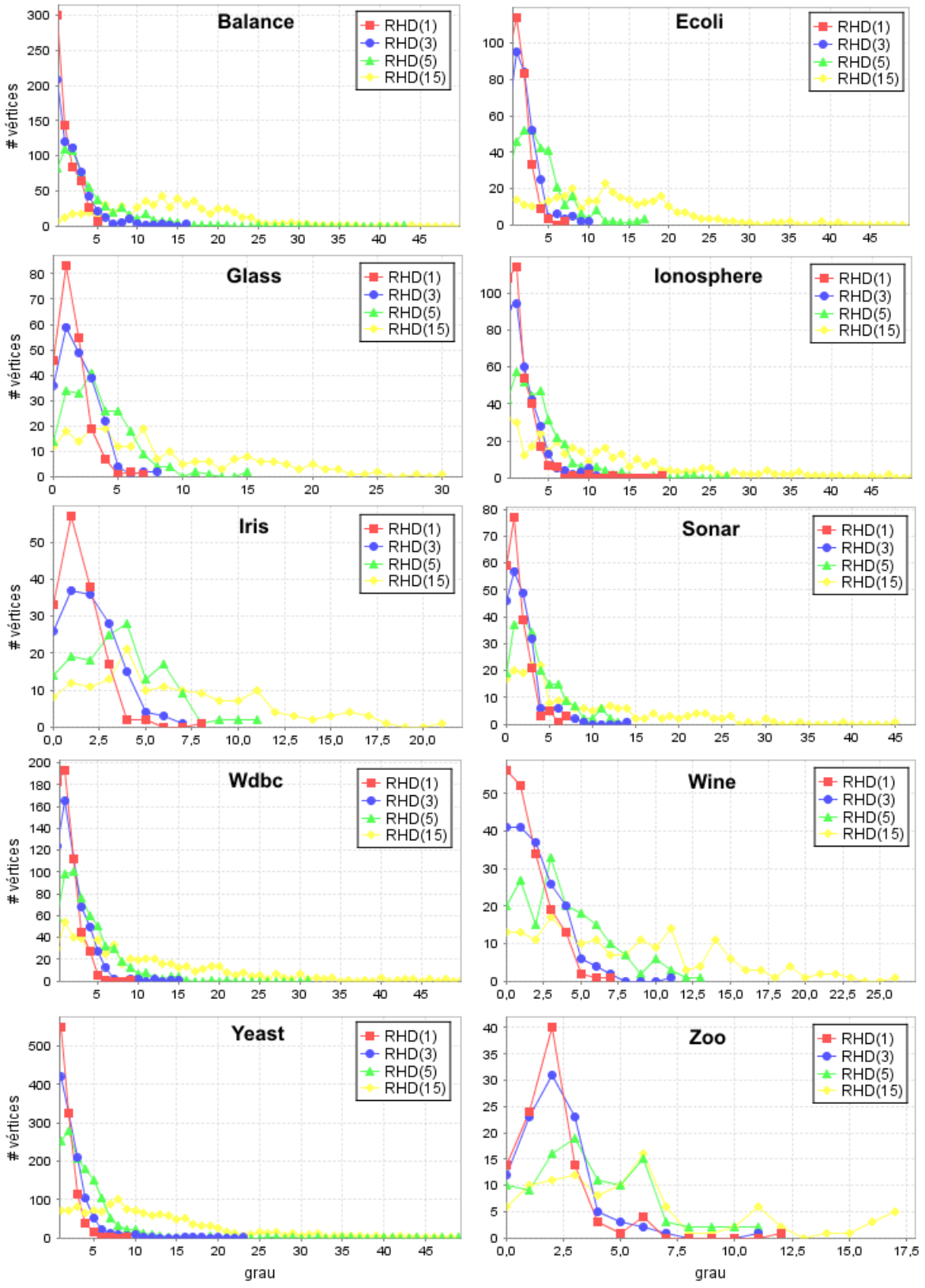


Figura 4.4: Distribuição do grau para os conjuntos atributo-valor numéricos.

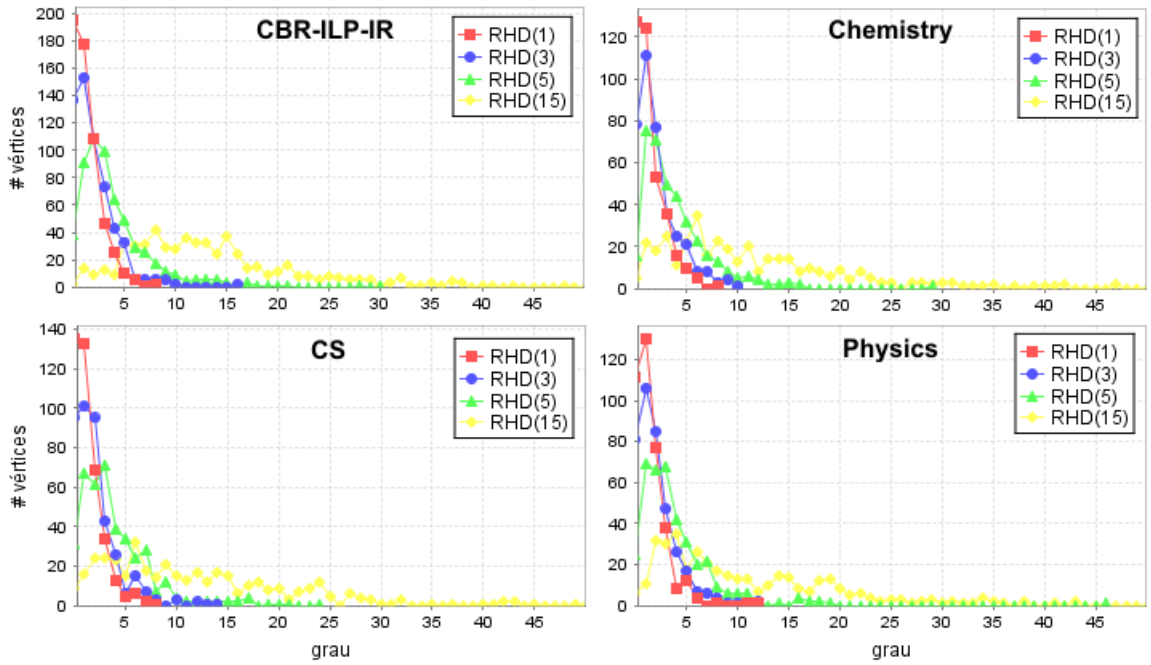


Figura 4.5: Distribuição do grau para os conjuntos atributo-valor textuais.

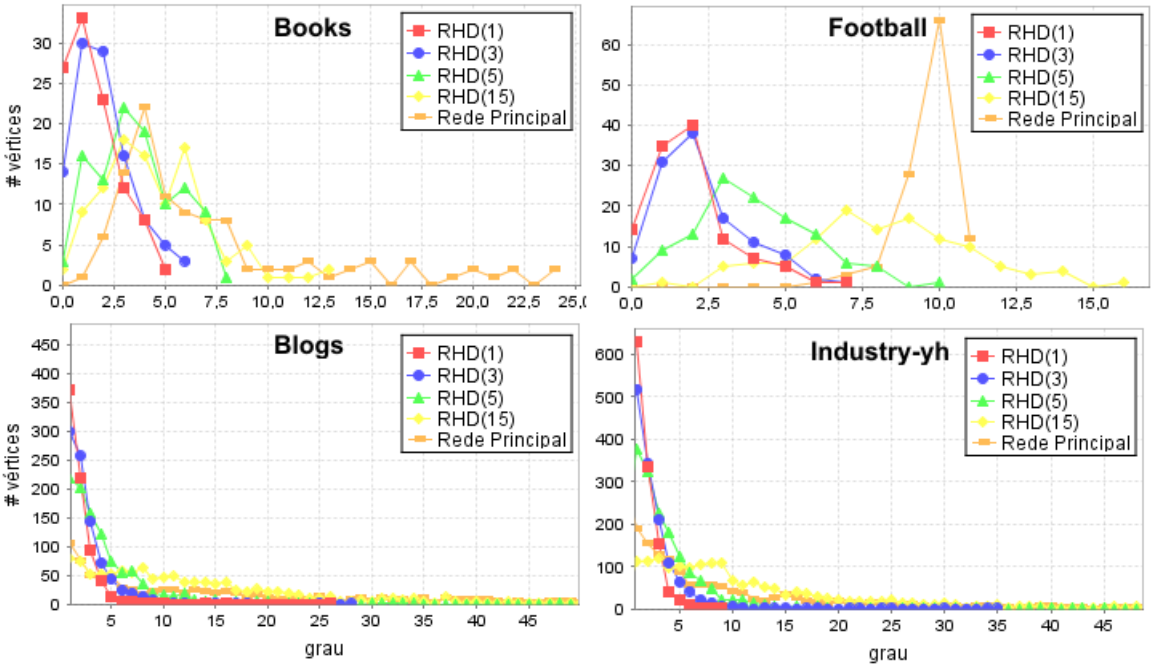


Figura 4.6: Distribuição do grau para os conjuntos relacionais.

A distribuição do grau apresenta uma curva aproximada do modelo livre de escala para os graus de entrada 3 e 5, e, para o grau de entrada 15, um modelo próximo do mundo pequeno em alguns casos (por exemplo nos conjuntos *Balance* e *Ecoli*) e entre o livre de escala e o mundo pequeno em outros casos (por exemplo nos conjuntos *Ionosphere* e *Wdbc*), apresentando um decaimento na curva, porém não em lei de potência. Essas observações são melhores identificadas nos conjuntos com maiores quantidade de exemplos.

Classificadores

Os resultados são apresentados separados em três partes, na primeira são apresentados os classificadores que consideram as redes hierárquicas, na segundo os classificadores que consideram as redes k-associados, e a última apresentando os resultados do classificador k-associados baseado em comitê, pois este não classifica todos exemplos e necessita ser analisado separadamente.

Em todas etapas os classificadores são comparados com o classificador k-NN (k = 1, 3, 5 e 15) e Naive Bayes, para os conjuntos proposicionais, e nBC, para os conjuntos relacionais. A partir dos resultados foi utilizado o teste estatístico não-paramétrico de Kruskal-Wallis para determinação de significância estatística na diferença de desempenho dos métodos de classificação, aplicando o pós-teste de múltiplas comparações de Dunn, obtendo os pares de classificadores nos quais um possui diferença estatística significativa comparado ao outro.

Classificadores baseados em redes hierárquicas

Nessa primeira parte dos resultados de classificação são apresentados os erros dos classificadores que utilizam as redes hierárquicas determinísticas e probabilísticas. Portanto, são aplicados os classificadores cbRN e o nBC utilizando as redes hierárquicas construídas. Todos os resultados são a média obtida pela aplicação de cada classificador em três execuções do processo de validação cruzada 10-*fold* já descrito.

Nas Tabelas 4.9 e 4.10 são apresentados o erro de cada classificador aplicado as redes hierárquicas determinísticas e probabilísticas, respectivamente. Além dos resultados do k-NN, Naive Bayes, e nBC para comparação. Resultados em negrito representam o menor erro para o conjunto de dados, e em cinza são os classificadores que não apresentaram diferença estatística para o classificador de menor erro, sendo o erro apresentado em porcentagem, assim como o desvio padrão.

Observa-se que os classificadores baseados nas redes hierárquicas determinísticas obtiveram um comportamento melhor comparado aos baseados nas redes hierárquicas probabilísticas. E comparando com os classificadores k-NN, Naive Bayes e nBC, o cbRH para as RHD estiveram entre os melhores estatisticamente em 15 dos 18 conjuntos de dados, contra 9 casos para o nBC.

Classificadores baseados em redes k-associados

A seguir são apresentados os erros dos classificadores que utilizam as redes k-associados e k-associados ótima. Os classificadores são *visão teste-rede* (Teste-Rede), *visão rede-teste* (Rede-Teste), e *baseado em comitê com maior probabilidade* (Maior Prob.). O classificador *baseado em comitê* é apresentado separadamente devido a não classificar todos exemplos de teste.

Tabela 4.9: Erros em porcentagem dos classificadores com redes hierárquicas determinísticas

Conjunto de dados	Naive Bayes	nBC Rede	kNN (1) erro	kNN (3) erro	kNN (5) erro	kNN (15) erro	cbRH(15) erro	nBC(15) erro
Balance	**9,3±3,6**	–	21,2±4,3	19,3±3,8	15,9±4,2	10,0±4,2	18,3±5,1	12,7±0,9
Ecoli	**14,0±4,9**	–	18,9±5,9	15,9±4,9	14,2±4,5	14,9±5,8	15,2±4,6	21,8±1,4
Glass	16,1±10,2	–	10,1±6,0	**9,8±5,3**	12,6±6,0	18,2±9,7	13,7±6,9	18,8±1,6
Ionosphere	6,5±5,4	–	1,1±1,6	1,2±1,8	2,4±2,5	8,3±4,3	3,7±3,1	2,2±0,7
Iris	4,4±5,6	–	4,7±4,7	4,7±4,7	**4,2±5,1**	4,4±5,9	4,4±5,1	5,3±1,3
Sonar	31,4±10,8	–	**12,9±6,9**	16,6±8,1	17,1±8,7	31,3±12,1	23,6±8,9	24,1±2,5
Wdbc	6,6±2,8	–	4,5±2,9	3,0±2,1	3,2±2,4	**2,8±1,9**	3,5±2,3	4,5±0,4
Wine	3,2±4,1	–	4,7±5,7	3,5±4,5	4,9±4,6	4,1±4,9	4,7±5,7	4,9±1,2
Yeast	42,6±3,9	–	47,6±3,6	46,4±3,2	43,3±3,8	41,4±3,7	42,5±4,3	54,9±1,1
Zoo	**3,3±5,4**	–	4,0±6,2	8,2±8,2	6,9±8,3	13,5±11,8	5,6±6,8	14,1±1,8
Books	–	20,3±0,9	–	–	–	–	22,4±10,0	21,8±2,3
Football	–	92,2±1,3	–	–	–	–	93,6±7,3	91,0±3,0
Blogs	–	52,0±0,0	–	–	–	–	**30,6±4,7**	37,2±3,1
Industry-yh	–	71,9±0,0	–	–	–	–	76,7±3,3	71,9±0,3
CBR-ILP-IR	3,0±2,5	–	2,1±1,8	1,9±2,0	1,5±1,9	1,2±1,4	2,1±1,8	1,7±0,4
Chemistry	9,6±3,6	–	**3,7±2,6**	3,8±3,4	4,7±3,9	5,2±4,1	4,4±2,8	7,0±0,9
CS	11,1±4,0	–	9,1±4,4	11,1±5,0	10,2±4,9	9,2±5,4	8,7±4,2	15,3±0,9
Physics	3,2±2,7	–	3,4±2,8	3,5±3,7	4,6±3,6	4,5±3,2	3,0±2,6	4,3±0,8

Tabela 4.10: Erros em porcentagem dos classificadores com redes hierárquicas probabilísticas

Conjunto de dados	Naive Bayes	nBC Rede	kNN (1) erro	kNN (3) erro	kNN (5) erro	kNN (15) erro	cbRH(15) erro	nBC(15) erro
Balance	**9,3±3,6**	–	21,2±4,3	19,3±3,8	15,9±4,2	10,0±4,2	19,4±5,0	16,0±1,4
Ecoli	**14,0±4,9**	–	18,9±5,9	15,9±4,9	14,2±4,5	14,9±5,8	15,4±4,7	22,9±1,7
Glass	16,1±10,2	–	10,1±6,0	**9,8±5,3**	12,6±6,0	18,2±9,7	17,7±6,4	21,3±3,7
Ionosphere	6,5±5,4	–	1,1±1,6	1,2±1,8	2,4±2,5	8,3±4,3	5,6±4,5	5,1±2,1
Iris	4,4±5,6	–	4,7±4,7	4,7±4,7	**4,2±5,1**	4,4±5,9	**4,2±4,8**	5,5±1,6
Sonar	31,4±10,8	–	**12,9±6,9**	16,6±8,1	17,1±8,7	31,3±12,1	27,7±10,5	31,6±3,2
Wdbc	6,6±2,8	–	4,5±2,9	3,0±2,1	3,2±2,4	**2,8±1,9**	4,6±3,0	4,8±0,8
Wine	3,2±4,1	–	4,7±5,7	3,5±4,5	4,9±4,6	4,1±4,9	5,2±5,5	5,8±1,6
Yeast	42,6±3,9	–	47,6±3,6	46,4±3,2	43,3±3,8	41,4±3,7	44,5±3,2	57,5±1,5
Zoo	**3,3±5,4**	–	4,0±6,2	8,2±8,2	6,9±8,3	13,5±11,8	7,3±8,3	13,9±2,3
Books	–	20,3±0,9	–	–	–	–	22,1±10,3	20,3±2,0
Football	–	92,2±1,3	–	–	–	–	94,2±5,9	91,3±2,5
Blogs	–	52,0±0,0	–	–	–	–	**30,6±4,3**	35,1±2,8
Industry-yh	–	71,9±0,0	–	–	–	–	76,8±2,6	71,9±0,3
CBR-ILP-IR	3,0±2,5	–	2,1±1,8	1,9±2,0	1,5±1,9	1,2±1,4	2,2±2,0	7,2±2,3
Chemistry	9,6±3,6	–	**3,7±2,6**	3,8±3,4	4,7±3,9	5,2±4,1	5,6±3,2	10,2±2,2
CS	11,1±4,0	–	9,1±4,4	11,1±5,0	10,2±4,9	9,2±5,4	8,7±4,4	25,0±5,0
Physics	3,2±2,7	–	3,4±2,8	3,5±3,7	4,6±3,6	4,5±3,2	3,5±2,8	8,3±2,2

Da mesma forma que os resultados apresentados para os classificadores hierárquicos, a média também foi obtida pela aplicação dos classificadores em três execuções do processo de validação cruzada 10-*fold*. Na Tabelas 4.11 e 4.12, os resultados em negrito destacam o melhor resultado para o conjunto de dados, e em cinza são destacados todos classificadores que não apresentaram diferença estatística para o classificador de menor erro.

Tabela 4.11: Erros dos classificadores com redes k-associados ($k = 15$)

Conjunto de dados	Naive Bayes	nBC Rede	kNN (1) erro	kNN (3) erro	kNN (5) erro	kNN (15) erro	Teste-Rede erro	Rede-Teste erro	Maior Prob. erro
Balance	**9,3±3,6**	–	21,2±4,3	19,3±3,8	15,9±4,2	10,0±4,2	10,3±3,9	10,2±4,1	10,1±3,9
Ecoli	**14,0±4,9**	–	18,9±5,9	15,9±4,9	14,2±4,5	14,9±5,8	16,3±5,1	15,6±5,3	14,5±5,4
Glass	16,1±10,2	–	10,1±6,0	**9,8±5,3**	12,6±6,0	18,2±9,7	21,8±9,3	18,7±8,6	19,4±10,0
Ionosphere	6,5±5,4	–	1,1±1,6	1,2±1,8	2,4±2,5	8,3±4,3	9,1±5,1	15,1±6,4	4,1±4,3
Iris	4,4±5,6	–	4,7±4,7	4,7±4,7	**4,2±5,1**	4,4±5,9	4,4±5,9	6,0±5,6	4,9±5,5
Sonar	31,4±10,8	–	**12,9±6,9**	16,6±8,1	17,1±8,7	31,3±12,1	30,7±10,6	27,0±10,4	28,5±11,4
Wdbc	6,6±2,8	–	4,5±2,9	3,0±2,1	3,2±2,4	**2,8±1,9**	**2,8±1,9**	4,7±2,9	2,9±2,0
Wine	3,2±4,1	–	4,7±5,7	3,5±4,5	4,9±4,6	4,1±4,9	4,3±5,0	4,7±4,5	**2,6±3,5**
Yeast	42,6±3,9	–	47,6±3,6	46,4±3,2	43,3±3,8	41,4±3,7	41,8±3,5	41,6±4,1	**40,5±3,7**
Zoo	**3,3±5,4**	–	4,0±6,2	8,2±8,2	6,9±8,3	13,5±11,8	17,2±13,1	16,5±11,8	14,2±10,6
Books	–	20,3±0,9	–	–	–	–	18,6±10,7	20,5±10,4	18,9±10,6
Football	–	92,2±1,3	–	–	–	–	91,3±6,8	92,2±6,9	92,5±7,1
Blogs	–	52,0±0,0	–	–	–	–	32,0±4,2	44,7±5,5	31,8±3,7
Industry-yh	–	71,9±0,0	–	–	–	–	**71,2±3,1**	71,9±2,8	71,5±2,9
CBR-ILP-IR	3,0±2,5	–	2,1±1,8	1,9±2,0	1,5±1,9	1,2±1,4	1,2±1,4	1,9±1,8	**1,1±1,5**
Chemistry	9,6±3,6	–	**3,7±2,6**	3,8±3,4	4,7±3,9	5,2±4,1	6,2±5,2	3,9±3,1	3,8±3,3
CS	11,1±4,0	–	9,1±4,4	11,1±5,0	10,2±4,9	9,2±5,4	10,3±5,9	9,7±4,4	**7,7±3,8**
Physics	3,2±2,7	–	3,4±2,8	3,5±3,7	4,6±3,6	4,5±3,2	5,4±3,6	3,8±2,8	**2,2±2,2**

Tabela 4.12: Erros dos classificadores com redes k-associados ótima

Conjunto de dados	Naive Bayes	nBC Rede	kNN (1) erro	kNN (3) erro	kNN (5) erro	kNN (15) erro	Teste-Rede erro	Rede-Teste erro	Maior Prob. erro
Balance	**9,3±3,6**	–	21,2±4,3	19,3±3,8	15,9±4,2	10,0±4,2	13,7±3,5	11,3±3,9	12,3±3,6
Ecoli	**14,0±4,9**	–	18,9±5,9	15,9±4,9	14,2±4,5	14,9±5,8	16,9±7,5	22,4±9,1	16,8±7,6
Glass	16,1±10,2	–	10,1±6,0	**9,8±5,3**	12,6±6,0	18,2±9,7	14,3±8,3	41,7±14,0	16,2±8,6
Ionosphere	6,5±5,4	–	1,1±1,6	1,2±1,8	2,4±2,5	8,3±4,3	**0,7±1,4**	35,6±8,6	8,5±6,9
Iris	4,4±5,6	–	4,7±4,7	4,7±4,7	**4,2±5,1**	4,4±5,9	4,4±5,6	9,3±8,1	4,9±5,5
Sonar	31,4±10,8	–	**12,9±6,9**	16,6±8,1	17,1±8,7	31,3±12,1	18,9±9,7	36,6±9,9	20,5±9,9
Wdbc	6,6±2,8	–	4,5±2,9	3,0±2,1	3,2±2,4	**2,8±1,9**	4,1±2,1	17,8±5,5	3,5±2,2
Wine	3,2±4,1	–	4,7±5,7	3,5±4,5	4,9±4,6	4,1±4,9	3,4±4,0	19,8±11,6	6,1±6,4
Yeast	42,6±3,9	–	47,6±3,6	46,4±3,2	43,3±3,8	41,4±3,7	43,9±4,0	48,3±4,0	43,4±4,3
Zoo	**3,3±5,4**	–	4,0±6,2	8,2±8,2	6,9±8,3	13,5±11,8	7,6±8,5	13,6±10,4	5,9±7,7
Books	–	20,3±0,9	–	–	–	–	**18,3±10,3**	22,9±10,6	18,9±11,2
Football	–	92,2±1,3	–	–	–	–	**90,5±9,3**	93,1±7,4	93,9±6,5
Blogs	–	52,0±0,0	–	–	–	–	31,3±4,9	43,4±5,3	32,8±5,8
Industry-yh	–	71,9±0,0	–	–	–	–	80,9±3,7	75,3±5,8	74,1±4,4
CBR-ILP-IR	3,0±2,5	–	2,1±1,8	1,9±2,0	1,5±1,9	1,2±1,4	1,9±1,9	8,2±3,8	2,0±2,1
Chemistry	9,6±3,6	–	**3,7±2,6**	3,8±3,4	4,7±3,9	5,2±4,1	6,4±4,8	16,9±7,0	6,6±4,4
CS	11,1±4,0	–	9,1±4,4	11,1±5,0	10,2±4,9	9,2±5,4	10,2±4,8	22,6±7,3	10,8±4,7
Physics	3,2±2,7	–	3,4±2,8	3,5±3,7	4,6±3,6	4,5±3,2	4,3±4,4	17,7±6,4	5,1±4,0

As redes K-associados são equivalentes nas visões *rede-teste* e *teste-rede*, porém o classificador *baseado em comitê com maior probabilidade* possui 15 casos entre os melhores resultados. Considerando apenas os conjuntos proposicionais, o classificador *baseado em comitê com maior probabilidade* situa-se entre os melhores em 11 de 14 conjuntos, contra o Naive Bayes que é melhor em 8 de 14, e próximo dos resultados do melhor k-NN.

Já para as redes K-Associados ótimas, os classificadores teste-rede e baseado em comitê com maior probabilidade superam ou igualam os melhores resultados dos outros classificadores.

Classificador baseado em comitê das redes k-associados

O classificador *baseado em comitê* das redes k-associados realizam a classificação de um exemplo para casos em que os classificadores *visão teste-rede* e *visão rede-teste* concordam. Devido a isso, não necessariamente todos exemplos de teste são classificados.

Sendo assim, os resultados para o classificador *baseado em comitê* são apresentados na Tabela 4.13 para as redes k-associados e k-associados ótima, com seus respectivos valores de cobertura.

Observa-se que os erros foram bem inferiores dos classificadores comparados, porém, em alguns casos, a cobertura também cai significativamente.

Após análise dos erros de todos classificadores, é possível notar que para os conjuntos de dados numéricos houve uma variação muito grande dos melhores classificadores. Porém, para os conjuntos textuais houve um domínio do classificador *baseado em comitê com maior probabilidade* para as redes k-associados, obtendo o menor erro em 3 dos 4 conjuntos. Além disso, estatisticamente ele se apresentou entre os melhores classificadores para 15 dos 18 conjuntos.

O erro dos classificadores baseados nas redes hierárquicas e nas redes k-associados, para, respectivamente, grau médio e valor k de entrada 1, 3 e 5, são apresentados no Apêndice C.

Tabela 4.13: Erros do classificador baseado em comitê das redes k-associados e redes k-associados ótima

Conjunto de dados	**Naive Bayes**	**nBC Rede**	**kNN (1) erro**	**kNN (3) erro**	**kNN (5) erro**	**kNN (15) erro**	**Comitê (kA) erro | cob.**	**Comitê (kAO) erro | cob.**
Balance	9,3±3,6	–	21,2±4,3	19,3±3,8	15,9±4,2	10,0±4,2	8,2±4,0|0.972	7,6±3,9|0.913
Ecoli	14,0±4,9	–	18,9±5,9	15,9±4,9	14,2±4,5	14,9±5,8	11,2±4,2|0.901	12,8±6,7|0.855
Glass	16,1±10,2	–	10,1±6,0	9,8±5,3	12,6±6,0	18,2±9,7	13,7±7,9|0.884	9,3±9,1|0.609
Ionosphere	6,5±5,4	–	1,1±1,6	1,2±1,8	2,4±2,5	8,3±4,3	0,0±0,0|0.780	1,0±2,1|0.650
Iris	4,4±5,6	–	4,7±4,7	4,7±4,7	4,2±5,1	4,4±5,9	4,3±5,5|0.979	3,8±5,4|0.939
Sonar	31,4±10,8	–	12,9±6,9	16,6±8,1	17,1±8,7	31,3±12,1	23,8±10,3|0.799	14,9±11,4|0.648
Wdbc	6,6±2,8	–	4,5±2,9	3,0±2,1	3,2±2,4	2,8±1,9	1,6±1,2|0.957	1,8±1,5|0.810
Wine	3,2±4,1	–	4,7±5,7	3,5±4,5	4,9±4,6	4,1±4,9	2,0±3,2|0.947	2,3±3,6|0.807
Yeast	42,6±3,9	–	47,6±3,6	46,4±3,2	43,3±3,8	41,4±3,7	36,6±4,1|0.803	35,1±5,1|0.624
Zoo	3,3±5,4	–	4,0±6,2	8,2±8,2	6,9±8,3	13,5±11,8	11,0±11,1|0.893	0,7±2,8|0.847
Books	–	20,3±0,9	–	–	–	–	16,0±10,9|0.949	16,9±11,1|0.931
Football	–	92,2±1,3	–	–	–	–	92,8±8,1|0.717	95,3±9,8|0.427
Blogs	–	52,0±0,0	–	–	–	–	18,3±6,5|0.361	25,0±8,4|0.516
Industry-yh	–	71,9±0,0	–	–	–	–	70,0±3,4|0.703	69,8±7,4|0.325
CBR-ILP-IR	3,0±2,5	–	2,1±1,8	1,9±2,0	1,5±1,9	1,2±1,4	0,4±0,8|0.979	1,0±1,3|0.921
Chemistry	9,6±3,6	–	3,7±2,6	3,8±3,4	4,7±3,9	5,2±4,1	1,7±2,7|0.931	2,2±2,8|0.821
CS	11,1±4,0	–	9,1±4,4	11,1±5,0	10,2±4,9	9,2±5,4	3,8±2,9|0.880	5,2±3,9|0.775
Physics	3,2±2,7	–	3,4±2,8	3,5±3,7	4,6±3,6	4,5±3,2	1,2±1,6|0.933	1,3±2,3|0.813

Capítulo

5

Conclusões

A grande maioria dos algoritmos de aprendizado de máquina utiliza dados estruturados em uma representação proposicional para construção de modelos computacionais. Tal representação limita-se a descrever características individuais dos objetos representados, não levando em consideração relações existentes entre os objetos. Com os dados representados relacionalmente é possível agregar conceitos e técnicas de redes complexas no processo de descoberta de conhecimento.

Neste trabalho foram apresentadas as tarefas e os resultados alcançados durante esta investigação. Três aspectos desta investigação devem ser destacados. O primeiro diz respeito a construção de uma representação relacional a partir de conjuntos de dados nos quais é possível calcular um grau de similaridade entre os exemplos, denominada Rede Hierárquica. O segundo diz respeito ao algoritmo de classificação baseado nas redes hierárquicas. E o terceiro aspecto é relacionado ao modelo de redes denominado K-Associados.

Para avaliação das técnicas propostas utilizou-se uma medida de pureza para avaliar as redes construidas, e para os classificadores foi realizada a comparação com os classificadores *k-nearest neighbors* (k-NN), Naive Bayes, e o Classificador Bayesiano baseado apenas na rede (nBC). Com todos procedimentos de avaliação sendo realizados efetuando três execuções da validação cruzada 10-*fold*, para obtenção de uma média mais estável dos resultados.

As redes hierárquicas construídas a partir dos conjuntos numéricos e conjuntos textuais, utilizando, respectivamente, distância euclidiana e similaridade cosseno como medidas de similaridade entre os exemplos, apresentaram valores de pureza próximos da pureza dos conjuntos de dados, o que demonstra que as redes refletem, isto é, preservam as relações

de vizinhança do espaço original. Lembrando que a principal diferença de considerar os adjacentes, na pureza da rede, e os vizinhos mais próximos, na pureza do conjunto, é que para o conjunto sempre se considera um valor constante de exemplos, já para a rede o número de adjacentes varia de 1 até dezenas ou centenas.

A pureza das redes k-associados e k-associados ótima apresentaram resultados ainda mais próximos das purezas dos conjuntos de dados, e demonstraram ser particularmente útil na avaliação da pureza dos componentes individualmente.

Considerando os conjuntos de dados relacionais, as redes, avaliadas pelas medidas de pureza, não apresentaram bons resultados, possivelmente a medida de similaridade Jaccard não é uma boa medida para se identificar similaridade entre exemplos dos conjuntos relacionais, talvez devido a esses conjuntos apresentarem um grau médio bastante elevado, fazendo com que um exemplo tenha alta similaridade com muitos outros.

Na avaliação das propriedades de redes complexas relacionadas ao grau dos vértices, aos caminhos e ao coeficiente de agrupamento e modularidade Q, foi possível observar que apresentaram um grau médio muito próximo do solicitado na construção das redes, com baixos valores de média do menor caminho e diâmetro na rede, tendo características de redes mundo pequeno. Também apresentaram alto valor de modularidade Q, demonstrando fortes estruturas de comunidades, úteis principalmente ao verificar que a pureza das redes também foi bastante alta, possibilitando a formação de comunidades com alta pureza. Em relação ao coeficiente de agrupamento foram obtidos altos valores principalmente para as redes hierárquicas determinísticas, possuindo muito mais triângulos nas redes comparadas às redes hierárquicas probabilísticas. Considerando os conjuntos relacionais, tais medidas poderiam ser computadas para as redes originais. Portanto, permitindo comparação diretas. Como já comentado as redes construídas não se mostram semelhantes, possivelmente por conta da medida de similaridade usada.

Em relação a distribuição do grau, observando os gráficos gerados, nota-se uma forte tendência a um modelo livre de escala, com uma curva seguindo uma lei de potência, para as redes hierárquicas formadas com baixos valores de grau médio de entrada, e a um modelo mundo pequeno, com a curva ligeiramente se aproximando de uma distribuição de Poison, conforme se aumenta o grau médio de entrada. Nos conjuntos relacionais observa-se que a distribuição do grau da rede original do conjunto se assemelha mais a distribuição do grau das redes hierárquicas formadas por maiores valores de grau médio de entrada, possivelmente devido ao grau médio das redes originais serem bastante elevados.

A avaliação dos classificadores foram divididas em três etapas, inicialmente avaliou-se os classificadores que consideram as redes hierárquicas, em seguida os classificadores que consideram as redes k-associados, e por fim, o classificador k-associados *baseado em comitê*, pois este não classifica todos exemplos. Nas duas primeiras etapas foi realizada uma análise estatística utilizando o teste estatístico de Kruskal-Wallis, com o pós-teste de múltiplas comparações de Dunn.

Comparando os classificadores baseados nas redes hierárquicas, cbRH e nBC, observou-se que quando consideradas as redes hierárquicas determinísticas, em geral, se obteve melhores resultados comparados às redes hierárquicas probabilísticas.

Em relação aos classificadores baseados nas redes k-associados, *visão teste-rede*, *visão rede-teste* e *baseado em comitê com maior probabilidade*, as redes k-associados apresentaram bons resultados em todos, e as redes k-associados ótima um resultado ruim no classificador *visão rede-teste*, mas melhores resultados nos classificadores *visão teste-rede* e *baseado em comitê com maior probabilidade*, esse último superando ou igualando os melhores resultados dos demais classificadores.

Tanto o classificador cbRH para as RHD construídas com grau médio 15, como o classificador *baseado em comitê com maior probabilidade* para as redes k-associados ótima, se igualaram ou superaram os melhores resultados dos outros classificadores.

O classificador *baseado em comitê* apresentou os melhores resultados, mesmo que para alguns conjuntos de dados a cobertura tenha sido baixa. Portanto, este classificador pode ser explorado em atividades nas quais o menor erro é mais importante mesmo que alguns não sejam classificados.

5.1 Principais contribuições

As principais contribuições deste trabalho estão relacionadas com:

1. desenvolvimento da técnica para criação da rede hierárquica baseada em similaridade;
2. desenvolvimento da técnica para criação da rede k-associados;
3. desenvolvimento do classificador baseado na rede hierárquica;
4. desenvolvimento dos classificadores baseados nas redes k-associados.

As técnicas comentadas deram origem às seguintes publicações:

1. Motta, R., Almeida, L. J., Lopes, A. A.: Redes probabilísticas baseadas em similaridade na exploração de comunidades. In: I Workshop on Web and Text Intelligence (SBIA-WTI08), Salvador, Brasil, pp. 1-8 (2008)
2. Motta, R., Lopes, A. A.: Rede complexa probabilística baseada em similaridade na classificação de dados com ruídos. In: I Workshop on Web and Text Intelligence (SBIA-WTI08), Salvador, Brasil, pp. 1-8 (2008)
3. Lopes, A. A. ; Bertini, J. R. ; Motta, R.; Liang, Z.. Classification Based on the Optimal K-Associated Network. In: I International Conference on Complex Sciences: Theory and Applications (Complex'2009), Xangai, China, pp. 1-12 (2009)

E às seguintes submissões:

1. Lopes, A. A., R. Motta, F. Paulovich, and R. Minghim. Objective Evaluation of Point Placement layouts based on Complex Network Measures. IEEE Computer Graphics and Applications Journal, 10 pages. (2009)

2. Motta, R., Lopes, A. A., Oliveira, M. C.: Centrality Measures from Complex Networks in Active Learning. Discovery Science, 14 pages. (2009)

As técnicas estudadas foram implementadas em um sistema denominado *ComplexNet*. Tal sistema possibilita o uso de conjuntos de dados representados por tabelas atributo-valor ou modelados em grafos, disponibilizando tarefas de preparação dos dados, algoritmos de classificação proposicional e relacional, aplicação de propriedades de redes complexas, e técnicas de visualização. As propriedades de redes complexas implementadas são referentes tanto a características individuais como globais da rede, contendo desde informações relacionadas ao vértices, como grau do vértice, coeficiente de agrupamento e medidas de vértices centrais, até medidas referentes a rede como um todo, como caminhos na rede e técnicas de identificação de comunidade.

5.2 Limitações

Este trabalho possui algumas limitações quanto à sua realização e a utilização das técnicas aqui propostas. Uma primeira limitação se refere à utilização de apenas quatro conjuntos de dados relacionais e quatro conjuntos textuais. Esta opção foi devido a utilização de três diferentes conjuntos de dados, numérico, textual e relacional, com isso buscou-se uma quantidade razoável de cada tipo de conjunto, obtendo a maior variação possível, de quantidade de exemplos, de atributo ou de classes.

Outra limitação que pode ser observada é relacionada as medidas de similaridade utilizadas para os experimentos. A definição de se utilizar distância euclidiana, similaridade cosseno e similaridade Jaccard para os conjuntos atributo-valor, textual e relacional, respectivamente, foi devido a essas medidas serem comumente utilizadas para estes tipos de conjuntos de dados, mas nada impede que se utilize outras medidas.

Além disso, muitos dos algoritmos aqui propostos possuem um alto custo de processamento, o que dificulta sua aplicação em conjuntos de dados com grande quantidade de exemplos.

5.3 Trabalhos futuros

Como proposta de trabalhos futuros, primeiramente pretende-se otimizar o processo de construção das redes propostas, redes hierárquicas e k-associados, possibilitando adicionar

novos exemplos às redes já construídas, tornando desnecessária a reconstrução de redes em conjuntos dinâmicos.

Além disso, pretende-se explorar a utilização das redes k-associados em um processo de avaliação de projeção multidimensional, aplicando medidas de pureza, proximidade entre componentes, quantidade de componentes, entre outras.

Em relação aos classificadores, pretende-se utilizar os classificadores relacionais propostos em conjuntos com classificadores que utilizam a representação atributo-valor, buscando obter melhores resultados realizando um processo de votação. Observa-se que neste trabalho os melhores classificadores foram um proposicional e um relacional, que poderiam ser utilizados em conjunto.

Também relacionado a classificação, tais técnicas poderiam ser utilizadas em tarefas de aprendizado ativo ou aprendizado semi-supervisionado baseado em certeza, principalmente considerando o classificador *baseado em comitê*, que apresentou ótimos resultados.

Por fim, pretende-se ainda avaliar o uso de técnicas de multinivel em redes em conjunto com as técnicas aqui propostas para diminuir o custo dos algoritmos, tais técnicas buscam minimizar a quantidade de vértices, arestas ou ambos nas redes para aplicação de algoritmos mais custosos.

Referências Bibliográficas

Adamic, L. & N. Glance (2005). The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, New York, NY, USA, pp. 36–43. ACM Press.

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, & A. I. Verkamo (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA, USA, pp. 307–328. American Association for Artificial Intelligence.

Bagrow, J. & E. Bollt (2005). A local method for detection communities. *Physical Review E 72*, 046108.

Balakrishnan, H. & N. Deo (2006). Discovering communities in complex networks. In R. Menezes (Ed.), *ACM Southeast Regional Conference*, pp. 280–285. ACM.

Barabási, A.-L. & R. Albert (1999). Emergence of scaling in random networks. *Science 286*(5439), 509–512.

Barrat, A., M. Barthélemy, R. Pastor-Satorras, & A. Vespignani (2004). The architeture of complex weighted networks. *Proceedings of the National Academy of Science USA 101*(11), 3747–3752.

Barrat, A. & M. Weigt (2000). On the properties of small-world network models. *European Physical Journal B 13*, 547.

Blum, A. & T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, New York, NY, USA, pp. 92–100. ACM.

Chakrabarti, S., B. Dom, & P. Indyk (1998). Enhanced hypertext categorization using hyperlinks. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, New York, NY, USA, pp. 307–318. ACM.

Choi, H.-J. & M. Krishnamoorthy (2007). Categorization of blogs through similarity analysis. In *ISI*, pp. 160–165. IEEE.

Clauset, A. (2005). Finding local community structure in networks. *Physical Review E 72*, 026132.

Clauset, A., M. Newman, & C. Moore (2004). Finding community structure in very large networks. *Physical Review E 70*(1), 066111.

Costa, L., F. A. Rodrigues, G. Travieso, & P. V. Boas (2007). Characterization of complex network: A survey of measurements. *Advances of Physics 56*(1), 167–242.

Danon, L., A. Díaz-Guilera, & A. Arenas (2006). The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment 2006*(11), 577–585.

Duch, J. & A. Arenas (2005). Community detection in complex networks using extremal optimization. *Physical Review E 72*, 027104.

Erdös, P. & A. Rényi (1959). On random graphs. *Publications Mathematicae 6*, 290–297.

Erdös, P. & A. Rényi (1960). On the evolution of random graphs. *Publication Mathematical Institute of the Hungarian Academy of Sciences 5*, 17–61.

Erdös, P. & A. Rényi (1961). On the strenght of connectedness of random graph. *Acta Mathematica Scientia Hungary 12*, 261–267.

Fawcett, T. & F. Provost (1999). Activity monitoring: Noticing interesting changes in behavior. In *In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 53–62.

Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth (1996). The kdd process for extracting useful knowledge from volumes of data. *Communication of the ACM 39*(11), 27–34.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry 40*(1), 35–41.

Gantz, J., C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, & A. Toncheva (2008). The diverse and exploding digital universe.

Geman, S. & D. Geman (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

Girvan, M. & M. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America 49*(2), 247–252.

Gustafsson, M., M. Hornquist, & A. Lombardi (2006). Comparison and validation of community structures in complex networks. *Physica A: Statistical Mechanics and its Applications 367*, 559–576.

Hosmer, D. & S. Lemeshow (1989). *Applied logistic regression*. Wiley.

Jain, A. K., M. N. Murty, & P. J. Flynn (1999). Data clustering: a review. *ACM Computing Surveys 31*(3), 264–323.

Jensen, D., J. Neville, & B. Gallagher (2004). Why collective inference improves relational classification. In *In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 593–598.

Liu, X., J. Bollen, M. L. Nelson, & H. V. de Sompel (2005). Co-authorship networks in the digital library research community. *Inf. Process. Manage. 41*(6), 1462–1480.

Lopes, A. A., J. Bertini, R. Motta, & L. Zhao (2009). Classification based on the optimal k-associated network. In *1st International Conference on Complex Sciences: Theory and Applications, Complex09*, Shanghai, China, pp. 1–12.

Lu, Q. & L. Getoor (2003). Link-based classification. In T. Fawcett, N. Mishra, T. Fawcett, & N. Mishra (Eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 496–503. AAAI Press.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development 2*(2), 159–165.

Macskassy, S. & F. Provost (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research 8*, 935–983.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Education (ISE Editions).

Newman, M. (2001). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E 64*(1), 016132.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review 45*, 167–256.

Newman, M. (2004a). Detecting community structure in networks. *European Physical Journal B 38*(2), 321–330.

Newman, M. (2004b). Fast algorithm for detecting community structure in networks. *Physical Review E 69*, 066133.

Newman, M. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E 74*(1), 036104.

Oliveira, M. C. F. & H. Levkowitz (2003). From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics 09*(3), 378–394.

Quinlan, J. R. (1986). Induction of decision trees. In *Machine Learning*, pp. 81–106.

Radicchi, F., C. Castellano, F. Cecconi, V. Loreto, & D. Parisi (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America 101*(9), 2658–2663.

Raedt, L. D. (2008). *Logical and Relational Learning*. Cognitive Technologies. Springer.

Rodrigues, F. A. (2007). *Caracterização, classificação e análise de redes complexas*. Tese de Doutorado, USP, São Carlos, SP. Tese de Doutorado, ICMC-USP.

Wakita, K. & T. Tsurumi (2007). Finding community structure in mega-scale social networks. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pp. 1275–1276. ACM.

Washio, T., L. D. Raedt, & J. N. Kok (2004). Advances in mining graphs, trees and sequences: Preface. *Fundam. Inf. 66*(1-2), 5–8.

Wasserman, S. & K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Watts, D. & S. Strogatz (1998). Collective dynamics of small-world networks. *Nature 393*(6684), 440–442.

Waxman, B. (1988). Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications 6*(9), 1617–1622.

Zhang, P., M. Li, J. Wu, Z. Di, & Y. Fan (2006). The analysis and dissimilarity comparison of community structure. *Proceedings of the National Academy of Sciences 367*, 577–585.

Zheng, Z. (1998). Naive bayesian classifier committees. In *Proceedings of the 10th European Conference on Machine Learning*, pp. 196–207. Springet-Verlag.

Apêndice

A

Tabelas com os resultados da avaliação das redes

Este apêndice contém as tabelas completas das purezas obtidas para as redes construídas comparadas com a pureza dos conjuntos de dados. A Tabela A.1 contém a pureza dos conjuntos de dados, a Tabela A.2 contém a pureza das redes hierárquicas determinísticas, a Tabela A.3 contém a pureza das redes hierárquicas probabilísticas, a Tabela A.4 contém a pureza das redes k-associados e k-associados ótima.

Tabela A.1: Pureza dos conjuntos de dados

Conjunto de dados	**Rede Principal**	**kNN(1)**	**kNN(3)**	**kNN(5)**	**kNN(15)**
Balance	–	0,783(0,012)	0,785(0,005)	0,786(0,003)	0,776(0,002)
Ecoli	–	0,81(0,0)	0,803(0,0)	0,799(0,0)	0,767(0,0)
Glass	–	0,911(0,0)	0,88(0,0)	0,835(0,0)	0,757(0,0)
Ionosphere	–	0,989(0,0)	0,985(0,0)	0,973(0,0)	0,942(0,0)
Iris	–	0,953(0,0)	0,944(0,0)	0,941(0,0)	0,929(0,0)
Sonar	–	0,875(0,0)	0,837(0,0)	0,795(0,0)	0,639(0,0)
Wdbc	–	0,953(0,0)	0,953(0,0)	0,949(0,0)	0,933(0,0)
Wine	–	0,949(0,0)	0,948(0,0)	0,935(0,0)	0,916(0,0)
Yeast	–	0,53(0,0)	0,499(0,0)	0,493(0,0)	0,467(0,0)
Zoo	–	0,967(0,006)	0,926(0,002)	0,916(0,001)	0,796(0,001)
Books	0,819	–	–	–	–
Football	0,636	–	–	–	–
Blogs	0,904	–	–	–	–
Industry-yh	0,457	–	–	–	–
CBR-ILP-IR	–	0,979(0,0)	0,965(0,0)	0,959(0,0)	0,929(0,0)
Chemistry	–	0,965(0,0)	0,933(0,0)	0,912(0,0)	0,853(0,0)
CS	–	0,917(0,0)	0,853(0,0)	0,829(0,0)	0,755(0,0)
Physics	–	0,971(0,0)	0,938(0,0)	0,915(0,0)	0,844(0,0)

Tabela A.2: Pureza das redes hierárquicas determinísticas

Conjunto de dados	**RHD(1)**	**RHD(3)**	**RHD(5)**	**RHD(15)**
Balance	0,779(0,011)	0,777(0,01)	0,772(0,01)	0,764(0,007)
Ecoli	0,799(0,014)	0,797(0,015)	0,784(0,013)	0,71(0,039)
Glass	0,866(0,01)	0,849(0,012)	0,841(0,015)	0,817(0,015)
Ionosphere	0,98(0,004)	0,98(0,0030)	0,973(0,004)	0,962(0,009)
Iris	0,934(0,01)	0,931(0,012)	0,921(0,014)	0,924(0,012)
Sonar	0,807(0,015)	0,8(0,013)	0,77(0,012)	0,725(0,01)
Wdbc	0,948(0,005)	0,944(0,006)	0,94(0,006)	0,937(0,004)
Wine	0,923(0,01)	0,923(0,01)	0,918(0,007)	0,924(0,007)
Yeast	0,5(0,0080)	0,498(0,008)	0,487(0,008)	0,453(0,006)
Zoo	0,896(0,011)	0,887(0,013)	0,849(0,019)	0,847(0,018)
Books	0,687(0,023)	0,684(0,026)	0,665(0,024)	0,668(0,025)
Football	0,088(0,015)	0,086(0,013)	0,099(0,011)	0,09(0,007)
Blogs	0,651(0,007)	0,654(0,008)	0,652(0,007)	0,642(0,006)
Industry-yh	0,186(0,005)	0,185(0,004)	0,185(0,005)	0,186(0,003)
CBR-ILP-IR	0,956(0,005)	0,955(0,005)	0,952(0,006)	0,924(0,008)
Chemistry	0,933(0,007)	0,926(0,009)	0,904(0,008)	0,887(0,007)
CS	0,855(0,009)	0,836(0,009)	0,8(0,009)	0,756(0,008)
Physics	0,943(0,008)	0,936(0,008)	0,912(0,008)	0,871(0,006)

Tabela A.3: Pureza das redes hierárquicas probabilísticas

Conjunto de dados	**RHP(1)**	**RHP(3)**	**RHP(5)**	**RHP(15)**
Balance	0,763(0,011)	0,76(0,012)	0,761(0,009)	0,739(0,015)
Ecoli	0,799(0,017)	0,789(0,014)	0,773(0,012)	0,679(0,033)
Glass	0,831(0,014)	0,83(0,015)	0,812(0,014)	0,773(0,033)
Ionosphere	0,979(0,005)	0,977(0,005)	0,971(0,007)	0,904(0,025)
Iris	0,931(0,009)	0,925(0,01)	0,909(0,018)	0,906(0,017)
Sonar	0,737(0,019)	0,737(0,02)	0,726(0,019)	0,688(0,016)
Wdbc	0,944(0,008)	0,944(0,006)	0,944(0,006)	0,923(0,01)
Wine	0,924(0,011)	0,924(0,009)	0,912(0,013)	0,901(0,016)
Yeast	0,459(0,014)	0,458(0,011)	0,454(0,008)	0,439(0,01)
Zoo	0,894(0,012)	0,883(0,015)	0,855(0,016)	0,852(0,016)
Books	0,669(0,031)	0,668(0,029)	0,685(0,023)	0,7(0,019)
Football	0,097(0,026)	0,094(0,021)	0,089(0,011)	0,082(0,006)
Blogs	0,639(0,009)	0,642(0,01)	0,637(0,008)	0,632(0,008)
Industry-yh	0,189(0,009)	0,189(0,007)	0,19(0,0060)	0,187(0,005)
CBR-ILP-IR	0,921(0,009)	0,927(0,01)	0,931(0,007)	0,835(0,038)
Chemistry	0,894(0,012)	0,898(0,011)	0,886(0,013)	0,811(0,034)
CS	0,794(0,015)	0,787(0,018)	0,788(0,0090)	0,674(0,021)
Physics	0,901(0,015)	0,906(0,009)	0,894(0,009)	0,791(0,034)

Tabela A.4: Pureza das redes k-associados e k-associados ótima

Conjunto de dados	**kA(1)**	**kA(3)**	**kA(5)**	**kA(15)**	**kAO**
Balance	0,784(0,013)	0,784(0,007)	0,785(0,006)	0,774(0,005)	0,871(0,016)
Ecoli	0,808(0,011)	0,803(0,009)	0,794(0,009)	0,763(0,008)	0,79(0,01)
Glass	0,91(0,01)	0,871(0,009)	0,823(0,007)	0,75(0,006)	0,865(0,018)
Ionosphere	0,989(0,003)	0,982(0,003)	0,97(0,003)	0,937(0,003)	0,955(0,009)
Iris	0,953(0,009)	0,944(0,009)	0,943(0,007)	0,922(0,007)	0,95(0,007)
Sonar	0,868(0,009)	0,826(0,008)	0,777(0,01)	0,629(0,007)	0,757(0,019)
Wdbc	0,952(0,005)	0,952(0,005)	0,949(0,004)	0,931(0,003)	0,948(0,004)
Wine	0,951(0,006)	0,946(0,005)	0,934(0,005)	0,915(0,005)	0,946(0,005)
Yeast	0,526(0,008)	0,497(0,007)	0,49(0,005)	0,464(0,005)	0,489(0,006)
Zoo	0,962(0,013)	0,923(0,009)	0,908(0,01)	0,779(0,011)	0,973(0,024)
Books	0,728(0,031)	0,686(0,02)	0,695(0,018)	0,704(0,015)	0,714(0,021)
Football	0,091(0,021)	0,099(0,011)	0,096(0,008)	0,079(0,004)	0,087(0,023)
Blogs	0,656(0,011)	0,656(0,006)	0,644(0,004)	0,622(0,004)	0,64(0,005)
Industry-yh	0,19(0,005)	0,19(0,004)	0,192(0,003)	0,189(0,002)	0,195(0,007)
CBR-ILP-IR	0,977(0,003)	0,963(0,003)	0,957(0,003)	0,925(0,003)	0,962(0,004)
Chemistry	0,963(0,006)	0,929(0,005)	0,905(0,004)	0,844(0,004)	0,919(0,012)
CS	0,905(0,009)	0,848(0,009)	0,822(0,008)	0,742(0,005)	0,826(0,015)
Physics	0,968(0,006)	0,933(0,005)	0,91(0,006)	0,834(0,006)	0,919(0,016)

Apêndice

B

Tabelas com os resultados da caracterização das redes

Este apêndice contém as tabelas completas das características das redes construídas. A Tabela B.1 contém informações relacionadas ao grau dos vértices, a Tabela B.2 contém informações relacionadas a caminhos nas redes, e a Tabela B.3 contém o coeficiente de agrupamento e a modularidade Q das redes.

Tabela B.1: Grau mínimo, máximo e médio das redes

Conjunto de dados	(grau médio | maior grau | menor grau) **Rede Conj.**		**RH(1)**	**RH(3)**	**RH(5)**	**RH(15)**
Balance		D	2,03|6|1	3,00|17|1	5,00|44|1	14,98|52|1
		P	2,08|6|1	3,00|10|1	4,99|13|1	8,17|26|1
Ecoli		D	2,32|8|1	3,01|11|1	5,00|18|1	14,29|149|1
		P	2,25|7|1	3,00|8|1	4,99|18|1	13,46|149|1
Glass		D	2,41|8|1	2,95|9|1	4,55|16|1	9,24|31|1
		P	2,38|6|1	2,98|9|1	4,51|11|1	10,05|32|1
Ionosphere		D	2,54|20|1	3,03|20|1	4,81|28|1	12,31|74|1
		P	2,49|9|1	2,96|11|1	4,80|21|1	11,69|60|1
Iris		D	2,40|9|1	2,99|8|1	4,60|12|1	7,48|22|1
		P	2,43|7|1	2,93|7|1	4,71|12|1	8,08|19|1
Sonar		D	2,36|8|1	2,90|15|1	4,47|14|1	9,42|46|1
		P	2,33|8|1	2,91|10|1	4,41|14|1	9,45|29|1
Wdbc		D	2,25|10|1	2,95|16|1	4,70|32|1	11,72|85|1
		P	2,17|6|1	2,94|11|1	4,72|17|1	11,28|52|1
Wine		D	2,42|8|1	2,98|12|1	4,81|14|1	8,76|27|1
		P	2,28|6|1	2,93|9|1	4,39|14|1	8,08|21|1
Yeast		D	2,24|10|1	3,00|24|1	5,00|61|1	14,70|287|1
		P	2,07|6|1	3,00|10|1	5,00|42|1	13,95|139|1
Zoo		D	2,97|13|1	3,17|12|1	4,75|12|1	6,71|18|1
		P	2,46|7|1	3,01|7|1	4,71|12|1	7,45|22|1
Books	8,40|25|2	D	2,50|6|1	3,01|7|1	4,59|9|1	5,64|14|1
		P	2,46|6|1	2,95|6|1	4,46|10|1	6,74|17|1
Football	10,66|12|7	D	2,89|8|1	3,27|8|1	4,99|11|1	9,18|17|2
		P	2,61|6|1	3,18|6|1	5,01|11|1	11,41|22|4
Blogs	27,36|351|1	D	2,22|27|1	3,01|29|1	4,95|63|1	14,36|108|1
		P	2,05|7|1	2,99|11|1	4,90|18|1	12,21|86|1
Industry-yh	15,74|250|1	D	2,20|10|1	3,01|36|1	5,00|82|1	15,00|173|1
		P	2,04|7|1	3,01|13|1	5,01|23|1	13,33|189|1
CBR-ILP-IR		D	2,31|9|1	2,99|17|1	4,98|31|1	14,44|60|1
		P	2,09|6|1	3,00|9|1	4,86|17|1	13,07|55|1
Chemistry		D	2,32|9|1	3,02|11|1	4,90|30|1	12,49|58|1
		P	2,16|7|1	2,99|14|1	4,80|24|1	12,17|71|1
CS		D	2,28|9|1	3,01|15|1	4,98|25|1	13,31|82|1
		P	2,10|7|1	3,01|12|1	4,79|19|1	11,51|51|1
Physics		D	2,42|13|1	3,00|13|1	4,89|47|1	12,49|73|1
		P	2,22|9|1	2,99|15|1	4,93|28|1	9,55|55|1

Tabela B.2: Média do menor caminho e diâmetro das redes

Conjunto	(média do menor caminho | diâmetro)					
de dados	**Rede Conj.**		**RH(1)**	**RH(3)**	**RH(5)**	**RH(15)**
Balance		D	11,08|36	7,09|23	5,01|15	3,06|8
		P	8,79|24	6,30|16	4,47|11	3,80|10
Ecoli		D	7,16|17	5,68|15	4,29|10	3,00|7
		P	7,55|24	5,84|15	4,27|11	2,87|6
Glass		D	8,61|29	7,58|23	5,64|18	4,29|13
		P	7,89|24	6,70|21	5,27|17	3,73|11
Ionosphere		D	6,61|15	5,71|14	4,24|9	3,02|9
		P	7,05|20	6,08|14	4,25|9	3,00|9
Iris		D	7,14|17	6,09|14	5,19|12	4,52|10
		P	7,36|17	6,10|15	5,21|13	4,11|10
Sonar		D	6,44|16	5,69|14	4,08|9	3,12|7
		P	6,40|14	5,24|13	4,08|11	3,00|7
Wdbc		D	7,47|18	6,03|14	4,46|12	3,21|8
		P	7,88|19	6,00|16	4,47|10	3,17|7
Wine		D	7,37|19	6,16|17	4,31|11	3,56|10
		P	7,06|19	5,92|15	4,22|13	3,42|9
Yeast		D	8,38|19	6,46|16	4,84|11	3,28|8
		P	9,38|22	6,63|15	4,85|11	3,24|7
Zoo		D	4,32|10	3,98|8	3,29|8	2,89|7
		P	4,77|11	4,10|9	3,26|8	2,72|6
Books	2,76|5	D	5,20|12	4,75|12	3,50|9	3,17|8
		P	5,40|13	4,46|10	3,60|8	2,92|7
Football	2,22|4	D	4,39|11	4,06|9	3,18|6	2,51|4
		P	4,90|12	4,13|9	3,19|6	2,29|4
Blogs	2,83|7	D	9,07|22	6,74|16	5,12|13	3,38|9
		P	9,35|22	6,61|19	4,87|12	3,47|9
Industry-yh	3,41|9	D	9,08|23	6,93|17	5,10|12	3,39|8
		P	9,79|22	6,74|16	4,98|12	3,45|7
CBR-ILP-IR		D	7,15|19	5,72|13	4,22|9	2,76|5
		P	8,62|24	5,80|14	4,27|10	2,83|6
Chemistry		D	8,30|20	5,97|14	4,40|10	3,14|8
		P	8,70|24	6,28|16	4,34|11	2,88|7
CS		D	7,77|21	5,99|15	4,22|10	2,90|7
		P	8,64|26	5,76|16	4,33|9	2,91|6
Physics		D	7,81|21	6,41|18	4,36|10	3,04|7
		P	8,17|22	6,00|16	4,36|11	3,14|7

Tabela B.3: Coeficiente de agrupamento e modularidade Q das redes

Conjunto de dados	(coeficiente de agrupamento | modularidade Q) **Rede Conj.**		**RH(1)**	**RH(3)**	**RH(5)**	**RH(15)**
Balance		D	0,00|0,91	0,13|0,81	0,30|0,71	0,42|0,59
		P	0,01|0,91	0,07|0,78	0,13|0,66	0,36|0,79
Ecoli		D	0,13|0,88	0,27|0,83	0,44|0,78	0,51|0,58
		P	0,04|0,86	0,12|0,80	0,19|0,67	0,38|0,58
Glass		D	0,13|0,83	0,22|0,83	0,43|0,81	0,50|0,68
		P	0,08|0,83	0,12|0,79	0,25|0,75	0,39|0,58
Ionosphere		D	0,11|0,84	0,21|0,83	0,37|0,77	0,50|0,60
		P	0,06|0,84	0,08|0,78	0,25|0,75	0,38|0,58
Iris		D	0,15|0,83	0,28|0,79	0,44|0,75	0,53|0,70
		P	0,11|0,80	0,19|0,79	0,30|0,70	0,43|0,64
Sonar		D	0,06|0,82	0,20|0,80	0,34|0,71	0,44|0,51
		P	0,05|0,82	0,12|0,75	0,21|0,64	0,31|0,49
Wdbc		D	0,05|0,88	0,16|0,83	0,29|0,75	0,35|0,56
		P	0,03|0,88	0,09|0,77	0,15|0,66	0,23|0,51
Wine		D	0,09|0,80	0,20|0,76	0,33|0,70	0,43|0,61
		P	0,02|0,80	0,13|0,75	0,22|0,70	0,36|0,65
Yeast		D	0,09|0,94	0,24|0,88	0,40|0,81	0,46|0,64
		P	0,01|0,93	0,06|0,77	0,10|0,63	0,19|0,54
Zoo		D	0,35|0,79	0,40|0,78	0,51|0,76	0,54|0,67
		P	0,11|0,76	0,24|0,76	0,38|0,70	0,47|0,60
Books	0,49|0,51	D	0,17|0,79	0,34|0,77	0,43|0,69	0,48|0,67
		P	0,12|0,76	0,18|0,74	0,37|0,68	0,46|0,63
Football	0,40|0,60	D	0,20|0,80	0,37|0,80	0,54|0,80	0,64|0,68
		P	0,05|0,80	0,22|0,80	0,44|0,77	0,54|0,60
Blogs	0,32|0,43	D	0,08|0,93	0,16|0,85	0,30|0,77	0,36|0,65
		P	0,01|0,93	0,04|0,77	0,10|0,70	0,20|0,65
Industry-yh	0,17|0,28	D	0,10|0,95	0,21|0,86	0,39|0,74	0,42|0,50
		P	0,00|0,94	0,03|0,71	0,07|0,55	0,20|0,46
CBR-ILP-IR		D	0,10|0,90	0,18|0,84	0,34|0,74	0,33|0,59
		P	0,01|0,89	0,05|0,75	0,10|0,66	0,18|0,53
Chemistry		D	0,11|0,88	0,24|0,82	0,38|0,73	0,43|0,61
		P	0,01|0,87	0,11|0,78	0,22|0,68	0,25|0,55
CS		D	0,11|0,89	0,23|0,82	0,38|0,70	0,38|0,52
		P	0,02|0,88	0,09|0,73	0,22|0,64	0,27|0,47
Physics		D	0,20|0,89	0,31|0,86	0,45|0,74	0,47|0,58
		P	0,05|0,86	0,10|0,76	0,22|0,65	0,36|0,59

Apêndice

C

Tabelas com os resultados dos classificadores propostos

Este apêndice contém as tabelas dos erros dos classificadores propostos. As Tabelas C.1 e C.2 contém, respectivamente, os erros dos classificadores com as redes hierárquicas determinísticas e probabilísticas, para valores de grau médio de entrada igual a 1, 3 e 5. As Tabelas C.3, C.4 e C.5 contém, respectivamente, os erros dos classificadores com as redes k-associados para k igual a 1, 3 e 5. E, por fim, a Tabela C.6 contém os erros dos classificadores paras as redes k-associados ótima.

Tabela C.1: Erros em porcentagem dos classificadores com redes hierárquicas determinísticas, com o grau médio de entrada igual a 1, 3 e 5.

Conjunto de dados	**cbRH (1) erro**	**nBC - RH(1) erro**	**cbRH (3) erro**	**nBC - RH(3) erro**	**cbRH (5) erro**	**nBC - RH(5) erro**
Balance	21,0±5,1	16,5±1,4	20,0±4,5	16,7±0,9	20,3±5,2	15,9±0,9
Ecoli	15,6±4,4	20,0±1,7	14,7±5,0	19,3±1,4	15,2±5,0	19,3±1,2
Glass	10,7±5,2	13,2±2,0	10,9±6,0	15,6±2,7	12,0±6,4	15,1±2,3
Ionosphere	1,3±1,8	1,3±0,4	1,3±1,8	1,2±0,4	1,8±2,4	1,3±0,4
Iris	4,0±4,8	4,8±1,3	3,3±3,8	4,9±1,2	4,7±5,3	5,5±1,5
Sonar	16,3±7,9	17,1±2,1	17,1±7,6	16,9±2,0	19,7±8,4	18,3±2,2
Wdbc	4,2±2,4	4,4±0,7	3,8±1,9	4,4±0,7	4,0±2,3	3,8±0,5
Wine	4,7±5,7	6,5±1,3	4,7±5,7	5,7±1,4	4,8±5,6	5,2±1,3
Yeast	45,8±3,8	49,0±1,0	45,2±4,1	48,8±1,1	45,1±4,5	49,9±2,0
Zoo	3,3±5,4	9,2±1,6	3,3±5,4	10,9±1,8	5,3±6,8	13,7±1,7
Books	25,9±12,9	22,1±1,7	24,0±13,5	21,1±1,7	22,7±9,7	22,7±2,5
Football	92,7±7,9	91,6±2,6	91,8±7,2	91,4±2,7	90,6±7,4	91,9±2,5
Blogs	32,4±3,4	34,1±1,2	32,4±3,5	32,0±1,1	31,2±4,1	31,9±1,4
Industry-yh	79,8±2,6	72,1±0,7	79,3±2,6	71,8±0,4	78,5±2,3	71,9±0,4
CBR-ILP-IR	2,1±1,8	3,6±0,5	2,2±1,8	3,5±0,6	2,3±2,1	2,8±0,5
Chemistry	3,5±2,2	5,3±0,9	4,1±2,6	5,4±1,1	4,1±2,2	5,7±1,0
CS	9,0±4,6	12,7±1,5	9,0±4,2	12,7±1,4	9,2±4,6	13,2±1,3
Physics	3,4±2,8	3,3±0,8	3,6±3,1	3,8±1,0	3,4±2,5	4,5±0,8

Tabela C.2: Erros em porcentagem dos classificadores com redes hierárquicas probabilísticas, com o grau médio de entrada igual a 1, 3 e 5.

Conjunto de dados	cbRH (1) erro	nBC - RH(1) erro	cbRH (3) erro	nBC - RH(3) erro	cbRH (5) erro	nBC - RH(5) erro
Balance	20,0±4,1	19,4±1,7	22,0±5,1	17,8±1,7	19,9±4,6	15,9±1,1
Ecoli	16,2±5,9	20,0±1,6	16,4±5,5	19,3±1,6	14,4±5,3	20,3±1,5
Glass	12,1±6,2	17,4±2,0	12,8±6,6	18,2±2,1	14,1±7,0	18,5±2,3
Ionosphere	1,3±1,9	1,2±0,5	1,4±1,9	1,2±0,6	1,6±2,1	1,6±0,4
Iris	3,6±4,2	6,0±1,4	4,0±4,1	5,3±1,4	5,8±5,7	5,8±1,6
Sonar	18,2±8,2	24,6±3,3	22,9±8,7	24,8±3,1	22,4±10,3	24,7±3,2
Wdbc	4,1±2,5	4,6±0,8	4,3±2,5	4,6±0,6	3,6±2,1	4,1±0,6
Wine	4,7±5,7	6,3±1,4	4,8±5,6	6,0±1,2	4,7±5,7	5,6±1,4
Yeast	46,6±4,0	54,7±1,4	45,6±5,0	54,8±1,6	45,0±3,8	55,7±0,9
Zoo	3,0±5,3	10,5±1,7	4,0±6,2	11,1±2,1	5,6±6,8	13,6±2,1
Books	23,3±13,6	23,6±2,6	23,3±12,7	23,0±2,2	21,7±10,7	21,4±2,2
Football	91,2±7,8	90,5±4,2	91,1±7,4	91,8±2,7	93,0±6,3	91,7±2,7
Blogs	33,1±3,8	34,7±1,5	32,0±3,5	33,6±1,5	30,9±4,2	31,8±1,1
Industry-yh	80,5±2,6	71,9±0,6	80,2±2,2	71,8±0,4	78,6±2,7	71,9±0,3
CBR-ILP-IR	2,1±1,8	6,4±0,9	2,3±1,9	5,0±0,9	2,3±1,9	3,1±0,7
Chemistry	4,0±2,3	8,8±1,6	4,3±3,4	7,5±1,1	4,1±2,3	7,4±1,3
CS	9,4±4,6	19,3±2,0	8,9±4,0	17,4±1,9	8,6±4,7	16,4±1,7
Physics	3,6±3,0	8,2±1,3	3,6±2,8	6,3±1,2	3,6±2,9	4,8±1,2

Tabela C.3: Erros em porcentagem dos classificadores com redes k-associados, com o valor de k igual a 1.

Conjunto de dados	Teste-Rede erro	Rede-Teste erro	Maior Prob. erro	Comitê erro | cob.
Balance	21,8±5,1	10,8±3,6	12,6±4,3	7,2±3,6|0.836
Ecoli	22,5±6,1	48,4±7,9	21,5±6,5	9,2±5,2|0.535
Glass	15,1±7,0	41,7±11,2	14,0±6,9	3,0±5,0|0.570
Ionosphere	1,1±1,6	49,3±9,6	1,3±1,8	0,0±0,0|0.501
Iris	4,9±4,9	35,8±14,6	4,7±4,7	2,0±3,8|0.653
Sonar	15,0±7,8	46,5±9,1	16,6±6,2	4,4±5,9|0.518
Wdbc	4,7±3,1	42,9±7,3	5,2±3,1	3,3±3,4|0.586
Wine	4,8±5,6	42,1±13,8	4,5±5,3	0,0±0,0|0.576
Yeast	59,4±3,7	70,7±3,6	57,0±4,0	36,1±5,9|0.368
Zoo	7,2±7,7	15,2±10,8	6,2±8,4	0,3±1,8|0.837
Books	28,8±15,4	51,7±15,8	28,4±15,6	15,5±18,4|0.500
Football	92,5±7,8	95,9±5,1	95,4±6,3	99,2±4,6|0.145
Blogs	34,8±3,7	59,1±4,4	37,0±4,0	29,1±6,4|0.470
Industry-yh	85,0±2,3	85,4±6,3	82,6±6,1	73,5±5,4|0.203
CBR-ILP-IR	3,0±2,6	41,3±5,9	3,1±2,3	0,4±0,9|0.583
Chemistry	5,7±3,1	40,3±7,0	4,7±2,9	1,9±2,7|0.596
CS	15,6±5,0	45,6±7,5	15,9±5,6	5,1±4,7|0.557
Physics	5,1±4,1	36,3±8,3	5,0±3,4	0,7±1,6|0.630

Tabela C.4: Erros em porcentagem dos classificadores com redes k-associados, com o valor de k igual a 3.

Conjunto de dados	Teste-Rede erro	Rede-Teste erro	Maior Prob. erro	Comitê erro | cob.
Balance	13,6±3,3	11,2±3,6	12,0±3,1	7,1±3,4|0.907
Ecoli	14,2±5,3	22,0±7,4	13,7±5,3	11,1±4,3|0.850
Glass	10,1±5,7	17,3±9,9	9,9±6,7	4,3±5,4|0.831
Ionosphere	1,2±1,8	30,3±8,7	2,3±2,3	0,0±0,0|0.693
Iris	4,7±4,7	9,8±7,8	4,7±4,7	3,3±4,8|0.924
Sonar	17,5±9,5	24,5±8,0	14,4±7,2	7,4±7,3|0.727
Wdbc	3,1±2,1	15,5±4,7	3,7±2,8	2,1±1,7|0.855
Wine	3,5±4,5	12,0±8,8	3,5±4,5	0,9±2,5|0.887
Yeast	43,8±3,7	48,6±4,1	43,4±4,1	37,9±4,9|0.720
Zoo	9,2±9,0	12,9±9,9	8,9±8,8	5,8±7,2|0.910
Books	21,7±10,7	29,5±11,7	21,1±9,9	19,6±10,2|0.856
Football	93,1±8,0	93,9±7,4	92,8±7,2	95,3±8,3|0.498
Blogs	30,8±4,1	42,6±4,9	33,2±3,2	25,3±4,6|0.539
Industry-yh	76,0±2,8	82,1±9,7	75,5±3,6	68,0±7,7|0.276
CBR-ILP-IR	2,5±2,3	13,6±3,9	2,6±2,3	0,2±0,6|0.852
Chemistry	3,8±3,6	16,6±5,4	3,3±3,2	1,5±2,6|0.828
CS	12,0±4,8	23,0±5,6	11,9±4,0	3,9±4,2|0.758
Physics	5,9±3,6	13,2±5,5	2,5±3,2	0,0±0,0|0.827

Tabela C.5: Erros em porcentagem dos classificadores com redes k-associados, com o valor de k igual a 5.

Conjunto de dados	**Teste-Rede erro**	**Rede-Teste erro**	**Maior Prob. erro**	**Comitê erro | cob.**
Balance	11,5±3,5	11,0±4,0	10,4±3,5	7,3±2,9|0.940
Ecoli	14,5±4,9	17,6±6,2	14,5±5,4	12,4±4,4|0.915
Glass	14,2±6,6	17,0±9,5	13,7±7,0	6,4±6,2|0.836
Ionosphere	2,4±2,5	25,1±7,3	3,2±2,7	0,0±0,0|0.737
Iris	4,2±5,1	7,8±6,3	5,3±5,9	3,7±5,3|0.957
Sonar	17,5±8,6	19,2±8,3	14,1±7,4	8,4±7,5|0.782
Wdbc	3,2±2,4	8,6±3,3	3,3±2,7	1,6±1,6|0.915
Wine	4,9±4,6	7,7±6,8	3,4±4,3	2,3±3,0|0.928
Yeast	42,9±4,3	44,5±2,8	42,7±3,9	37,2±3,5|0.762
Zoo	8,6±9,0	9,9±8,7	8,2±8,3	5,4±7,8|0.943
Books	19,5±10,8	21,5±10,5	20,2±11,3	19,4±10,6|0.967
Football	93,3±7,2	93,3±7,7	94,2±6,7	94,3±8,8|0.695
Blogs	31,3±3,4	45,6±5,1	34,2±3,9	25,6±5,1|0.478
Industry-yh	72,5±3,6	72,2±3,0	72,0±3,3	70,0±4,6|0.538
CBR-ILP-IR	1,6±2,1	6,7±3,1	1,5±1,9	0,3±0,7|0.925
Chemistry	4,8±4,2	9,7±5,1	2,2±2,3	1,3±1,9|0.880
CS	10,0±4,7	18,9±6,2	10,1±4,4	4,5±3,5|0.817
Physics	4,9±3,6	6,9±4,3	2,0±3,2	0,4±1,2|0.901

Tabela C.6: Erros em porcentagem dos classificadores com redes k-associados ótima, com o valor de k_{max} igual a 15.

Conjunto de dados	**Teste-Rede erro**	**Rede-Teste erro**	**Maior Prob. erro**	**Comitê erro | cob.**
Balance	13,7±3,5	11,3±3,9	12,3±3,6	7,6±3,9|0.913
Ecoli	16,9±7,5	22,4±9,1	16,8±7,6	12,8±6,7|0.855
Glass	14,3±8,3	41,7±14,0	16,2±8,6	9,3±9,1|0.609
Ionosphere	0,7±1,4	35,6±8,6	8,5±6,9	1,0±2,1|0.650
Iris	4,4±5,6	9,3±8,1	4,9±5,5	3,8±5,4|0.939
Sonar	18,9±9,7	36,6±9,9	20,5±9,9	14,9±11,4|0.648
Wdbc	4,1±2,1	17,8±5,5	3,5±2,2	1,8±1,5|0.810
Wine	3,4±4,0	19,8±11,6	6,1±6,4	2,3±3,6|0.807
Yeast	43,9±4,0	48,3±4,0	43,4±4,3	35,1±5,1|0.624
Zoo	7,6±8,5	13,6±10,4	5,9±7,7	0,7±2,8|0.847
Books	18,3±10,3	22,9±10,6	18,9±11,2	16,9±11,1|0.931
Football	90,5±9,3	93,1±7,4	93,9±6,5	95,3±9,8|0.427
Blogs	31,3±4,9	43,4±5,3	32,8±5,8	25,0±8,4|0.516
Industry-yh	80,9±3,7	75,3±5,8	74,1±4,4	69,8±7,4|0.325
CBR-ILP-IR	1,9±1,9	8,2±3,8	2,0±2,1	1,0±1,3|0.921
Chemistry	6,4±4,8	16,9±7,0	6,6±4,4	2,2±2,8|0.821
CS	10,2±4,8	22,6±7,3	10,8±4,7	5,2±3,9|0.775
Physics	4,3±4,4	17,7±6,4	5,1±4,0	1,3±2,3|0.813