Definição automática da quantidade de
atributos selecionados em tarefas de
agrupamento de dados

José Augusto Andrade Filho

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:

Definição automática da quantidade de atributos selecionados em tarefas de agrupamento de dados

José Augusto Andrade Filho

***Orientador:* Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho**

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

**USP – São Carlos
Novembro de 2013**

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

AAN553 Andrade Filho, José Augusto
d Definição automática da quantidade de atributos
selecionados em tarefas de agrupamento de dados /
José Augusto Andrade Filho; orientador André Carlos
Ponce de Leon Ferreira de Carvalho. -- São Carlos,
2013.
73 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2013.

1. Seleção de atributos. 2. Agrupamento de dados.
3. Definição de quantidade. I. de Carvalho, André
Carlos Ponce de Leon Ferreira, orient. II. Título.

A meus pais

Agradecimentos

Queria agradecer a todas as pessoas que me auxiliaram de alguma maneira neste doutorado.

A meu orientador, prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho, pela dedicação ao me orientar, por todo o apoio e pela abertura que me foi dada para discutir propostas, técnicas e caminhos.

Ao meu supervisor durante o estágio sanduíche, prof. Dr. Huan Liu, por ter me recebido bem e pela ajuda na definição do escopo do trabalho.

A meus colegas e amigos da USP, em especial ao Paulo, Rosane e ao prof. Rodrigo Mello pela importante ajuda e apoio durante a realização deste trabalho e pelos importantes momentos de confraternização.

A meus pais, Vilma e Guga, pelo apoio incondicional em todas as horas, por serem um modelo de conduta.

A minhas irmãs, Mônica e Shirley, pelos momentos divertidos e pela cooperação e apoio em todas as horas.

A minhas avós, tios (em especial a meu padrinho Chico), primos (em especial a Jéssica e Andrezza) e amigos (Stella, Tadeu, Jean, Fernanda e Mike) que tanto me apoiaram nessa jornada.

A CNPq, pelo apoio financeiro durante o doutorado e a CAPES pelo apoio financeiro que permitiu a realização do estágio sanduíche.

*“As far as the laws of mathematics
refer to reality, they are not certain;
and as far as they are certain, they
do not refer to reality.”*

Albert Einstein

Resumo

Conjuntos de dados reais muitas vezes apresentam um grande número de atributos preditivos ou de entrada, o que leva a uma grande quantidade de informação. Entretanto, essa quantidade de informação nem sempre significa uma melhoria em termos de desempenho de técnicas de agrupamento. Além disso, alguns atributos podem estar correlacionados ou adicionar ruído, reduzindo a qualidade do agrupamento de dados. Esse problema motivou o desenvolvimento de técnicas de seleção de atributos, que tentam encontrar um subconjunto com os atributos mais relevantes para agrupar os dados. Neste trabalho, o foco está no problema de seleção de atributos não supervisionados. Esse é um problema difícil, pois não existe informação sobre rótulos das classes. Portanto, não existe um guia para medir a qualidade do subconjunto de atributos. O principal objetivo deste trabalho é definir um método para identificar quanto atributos devem ser selecionados (após ordená-los com base em algum critério). Essa tarefa é realizada por meio da técnica de Falsos Vizinhos Mais Próximos, que tem sua origem na teoria do caos. Resultados experimentais mostram que essa técnica informa um bom número aproximado de atributos a serem selecionados. Quando comparado a outras técnicas, na maioria dos casos analisados, enquanto menos atributos são selecionados, a qualidade da partição dos dados é mantida.

Abstract

Real-world datasets commonly present high dimensional data, what leads to an increased amount of information. However, this does not always imply on an improvement in terms of clustering techniques performance. Furthermore, some features may be correlated or add unexpected noise, reducing the data clustering performance. This problem motivated the development of feature selection techniques, which attempt to find the most relevant subset of features to cluster data. In this work, we focus on the problem of unsupervised feature selection. This is a difficult problem, since there is no class label information. Therefore, there is no guide to measure the quality of the feature subset. The main goal of this work is to define a method to identify the number of features to select (after sorting them based on some criterion). This task is carried out by means of the False Nearest Neighbor, which has its root in the Chaos Theory. Experimental results show that this technique gives an good approximate number of features to select. When compared to other techniques, in most of the analyzed cases, while selecting fewer features, it maintains the quality of the data partition.

Lista de Figuras

2.1	Exemplo de <i>Manifold</i> (Mello, 2009)	6
2.2	Saída da função Logística – Primeiras 100 observações	7
2.3	Observações da função Logística reconstruída (dimensão embutida 2 e de separação 1)	7
2.4	Atrator de Lorenz: Amostra de observações	8
2.5	Atrator de Lorenz reconstruído no espaço de coordenadas de atraso (dimensão embutida 2 e de separação 5) (Mello, 2009)	8
2.6	Atrator de Lorenz reconstruído no espaço de coordenadas de atraso (dimensão embutida 3 e de separação 5) (Mello, 2009)	8
2.7	Atrator de Lorenz – Auto-Mutual Information	10
2.8	Atrator de Lorenz – Encontrado dimensões embutidas	11
2.9	Grupo bem separado.	13
2.10	Grupo baseado em centro.	13
2.11	Grupo contínuo.	14
2.12	Grupo baseado em densidade.	14
2.13	Etapas do processo de agrupamento de dados	14
2.14	Taxonomia de algoritmos de Agrupamento de Dados (Jain <i>et al.</i> , 1999).	16
3.1	Redefinindo um conjunto de dados como uma série temporal ($n = 2$ and $\tau = m = 3$).	35
4.1	Planejamento dos experimentos	46
4.2	Conjunto de dados Iris	48
4.3	Conjunto de dados Breast Tissue	48
4.4	Conjunto de dados Soybean	49
4.5	Conjunto de dados Yeast	49
4.6	Conjunto de dados Ecoli	50
4.7	Conjunto de dados Glass	50
4.8	Conjunto de dados Swissbank	50

4.9	Conjunto de datos Lung cancer	51
4.10	Conjunto de datos WDBC	51
4.11	Conjunto de datos AR10P	56
4.12	Conjunto de datos PIE10P	57
4.13	Conjunto de datos CLL-SUB-111	57
4.14	Conjunto de datos GLI-85	58

Lista de Tabelas

2.1	Atrator de Lorenz – Dados originais	12
2.2	Atrator de Lorenz – Dados reconstruídos de acordo com as dimensões embuti- das e de separação ($m = 3$ and $\tau = 5$)	12
3.1	Conjunto de dados de exemplo	36
3.2	Distâncias – F1	36
3.3	Distâncias – F1F2	37
3.4	Distâncias – F1F2F3	37
3.5	Distâncias – F1F2F3F4	38
3.6	Distâncias – F1F2F3F6	38
3.7	Distâncias – F1F5	38
3.8	Distâncias – F1F6	39
4.1	Principais parâmetros utilizados nos experimentos	46
4.2	Conjuntos de dados do Grupo I	47
4.3	Resultados CR – Base comparativa – Grupo I	52
4.4	Resultados CR – FQFNN e rankings – Grupo I	52
4.5	Resultados Silhueta – Base comparativa – Grupo I	53
4.6	Resultados Silhueta – FQFNN e rankings – Grupo I	53
4.7	Resultado da avaliação do teste de hipótese – CR – Grupo I	54
4.8	Resultado da avaliação do teste de hipótese – Silhueta – Grupo I	54
4.9	Diferença entre a quantidade de atributos selecionados pela técnica FQFNN+SPEC – Grupo I	55
4.10	Similaridade entre os atributos selecionados pela técnica FQFNN+SPEC – Grupo I	56
4.11	Conjuntos de dados do Grupo II	56
4.12	Resultados CR – Base comparativa – Grupo II	58
4.13	Resultados CR – FQFNN e rankings – Grupo II	59
4.14	Resultados Silhueta – Base comparativa – Grupo II	59

4.15	Resultados Silhueta – FQFNN e rankings – Grupo II	60
4.16	Resultado da avaliação do teste de hipótese – CR – Grupo II	60
4.17	Resultado da avaliação do teste de hipótese – Silhueta – Grupo II	61
4.18	Diferença entre a quantidade de atributos selecionados com FQFNN+SPEC – Grupo II	61
4.19	Similaridade entre os atributos selecionados com FQFNN+SPEC – Grupo II . .	62
4.20	Conjuntos de dados do Grupo III	62
4.21	Resultados para o conjunto de dados de espécies endêmicas	63
4.22	Avaliação de Silhueta para o conjunto de dados de espécies endêmicas – 2 grupos	63

Lista de Algoritmos

2.1	Funcionamento básico do <i>K-means</i>	18
2.2	Generalização de uma técnica baseada em Filtro	22
2.3	Generalização de uma técnica baseada em <i>wrapper</i>	23
2.4	Generalização de uma técnica de seleção de atributos híbrida	24
2.5	SPEC – Spectral Feature Selection	28
2.6	Algoritmo <i>k-medoids</i>	30
2.7	Estratégia de amostragem para o <i>k-medoids</i>	31
3.1	Algoritmo da técnica FQFNN	40
3.2	Pseudo-código da técnica SPECMI	42
3.3	Pseudo-código da geração do <i>ensemble</i>	42

Lista de Símbolos

X	Conjunto de exemplos.
π^e	Partição resultante de um algoritmo de agrupamento aplicado a um conjunto de X .
G_w	Grupo w de uma partição π .
x_i	i -ésimo exemplo do conjunto de dados X .
n	Número de exemplos. No contexto desta tese, $n = \tau$
τ	Dimensão de separação (<i>time-delay</i>). No contexto desta tese, $\tau = n$
F	Conjunto de atributos que compõe um exemplo x_i .
p	Número de atributos em um conjunto de dados.
f_j	j -ésimo atributo de um exemplo x_i .
F^*	Subconjunto de F .
m	Dimensão embutida, resultado da técnica de falsos vizinhos.

Sumário

1	Introdução	1
1.1	Contextualização	1
1.2	Objetivo	3
1.3	Estrutura	4
2	Trabalhos relacionados	5
2.1	Considerações Iniciais	5
2.2	Teoria do Caos	5
2.3	Agrupamento de Dados	13
2.3.1	Algoritmos de Agrupamento de Dados	15
2.3.2	Algoritmo <i>K-means</i>	17
2.3.3	Índices de Validação	18
2.4	Seleção de Atributos para Agrupamento de Dados	21
2.5	Considerações Finais	31
3	Técnica FQFNN: <i>Feature Quantification by False Nearest neighbors</i>	33
3.1	Considerações Iniciais	33
3.2	Falsos vizinhos mais próximos aplicado a conjuntos de dados	34
3.3	Análise da interação entre as distâncias	36
3.4	Técnica FQFNN	39
3.5	Técnicas para <i>ranking</i> de atributos	41
3.6	Considerações Finais	41
4	Experimentos e resultados	45
4.1	Considerações Iniciais	45
4.2	Planejamento dos Experimentos	45
4.3	Grupo I: conjuntos de dados da UCI	47
4.4	Grupo II: Conjunto de dados de imagens e <i>microarray</i>	55
4.5	Grupo III: Conjunto de dados de Espécies Endêmicas	59

4.6	Considerações Finais	62
5	Conclusão	65
5.1	Considerações	65
5.2	Contribuições	66
5.3	Publicações	67
5.4	Trabalhos futuros	67
	Referências	69

Introdução

1.1 Contextualização

Conjuntos de dados reais, muitas vezes apresentam um alto número de dimensões, isto é, com um grande número de atributos preditivos ou de entrada (Liu & Motoda, 2007; Jain, 2010). Esses conjuntos podem ser encontrados em dados de diversas áreas, como medicina, biologia, climatologia, física, química etc (Asuncion & Newman, 2007). Dentre esses atributos, é fácil encontrar aqueles que estão correlacionados a outros, que são irrelevantes ou que apresentam ruído. Atributos com esses problemas podem prejudicar, ou não influenciar no desempenho de uma técnica de aprendizado. Além disso, um número elevado de atributos, de acordo com a maldição da dimensionalidade (*curse of dimensionality*), leva à necessidade de um grande número de exemplos (Aggarwal *et al.*, 2001). A seleção de um subconjunto dos atributos que contivesse os mais relevantes, poderia reduzir a ocorrência dos problemas previamente apontados. A investigação e o desenvolvimento de técnicas para selecionar os atributos mais relevantes podem melhorar a qualidade do modelo criado a partir da técnica de aprendizado, bem como reduzir o tempo necessário para induzir tal modelo.

A tarefa de seleção de atributos pode ser formalmente definida do seguinte modo: considere um conjunto de dados X contendo n exemplos, $X = \{x_1, x_2, \dots, x_n\}$. Cada exemplo x_i pode ser descrito como um conjunto $F = \{f_1, f_2, \dots, f_p\}$ de atributos de entrada de tamanho p . A hipótese principal do processo de seleção de atributos é que existe um subconjunto $F^* \subset F$, onde $|F^*| < |F|$, que descreve X , de tal maneira que o modelo induzido por F^* melhora, ou mantém, a qualidade medida quando comparado com o modelo original (que usa F), ao mesmo tempo em que se reduz o tempo de processamento.

A seleção de atributos é geralmente associada com o aprendizado supervisionado (Devaney & Ram, 1997). Nesse caso, os rótulos de classe dos exemplos, os atributos alvo, podem ser

utilizados para determinar a qualidade do modelo induzido com o subconjunto de atributos. Entretanto, no cenário não-supervisionado, não existe informação para guiar a avaliação de qualidade do subconjunto de atributos, o que torna a tarefa de seleção de atributos mais difícil (Devaney & Ram, 1997; Dy & Brodley, 2000b; Dash *et al.*, 2002). Outro problema associado com a seleção de atributos no cenário não-supervisionado é que, mesmo que se conheça, de antemão, quais os atributos relevantes, ainda é difícil definir quantos desses atributos devem ser selecionados.

Diversas abordagens foram propostas na literatura para o problema de seleção de atributos no contexto não-supervisionado (Liu & Motoda, 2007). Elas são classificadas em três categorias: *wrappers*, filtros e híbridos. As abordagens baseadas em *wrappers* utilizam o desempenho de um determinado modelo preditivo para selecionar um subconjunto de atributos relevantes. Para isso, avaliam para um dado subconjunto, qual o efeito desse subconjunto no desempenho preditivo do modelo (Dy & Brodley, 2004; Boutsidis *et al.*, 2009). As abordagens baseadas em filtros, por outro lado, selecionam subconjuntos de atributos utilizando apenas os valores dos atributos de entrada dos objetos no conjunto de dados, procurando, por exemplo, identificar correlações entre atributos de entrada (Zhao & Liu, 2007; Liu & Yu, 2005; Dash *et al.*, 2002). Por fim, as abordagens híbridas incorporam características das duas abordagens anteriores (Covões & Hruschka, 2011).

Como as abordagens baseadas em *wrappers* dependem de um modelo específico, eles são computacionalmente mais custosas do que as baseadas em filtros. Além disso, abordagens baseadas em *wrappers* possuem um viés que tende a privilegiar o modelo utilizado para a seleção de atributos. Por essas razões, abordagens baseadas em filtros são mais utilizadas, especialmente, em problemas que envolvem uma grande quantidade de dados (*big data problems*). Apesar disso, foi empiricamente mostrado que abordagens baseadas em *wrappers*, na maioria dos casos, tendem a gerar modelos melhores para a seleção de atributos do que abordagens baseadas em filtros (Liu & Yu, 2005; Dash *et al.*, 2002).

Para utilizar os aspectos positivos das duas abordagens, foi proposta a abordagem híbrida, que combina características dos métodos baseados em *wrappers* e dos baseados em filtros. Métodos híbridos utilizam métricas independentes e técnicas de indução de modelos para determinar a relevância de um subconjunto de atributos (Covões & Hruschka, 2011).

Outra diferença entre as diversas técnicas utilizadas para seleção de atributos diz respeito à ordem com que os atributos são selecionados. Em uma abordagem gulosa, os atributos são selecionados sequencialmente. O primeiro atributo é o mais relevante. O segundo, é o próximo mais relevante, e assim por diante. Essa relevância pode ser absoluta, considerando todo o conjunto de atributos, ou relativa, mais relevante dados os atributos que já foram selecionados. Um problema dessa abordagem é que ela pode selecionar atributos correlacionados. A segunda abordagem, aqui chamada de combinatorial, procura selecionar de uma única vez o subconjunto de atributos que, juntos, formam o subconjunto mais relevante. O problema dessa abordagem é que, para um número grande de atributos, o número de possíveis subconjuntos pode ser muito

grande, tornando difícil a seleção do subconjunto.

Com base nos trabalhos encontrados na literatura, observou-se que a definição do número de atributos a serem selecionados é feita a partir do resultado da própria técnica, por meio de um limiar (*threshold*) (Dash *et al.*, 2002; Dy & Brodley, 2004), ou pela definição de um percentual de atributos a serem utilizados (Zhao & Liu, 2007). Isso pode tornar a definição do tamanho do subconjunto dependente da técnica de seleção ou um processo subjetivo, pois um limiar ou percentagem bom para um conjunto de dados, não necessariamente é bom para outro.

Para lidar com essa deficiência, é proposto nesta Tese um método para a definição da quantidade de atributos a serem utilizados na tarefa de seleção de atributos no contexto não-supervisionado, que utilize apenas informações do conjunto de dados. Para esse fim, uma técnica baseada em teoria do caos é utilizada. O teorema de imersão de Whitney (1936), que posteriormente foi estendido por Takens (1980), é a base teórica deste trabalho. O teorema de Whitney (1936) mapeia uma série de observações em um espaço de maior dimensão. Posteriormente, Takens (1980) estendeu o teorema de Whitney e formalizou como observações em um espaço de menor dimensão podem ser organizadas para serem mapeadas em um espaço de maior dimensão. Nesse mapeamento, Takens observou que as distâncias entre observações permanecem iguais, mesmo quando o número de dimensões aumenta. Isso acontece porque, após mapear as observações no número correto de dimensões, não há necessidade de aumentá-las. Portanto, nesse espaço de maior dimensão, a informação está melhor representada. Para tanto, Takens (1980) considera os conceitos de dimensão embutida e de separação. O cálculo da dimensão embutida é feito por meio da técnica FNN (*False Nearest Neighbor*), que foi proposta por Kennel *et al.* (1992a), com base no trabalho de Takens (1980).

No contexto desta Tese, é utilizado o mesmo princípio para identificar o número de atributos (dimensões) que melhor representa um conjunto de dados. Ao medir a variação das distâncias entre exemplos, considerando diferentes números de dimensões, é possível identificar quando novas informações não são adicionadas.

Para avaliar o desempenho da técnica proposta, duas medidas são utilizadas. A primeira, é uma medida de avaliação interna, silhueta (Kaufman & Rousseeuw, 1990), que tem como objetivo representar como os exemplos estão dispostos nas partições. A segunda, *Corrected Rand* (CR) (Hubert & Arabie, 1985), é uma medida de avaliação externa que, neste trabalho, compara a partição resultante de um algoritmo de agrupamento com as classes reais (atributos alvo). Uma outra comparação é feita e leva em consideração a quantidade de atributos selecionados pelas técnicas avaliadas.

1.2 Objetivo

O principal objetivo desta Tese é propor e investigar uma técnica para definição automática do número de atributos a serem selecionados, quando os dados não são rotulados. Para isso, é realizado um estudo da interação dos exemplos em um conjunto de dados ao variar a quantidade

de atributos, para, assim, sugerir quantos atributos devem ser utilizados.

Para realizar esse estudo, é desenvolvida uma adaptação da técnica de falsos vizinhos mais próximos (*False Nearest Neighbors* – FNN) (Kennel *et al.*, 1992a), que é baseada nos conceitos definidos no teorema de imersão de Whitney (1936) e Takens (1980).

1.3 Estrutura

O texto desta Tese está estruturado da seguinte forma: no Capítulo 2 são descritos os principais conceitos e definições úteis ao entendimento deste trabalho nos temas de agrupamento de dados e teoria do caos. Esse capítulo também contém uma análise de técnicas de seleção de atributos em problemas de agrupamento de dados. O Capítulo 3 apresenta a definição da proposta, bem como um caso de estudo que exemplifica a utilização da técnica proposta. Os experimentos realizados para a validação da técnica proposta e os resultados obtidos, são apresentados no Capítulo 4. Finalmente, o Capítulo 5 explicita as conclusões obtidas a partir dos resultados e aponta direções para trabalhos futuros.

Trabalhos relacionados

2.1 Considerações Iniciais

Esse capítulo apresenta conceitos que serão úteis na leitura deste trabalho. Inicialmente, é feita uma rápida introdução à teoria do caos, focando no teorema de imersão de Whitney (1936) e, principalmente, no algoritmo de falsos vizinhos mais próximos, proposto por Kennel *et al.* (1992a). Em seguida, conceitos básicos de agrupamentos de dados são expostos. Essa seção tem por objetivo uniformizar os termos de agrupamentos de dados que são utilizados no decorrer desta Tese. Por último, a Seção 2.4 descreve trabalhos relacionados à seleção de atributos em agrupamento de dados, tema desta Tese.

2.2 Teoria do Caos

A Teoria do Caos pode ser utilizada para extrair o comportamento de uma série temporal (Whitney, 1936; Takens, 1980). Se ao invés de uma série temporal, for utilizado um conjunto de dados convencional, estático, é possível utilizar as ferramentas dessa teoria com o intuito de auxiliar a identificação de atributos relevantes em um conjunto de dados. Essa seção apresenta conceitos que são utilizados nesta proposta. Essa seção apresenta alguns conceitos de teoria do caos.

Whitney (1936) utilizou *manifolds*, ou variedades, diferenciáveis como forma de reconstruir funções por meio de transformações para o espaço Euclidiano multidimensional. *Manifolds*, utilizados com frequência na geometria e topologia diferencial, representam um espaço matemático. Matematicamente, é possível dizer que $M \subset \mathbb{R}^k$ é um *manifold* diferenciável de dimensão m se, para cada ponto $p \in M$, existe uma vizinhança $U \subset M$ de p e um homeomorfismo $x : U \rightarrow U_0$, U_0 é um aberto de $\mathbb{R}^m$, como o homeomorfismo inverso $x^{-1} : U_0 \rightarrow U \subset \mathbb{R}^k$

Figura 2.1: Exemplo de *Manifold* (Mello, 2009)

é uma imersão da classe C^{inf} . Ou seja, para cada $u \in U_0$, a derivada $dx^{-1}(u) : \mathbb{R}^m \rightarrow \mathbb{R}^k$ é biunívoca. Portanto, sob essas circunstâncias, diz-se que (x, U) é uma carta local ao redor de p , e U é uma vizinhança coordenada de p (Palis Jr. & Melo, 1978; Mello, 2009).

A Figura 2.1 mostra um exemplo de parametrização de um plano $\mathbb{R}^{m-1} \times \mathbb{R}_+$ para outro $\mathbb{R}^k$. Dado um ponto q' , é possível, utilizando $\phi_q : H_0 \rightarrow H \cap M$, encontrar um ponto correspondente de q em M . O mesmo acontece com p' . Entretanto, nessa situação, obtém-se uma região nos limites de M . Esse exemplo ilustra o mapeamento de um ponto e sua vizinhança em um espaço de maior dimensão (Mello, 2009).

Whitney (1936) observou que esse mapeamento permite o entendimento de comportamentos não-observáveis ou pouco-representativos quando descritos como um espaço de menor dimensão. A partir disso, o autor propôs o teorema de imersão, em que qualquer *manifold* de n -dimensões pode ser mapeado em um espaço com $2n + 1$ -dimensões.

Baseado no teorema de imersão de Whitney (1936), Takens (1980) provou que, ao invés de mapear os estados de um sistema dinâmico em um espaço de $2n + 1$ -dimensões, o sistema pode ser reconstruído considerando atrasos (*time delays*). De acordo com o teorema de imersão de Takens (1980), uma série temporal $x_0, x_1, \dots, x_{n-1}$ pode ser reconstruída em um espaço multi-dimensional $x_n(m, \tau) = (x_n, x_{n+\tau}, \dots, x_{n+(m-1)\tau})$, também chamado espaço de coordenadas de atraso, onde m é a dimensão embutida e τ representa o atraso (ou dimensão de separação). Esse mapeamento, ou técnica de reconstrução, permite transformar observações de um sistema dinâmico (ou regras de saída) em um conjunto de pontos em um espaço Euclidiano de m -dimensões. Essa reconstrução suporta a obtenção de regras para o sistema dinâmico. Como consequência, a reconstrução simplifica o estudo de comportamentos e seus usos sob diferentes circunstâncias, como o estudo de órbitas, tendências e previsões (Alligood *et al.*, 1997; Mello, 2009).

Para melhor entender as dimensões embutida (número de dimensões) e de separação (atraso ou *time delay*), considere a saída da função Logística definida na Equação 2.1, reconstruída em

um espaço multidimensional onde $m = 2$ e $\tau = 1$, o que resulta em pares de pontos (x_t, x_{t+1}) (Figura 2.3). Após a reconstrução, o comportamento da função Logística, que se comportava como caminhada aleatória (Figura 2.2), pode ser estudada, entendida e modelada de um modo mais simples. Após a regressão dos dados, é possível obter a regra do sistema dinâmico e, portanto, entender transições, estimar e prever observações. Com essa regra e um x_t inicial, pode-se, por exemplo, definir a próxima série de observações, x_{t+1} , que serve como entrada para gerar x_{t+2} , e assim por diante (Mello, 2009).

$$x_{t+1} = b \cdot x_t \cdot (1 - x_t) \quad (2.1)$$

Figura 2.2: Saída da função Logística – Primeiras 100 observações

Figura 2.3: Observações da função Logística reconstruída (dimensão embutida 2 e de separação 1)

A dimensão embutida basicamente define o número de eixos para o espaço de coordenadas atraso. Isso determina o número de dimensões necessárias para desdobrar a série reconstruída.

Nesse caso, a série requer duas dimensões, outras situações podem precisar de mais. Esse comportamento é, por exemplo, observado no atrator de Lorenz cujas amostras de observação são apresentadas na Figura 2.4.

Figura 2.4: Atrator de Lorenz: Amostra de observações

Figura 2.5: Atrator de Lorenz reconstruído no espaço de coordenadas de atraso (dimensão embutida 2 e de separação 5) (Mello, 2009)

Figura 2.6: Atrator de Lorenz reconstruído no espaço de coordenadas de atraso (dimensão embutida 3 e de separação 5) (Mello, 2009)

Ao considerar a reconstrução da série utilizando a dimensão embutida igual a 2, obtém-se um espaço de coordenadas de atraso similar ao ilustrado pela Figura 2.5. Entretanto, após adicionar uma nova dimensão e reconstruí-la, logo, com $m = 3$, o comportamento completo da série é desdobrado, o que simplifica seu estudo e compreensão (Figura 2.6). É encerrado o processo de adicionar novas dimensões quando não existe nenhum novo comportamento; nessa situação, o processo é interrompido com $m = 3$, o que representa a melhor dimensão para o atrator de Lorenz (Mello, 2009).

Além da dimensão embutida, ainda existe a de separação, que permite a extração do comportamento periódico da série. Essa dimensão informa o atraso das observações históricas a serem modeladas e analisadas de modo a prever eventos futuros (basicamente permite apontar quão longe deve-se observar para obter um relacionamento de causa-consequência na série). Por exemplo, para prever a temperatura para uma dada região do mundo no dia 12 de dezembro de 2009, pode-se observar o relacionamento ou dinâmica dessa série sobre o tempo. Baseando-se nessa dinâmica, encontra-se o atraso (*time delay*) que auxilia a desdobrar o comportamento mais dissimilar em um sistema dinâmico. Dissimilaridades auxiliam a desdobrar os diferentes estados do sistema que são relacionados por uma regra (ou conjunto de equações) (Mello, 2009).

As dimensões embutida e de separação permitem o estudo das séries. Entretanto, é necessário encontrar essas dimensões para uma série qualquer, incluindo aquelas geradas por dados experimentais.

Fraser & Swinney (1986) estudaram e confirmaram que a técnica *Auto-Mutual Information* (AMI) apresenta melhores resultados ao estimar a dimensão de separação. Para obter a separação da série, aplica-se a AMI sob diferentes atrasos. Posteriormente, plota-se a curva em função dos atrasos (começando de 1 e aumentando) e adota-se o primeiro mínimo como dimensão de separação (Mello, 2009).

A informação mútua média é dada pela Equação 2.2, onde X e Y seguem, respectivamente, as funções de densidade de probabilidade P_X e P_Y , e X e Y acontecem em pares com densidade combinada P_{XY} (Kennel, 2002). Aplicando essa técnica na base de dados do atrator de Lorenz, previamente apresentada, obtém-se a Figura 2.7, de onde pode-se encontrar o primeiro mínimo em 5, confirmando os resultados apresentados em (Lorenz, 1963; Kennel *et al.*, 1992b).

$$I(X; Y) = \int P_{XY}(x, y) \log_2 \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)} dx dy \quad (2.2)$$

Após definir a dimensão de separação, é necessário encontrar a dimensão embutida. Takens (1980) e Mañé (1980) estudaram e confirmaram que o limite superior da dimensão embutida $D_e \in \mathbb{Z}^+$ pode ser estimado utilizando a dimensão fractal D_f , como $D_e > 2.0 \cdot D_f$. Entretanto, a dimensão resultante desse método geralmente é maior que o necessário. Por exemplo, a dimensão fractal do atrator de Lorenz é 2.06 (Medio & Gallo, 1993), consequentemente, o limite superior da dimensão embutida é $D_e > 2.0 \cdot 2.06$, o que resulta em 5. Embora, de

Figura 2.7: Atrator de Lorenz – Auto-Mutual Information

acordo com Kennel *et al.* (1992b), o atrator pode ser desdobrado em $D_e = 3$. Do ponto de vista matemático (Kennel *et al.*, 1992a,b), pode-se modelar esse sistema utilizando 3 ou 5 dimensões, pois, uma vez que todos os possíveis estados são encontrados, pode-se conduzir a análise de comportamento. Porém, ao trabalhar desnecessariamente com dimensões a mais, adiciona-se complexidade e tempo de processamento às etapas de modelagem e análise (Kennel *et al.*, 1992b).

Uma alternativa para obter-se a dimensão embutida mínima é pela computação de sistema invariantes (como o expoente de Lyapunov (Alligood *et al.*, 1997)) para diferentes dimensões, com a observação do resultado de saturação. A complexidade dessa abordagem motivou Kennel *et al.* (1992b) a propor o método de Falsos Vizinhos Mais Próximos (FNN - *False Nearest Neighbor*), que calcula os vizinhos mais próximos para cada ponto, no espaço de coordenadas de atraso (iniciando com dimensão embutida igual a 1). Em seguida, uma nova dimensão é adicionada e as distâncias entre os vizinhos mais próximos é novamente calculada. Quando a distância aumenta, os pontos são considerados falsos vizinhos, o que evidencia a necessidade de mais dimensões para desdobrar o comportamento da série (Mello, 2009).

Kennel *et al.* (1992b) consideram a dimensão embutida d onde o r -ésimo vizinho mais próximo de $y(n)$ é dada por $y^{(r)}(n)$. A distância Euclidiana entre os pontos $y(n)$ e seu r -ésimo vizinho mais próximo é dada pela Equação 2.3. Ao adicionar um nova dimensão, reconstrói-se a série em um espaço $d + 1$ e são adicionadas as coordenadas $(d + 1)$ em cada vetor $y(n)$, que é incluído na equação de distância Euclidiana (componente $x(n + dT)$ da Equação 2.4). Desse modo, o critério de medida da variação de distância após adicionar uma nova dimensão é descrito na Equação 2.5.

$$R_d^2(n, r) = \sum_{k=0}^{d-1} (x(n + kT) - x^{(r)}(n + kT))^2 \quad (2.3)$$

$$R_{d+1}^2(n, r) = R_d^2(n, r) + (x(n + dT) - x^{(r)}(n + dT))^2 \quad (2.4)$$

$$V_{n,r} = \sqrt{\frac{R_{d+1}^2(n, r) - R_d^2(n, r)}{R_d^2(n, r)}} = \frac{|x(n + dT) - x^{(r)}(n + dT)|}{R_d(n, r)} \quad (2.5)$$

De acordo com os autores, se $V_{n,r} > R_{tol}$, os pontos são considerados falsos vizinhos, onde R_{tol} é um limite.

Figura 2.8: Atrator de Lorenz – Encontrado dimensões embutidas

Aplicando o método FNN nos dados do atrator de Lorenz (utilizando dimensão de separação 5, previamente obtida), obtém-se o resultado apresentado na Figura 2.8. Essa figura apresenta a fração de falsos vizinhos por diferentes dimensões embutidas. Quando essa fração é igual a zero, encontra-se a melhor dimensão embutida. Nesse caso, a dimensão embutida é 3, o que confirma o resultado apresentado por Kennel *et al.* (1992b).

Após definir ambas as dimensões, aplica-se o teorema de Takens (1980), como previamente apresentado, onde a série temporal $x_0, x_1, \dots, x_{n-1}$ é reconstruída em um espaço multidimensional, ou espaço de coordenadas de atraso, $x_n(m, \tau) = (x_n, x_{n+\tau}, \dots, x_{n+(m-1)\tau})$ (o componente m representa a dimensão embutida, isto é, o número de dimensões para desdobrar a série, e τ é a de separação, ou seja, o atraso para considerar observações históricas). A reconstrução desdobra completamente o sistema dinâmico que permite a obtenção da regra. Para exemplificar esse desdobramento, considere os dados do atrator de Lorenz apresentado na Tabela 2.1. Após reconstruir os dados com dimensão embutida 3 e separação 5, obtém-se a Tabela 2.2 (a curva resultante dessa reconstrução é apresentada na Figura 2.6) ¹ (Mello, 2009).

Essa reconstrução permite desdobrar o comportamento da série e obter sua regra de comportamento, ou seja, o conjunto de equações que definem a órbita sobre o tempo. Após a obtenção

¹Essa tabela mostra mais observações que a Tabela 2.1. Isso é apenas um exemplo de desdobramento para obter a regra, ou função, o que gera esse sistema dinâmico.

Tabela 2.1: Atrator de Lorenz – Dados originais

Dimensão 1
–9.6559617
–6.9902085
–4.9834927
–3.5773619
–2.6589215
–2.1120568
–1.8411753
–1.7784935
–1.8834828
–2.1397586
–2.5521791
–3.1453527
–3.9638112
–5.0733551
–6.5619076
–8.5356685
–11.100864
–14.311700
–18.056232
–21.873802
–24.819411

Tabela 2.2: Atrator de Lorenz – Dados reconstruídos de acordo com as dimensões embutidas e de separação ($m = 3$ and $\tau = 5$)

Dimensão 1	Dimensão 2	Dimensão 3
–9.655962	–2.112057	–2.552179
–6.990209	–1.841175	–3.145353
–4.983493	–1.778494	–3.963811
–3.577362	–1.883483	–5.073355
–2.658921	–2.139759	–6.561908
–2.112057	–2.552179	–8.535668
–1.841175	–3.145353	–11.100864
–1.778494	–3.963811	–14.311700
–1.883483	–5.073355	–18.056232
–2.139759	–6.561908	–21.873802
–2.552179	–8.535668	–24.819410

desse esse conjunto, é possível estudar seus pontos fixos estáveis e instáveis, compreender suas tendências e também prever observações futuras.

2.3 Agrupamento de Dados

O agrupamento de dados é considerado o problema mais importante no aprendizado não-supervisionado (Kononenko & Kukar, 2007). Seu objetivo é encontrar alguma estrutura em uma base de dados, sem conhecimento prévio (Jain & Dubes, 1988; Mitchell, 1997; Kononenko & Kukar, 2007). Nessa estrutura, os objetos pertencentes a cada grupo (*cluster*) compartilham alguma característica ou propriedade relevante para o domínio do problema em estudo (Jain & Dubes, 1988; Faceli *et al.*, 2011). A definição do que é um grupo é intuitiva, assim, não existe uma formalização única e precisa para esse conceito. Ao contrário, existe uma grande variedade de definições na literatura e isso é resultado das diferentes visões/objetivos dos pesquisadores que trabalham com técnicas de agrupamento. Dentre essas definições, podem ser encontradas (Barbara, 2000; Faceli *et al.*, 2011):

- Grupo bem separado: é um conjunto de pontos, onde qualquer ponto em um grupo A está mais próximo (ou é mais similar) a cada outro ponto nesse grupo A , do que a qualquer ponto que não pertença a A (Figura 2.9).

Figura 2.9: Grupo bem separado.

- Grupo baseado em centro: é um conjunto de pontos onde qualquer ponto em um dado grupo A está mais próximo (ou é mais similar) ao centro do grupo A do que ao centro de qualquer outro grupo. O centro de um grupo pode ser representado como a média aritmética dos pontos desse grupo, ou pelo ponto mais representativo (Figura 2.10).

Figura 2.10: Grupo baseado em centro.

- Grupo contínuo: é um conjunto de pontos onde qualquer ponto em um grupo A está mais próximo (ou é mais similar) a um dos pontos nesse grupo do que a qualquer ponto não pertencente a A (Figura 2.11).
- Grupo baseado em densidade: é uma região densa de pontos, separada de outras regiões de alta densidade por regiões de baixa densidade (Figura 2.12).

Figura 2.11: Grupo contínuo.

Figura 2.12: Grupo baseado em densidade.

A partir das definições de grupo é possível definir um critério de agrupamento, isto é, de qual maneira selecionar a estrutura dos grupos que seja adequada a um conjunto de dados. Os algoritmos de agrupamento são baseados em um critério para agrupar os dados e utilizam medidas de proximidade juntamente com um método de busca para encontrar uma estrutura ótima ou subótima, de modo que, dado um critério de agrupamento, seja possível descrever o conjunto de dados (Jiang *et al.*, 2004).

O processo de agrupamento de dados compreende várias etapas. Essas etapas estão ilustradas na Figura 2.13, que é baseada nas informações apresentadas por Jain *et al.* (1999), Barbara (2000) e Faceli *et al.* (2011).

Figura 2.13: Etapas do processo de agrupamento de dados

Preparação

A representação de um objeto (físico ou noção abstrata) é comumente chamada de padrão, exemplo, amostra, instância ou ponto. A preparação dos dados para o agrupamento

envolve vários aspectos relacionados ao seu pré-processamento e à forma de representação apropriada para a utilização em um algoritmo de agrupamento (Faceli *et al.*, 2011). Neste trabalho, o termo exemplo é utilizado.

Durante o pré-processamento, normalizações, conversões de tipos e reduções do número de atributos por meio de seleção ou extração de características podem ser utilizadas (Jain *et al.*, 1999).

Para a representação, os objetos a serem agrupados são, geralmente, representados por uma matriz de objetos $X_{n \times p} = \{x_1, x_2, \dots, x_n\}$, onde $x_i = \{f_1, f_2, \dots, f_p\}$, n é o número de exemplos e p é o número de atributos que representam os exemplo, ou seja, a dimensionalidade dos objetos.

Proximidade

Consiste na definição de uma medida de proximidade apropriada ao domínio da aplicação. Essa medida pode ser de similaridade ou de dissimilaridade entre dois objetos. As medidas de proximidade, em geral, consideram todos os atributos igualmente importantes.

Jain & Dubes (1988) descrevem as medidas de proximidade mais adequadas para cada tipo de escala de atributo possível. Uma das medidas de proximidade mais comum é a distância Euclideana.

Agrupamento

Essa etapa representa a aplicação de um determinado algoritmo de agrupamento de dados. Uma taxonomia, baseada em Jain *et al.* (1999) é apresentada na seção 2.3.1. O algoritmo *k-means*, utilizado nos experimentos, é detalhado na seção 2.3.2.

Validação

Nessa etapa, o resultado de um algoritmo de agrupamento (etapa Agrupamento) é avaliado. Essa avaliação, de forma objetiva, determina se os agrupamentos são significativos, isto é, se a solução é representativa para o conjunto de dados. A seção 2.3.3 apresenta os índices utilizados neste trabalho.

Interpretação

O objetivo dessa etapa é examinar cada agrupamento formado com o objetivo de rotulá-lo, descrevendo a natureza desse agrupamento. Em geral, essa etapa é realizada por um especialista, que pode ter interesse em encontrar diferenças semânticas de acordo com os objetos e os valores em cada agrupamento (Faceli *et al.*, 2011).

2.3.1 Algoritmos de Agrupamento de Dados

Jain *et al.* (1999) definem uma taxonomia dos algoritmos de agrupamento de dados. Essa taxonomia é apresentada na Figura 2.14 e descrita a seguir.

Figura 2.14: Taxonomia de algoritmos de Agrupamento de Dados (Jain *et al.*, 1999).

Hierárquico Um algoritmo de agrupamento hierárquico tem como resultado um dendrograma. Esse dendrograma representa o agrupamento dos exemplos e os níveis de similaridade em que o agrupamento é modificado. Um dendrograma pode ser considerado em diversos níveis, cada nível representa os possíveis agrupamentos. Em geral, os métodos hierárquicos são variantes dos algoritmos Ligação simples e Ligação completa (Jain *et al.*, 1999). Em ambos os casos, dois grupos são fundidos para formar um grupo maior, baseando-se no critério de distância mínima.

- Ligação-simples: A distância entre dois grupos é a menor das distâncias entre todos os pares de exemplos desses grupos. Essa técnica tem uma tendência de gerar grupos alongados.
- Ligação-completa: Diferente do Ligação-simples, a distância entre dois grupos é a máxima entre pares de exemplos desses dois grupos. Essa técnica, em geral, fornece agrupamentos compactos.

Particional Um algoritmo de agrupamento particional tem como resultado uma partição dos dados. Técnicas particionais geralmente produzem grupos por meio da otimização de um critério, que pode ser definido localmente (em um subconjunto de exemplos) ou globalmente (considerando todos os exemplos). De modo geral, esse tipo de algoritmo é executado diversas vezes para diferentes estados iniciais. A melhor configuração é utilizada como saída desse algoritmo de agrupamento. Dentre os critérios utilizados em algoritmos particionais, podem ser citados (Jain *et al.*, 1999):

- Erro quadrático: o mais intuitivo e frequente critério é o erro quadrático, que tende a ter bons resultados com agrupamentos isolados e compactos.
- Teoria dos grafos: utiliza como critério o cálculo da *Minimal Spanning Tree (MST)* dos dados e na exclusão das arestas da MST com as maiores dimensões, para identificar diferentes grupos.
- Modelo de Mistura e *Mode Seeking*: As abordagens Modelo de Misturas e *Mode Seeking* podem ser utilizadas de diversas maneiras. A premissa básica é que os

exemplos a serem agrupados são escolhidos a partir de uma dentre várias distribuições de probabilidade. O objetivo é identificar os parâmetros de cada distribuição e seus valores. A maioria dos trabalhos supõe que os componentes individuais da mistura são Gaussianas.

Além das técnicas presentes na taxonomia, outros aspectos precisam ser levados em consideração pois, podem ser utilizados em diversos ramos da taxonomia (Jain *et al.*, 1999). São eles:

- **Aglomerativo vs. Divisivo:** Um algoritmo aglomerativo inicia com cada exemplo em um agrupamento distinto e, sucessivamente, combina agrupamentos até que um critério de parada seja atingido. Um método divisivo inicia com todos os exemplos em um único agrupamento e realiza divisões até atingir um critério de parada.
- **Hard vs. Fuzzy:** Em um método de agrupamento *Hard*, os exemplos pertencem a apenas um agrupamento. Em um método *Fuzzy*, são definidos graus de pertinência de cada exemplo para todos os agrupamentos. Um método *Fuzzy* pode ser convertido em um método *Hard*, ao atribuir cada padrão ao agrupamento com maior grau de pertinência.
- **Determinístico vs. Estocástico:** Esse aspecto é mais relevante para abordagens particionais que otimizam uma função de erro quadrático. Essa otimização pode ser alcançada utilizando técnicas tradicionais, ou por meio de uma busca aleatória do espaço de estados, que compreende todas as possibilidades de rotulação.

Para mais informações sobre o processo de agrupamento de dados, sugere-se a leitura dos materiais publicados em Jain & Dubes (1988) e em Faceli *et al.* (2011).

2.3.2 Algoritmo *K-means*

O algoritmo *K-means* é um algoritmo particional que requer a definição do número de agrupamentos k como entrada. Esse algoritmo utiliza uma técnica de realocação iterativa que encontra um ótimo local. Devido a sua simplicidade e variadas implementações é um algoritmo comumente utilizado, apesar da limitação de definir o número de grupos k previamente (Simovici, 2007).

O *K-means* tem como objetivo minimizar o erro quadrático, definido na Equação 2.6, onde μ_j é o vetor centróide do grupo G_j e $d(x_i, \mu_j)$ é a distância Euclideana entre x_i e μ_j , ou seja, o critério seguido pelo *K-means* é minimizar a distância entre cada ponto e o centróide do agrupamento ao qual o ponto pertence.

$$E = \sum_{j=1}^k \sum_{x_i \in G_w} d(x_i, \mu_j) \quad (2.6)$$

Para realizar a tarefa de agrupamento de dados, o algoritmo *K-means*, que tem como parâmetros o conjunto de dados X e número de grupos k e seu funcionamento básico é mostrado no Algoritmo 2.1. O critério utilizado pelo *K-means* é minimizar o erro quadrático entre os exemplos e o centroide do grupo ao qual o exemplo pertence. Portanto, o *K-means* tem como resultado uma partição formada por grupos de formato hipersférico de mesmo tamanho, ou por grupos bem separados.

Algoritmo 2.1: Funcionamento básico do *K-means*

Entrada: Conjunto de dados X

Entrada: Número de grupos k

Saída: Partição do conjunto X em k grupos

```

1  inicio
2      Definir aleatoriamente  $k$  centroides de grupos;
3      repita
4          para cada exemplo  $x_i \in X$  e cada grupo  $G_w, w = 1, \dots, k$  fazer
5              Calcular a distância entre  $x_i$  e o centroide do grupo  $G_w$ ;
6          fin
7          para cada exemplo  $x_i$  fazer
8              Associar  $x_i$  ao grupo com centroide mais próximo;
9          fin
10         para cada grupo  $G_w$  fazer
11             Recalcular o centroide;
12         fin
13     até que nenhuma alteração nas associações de exemplos a grupos seja realizada;
14     retorna Partição formada pelo  $G_k$  grupos
15 fin

```

Os critérios de parada do *K-means* mais comumente utilizados são: execução de um determinado número de iterações, e estabilização dos centroides. Esses critérios de parada podem ser utilizados em conjunto ou individualmente. Vale ressaltar que o resultado desse algoritmo depende de como os centroides são definidos inicialmente.

A complexidade do algoritmo *K-means* é $O(knp)$, onde p é o número de atributos do conjunto de dados X (Simovici, 2007).

2.3.3 Índices de Validação

Para validar uma partição, em geral, utilizam-se índices estatísticos que julgam, de uma maneira quantitativa e objetiva, o mérito das estruturas encontradas. Um índice quantifica a qualidade de uma partição, com base em alguma informação. Um critério de validação define como um determinado índice é aplicado para validar uma partição. Portanto, um critério de validação é uma estratégia para validar uma estrutura de agrupamento e um índice é uma estatística pela qual a validade é testada (Jain & Dubes, 1988).

Jain & Dubes (1988) definiram três tipos de critérios para investigar a validade de agrupamentos:

- **Crítérios Relativos:** comparam diversas partições para decidir qual delas é melhor, de acordo com algum critério. Esses critérios podem ser utilizados na comparação de algoritmos de agrupamento ou na definição do melhor valor para um parâmetro. Por exemplo, decidir qual o melhor valor de k dado um conjunto de partições com diferentes valores de k para o *K-means*, ou seja, quantos grupos devem ser considerados.
- **Crítérios Internos:** utilizados para medir a qualidade de partições geradas com base apenas nos dados originais.
- **Crítérios Externos:** avaliam uma partição com base em uma estrutura pré-identificada, que reflete o conhecimento prévio do pesquisador sobre a organização dos dados.

Neste trabalho, para a seleção de atributos é avaliado o índice interno Silhueta, bem como a correspondência da partição gerada com a partição real, por meio do índice externo Rand Corrigido (CR – *Corrected Rand*), ambos descritos a seguir.

Índice Silhueta

O índice de validação Silhueta é calculado para cada exemplo em um grupo. Esse índice avalia a qualidade das partições baseando-se na proximidade entre padrões de um grupo e na distância dos padrões de um grupo em relação ao grupo mais próximo (Kaufman & Rousseeuw, 1990).

Silhuetas identificam quais padrões estão bem situados em seus grupos e quais estão fora de um grupo apropriado. Podem ser utilizadas medidas de similaridades quanto de dissimilaridades (distâncias) (Faceli *et al.*, 2011). Dados $a(x_i)$, a dissimilaridade média do padrão x_i em relação a todos os outros padrões do grupo G_i , $d(x_i, G_j)$, a dissimilaridade média do padrão x_i em relação aos padrões do agrupamento C_j e $b(x_i)$, a menor dissimilaridade média de x_i em relação a todos os demais agrupamentos, dada pela Equação 2.7, a silhueta de um padrão $s(x_i)$, empregando dissimilaridade, é dada pela Equação 2.8 (Rousseeuw, 1987; Kaufman & Rousseeuw, 1990).

$$b(x_i) = \min_{C_j \neq C_i} d(x_i, C_j) \quad (2.7)$$

$$s(x_i) = \begin{cases} 1 - a(x_i)/b(x_i), & a(x_i) < b(x_i) \\ 0, & a(x_i) = b(x_i) \\ b(x_i)/a(x_i) - 1, & a(x_i) > b(x_i) \end{cases} \quad (2.8)$$

Para a utilização de similaridades, utilizam-se $a'(x_i)$ e $d'(x_i, C_j)$, as respectivas similaridades médias, $b'(x_i)$, dada pela Equação 2.9, e $s(x_i)$, dada pela Equação 2.10.

$$b'(x_i) = \min_{C_i \neq C_j} d'(x_i, C_j) \quad (2.9)$$

$$s(x_i) = \begin{cases} 1 - b'(x_i)/a'(x_i), & a'(x_i) > b'(x_i) \\ 0, & a'(x_i) = b'(x_i) \\ a'(x_i)/b'(x_i) - 1, & a'(x_i) < b'(x_i) \end{cases} \quad (2.10)$$

Um exemplo bem situado dentro do grupo tem uma silhueta com valor próximo de 1 e um valor -1 indica que um exemplo deveria ser associado a outro grupo.

A silhueta depende apenas da partição dos dados, não dependendo do algoritmo de agrupamento utilizado. Esse índice de validação é apropriado nos casos em que a proximidade está em uma escala de proporção, como a distância Euclideana, e para a identificação de grupos compactos e bem separados (Rousseeuw, 1987).

Nesta Tese, foi utilizada a implementação de Silhueta disponível no pacote *fpc* do *software* R, que é uma implementação do índice definido em Rousseeuw (1987) e Kaufman & Rousseeuw (1990).

Índice Rand Corrigido – CR

O Rand Corrigido (CR – *Corrected Rand*) (Hubert & Arabie, 1985) é um índice de validação externa que não é sensível ao número de agrupamentos (Jain & Dubes, 1988). O CR é um dos índices de validação externa mais usado em avaliações e comparações de algoritmos de agrupamento (Faceli, 2006).

Esse índice determina a similaridade entre duas partições π^a e π^b pela concordância, positiva ou negativa, na associação de pares de exemplos aos grupos, ou seja, o índice penaliza associações diferentes de pares de exemplos nas duas partições. De maneira geral, se dois exemplos x_1 e x_2 são associados ao mesmo grupo em π^a e a grupos distintos em π^b , o valor do índice é diminuído.

O índice CR é uma correção estatística do índice Rand e é dado pela Equação 2.11 (Hubert & Arabie, 1985), onde g_{ij} é o número de exemplos comuns aos grupos G_i de π^e e G_j de π^r , g_i é o número de exemplos no grupo G_i de π^e , g_j é o número de exemplos no grupo G_j de π^r e

k^e e k^r representam a quantidade de grupos nas partições π^e e π^r , respectivamente.

$$CR(\pi^e \pi^r) = \frac{\sum_{i=1}^{k^e} \sum_{j=1}^{k^r} \binom{g_{ij}}{2} - \left[\sum_{i=1}^{k^e} \binom{g_{i.}}{2} \sum_{j=1}^{k^r} \binom{g_{.j}}{2} \right] / \binom{n}{2}}{\left[\sum_{i=1}^{k^e} \binom{g_{i.}}{2} + \sum_{j=1}^{k^r} \binom{g_{.j}}{2} \right] / 2 - \left[\sum_{i=1}^{k^e} \binom{g_{i.}}{2} \sum_{j=1}^{k^r} \binom{g_{.j}}{2} \right] / \binom{n}{2}} \quad (2.11)$$

O índice Rand corrigido varia no intervalo $[-1, 1]$, apresentando valores próximos ou menores que 0 quando a semelhança se deve ao acaso e valor 1 quando as partições são idênticas (Hubert & Arabie, 1985). Esse índice apresenta um valor mínimo de -1 , que não é atingido. Em geral, partições muito diferentes resultam em um valor próximo de 0. Como valores menores que 0 não têm aplicação prática (Hubert & Arabie, 1985), a normalização para o intervalo apresenta vantagens. Nesta Tese, a implementação do índice CR utilizada está presente no pacote *fpc* do *software* R.

2.4 Seleção de Atributos para Agrupamento de Dados

Técnicas de seleção de atributos estão geralmente relacionadas ao aprendizado supervisionado (Devaney & Ram, 1997; Dash *et al.*, 2002), pois, a informação da classe, em geral, é utilizada para guiar a seleção dos atributos relevantes (Dash *et al.*, 2002). Entretanto, em problemas de agrupamento de dados, a informação da classe não está disponível, o que dificulta a seleção de atributos relevantes.

Em geral, as técnicas de seleção de atributos em agrupamentos de dados são classificadas em abordagens baseadas em *wrapper*, baseadas em filtros e técnicas híbridas. Técnicas baseadas na abordagem *wrapper* utilizam o resultado de algoritmos de agrupamento para identificar quais atributos são importantes. As técnicas baseadas em filtros utilizam informações do conjunto de dados para calcular a importância de um determinado atributo. As técnicas híbridas utilizam tanto informações do conjunto de dados, quanto o resultado de algoritmos de agrupamento (Dash & Liu, 2000; Liu & Yu, 2005; Covões & Hruschka, 2011).

Uma generalização da abordagem baseada em filtros é apresentada no Algoritmo 2.2. Nessa abordagem, para um dado conjunto de dados X , o algoritmo inicia a busca a partir de um subconjunto $F_{inicial}$ (que pode ser um conjunto vazio, o conjunto completo ou um conjunto aleatório de atributos) e faz uma varredura pelo espaço de atributos utilizando uma dada estratégia de busca. Cada subconjunto de atributos F^* é avaliado por uma medida independente M e, em seguida, comparado com o melhor valor até então. Se melhor, F^* passa a ser considerado o melhor subconjunto de atributos. A busca é finalizada quando um determinado critério de parada δ é atingido.

Ao variar a estratégia de busca e a medida de avaliação, nas etapas *gerar*(X) (linha 5) e

Algoritmo 2.2: Generalização de uma técnica baseada em Filtro

Entrada: Conjunto de dados de treinamento com p atributos: $X = f_0, f_1, \dots, f_p$

Entrada: Subconjunto de busca inicial: $F_{inicial}$

Entrada: Critério de parada: δ

Saída: Subconjunto de atributos ótimo: F_{melhor}

```
1 inicio
2    $F_{melhor} \leftarrow F_{inicial};$ 
   /* avalia  $F_{inicial}$  por meio de uma medida independente  $M$  */
3    $\gamma_{melhor} \leftarrow avalia(F_{inicial}, X, M);$ 
4   repita
   /* Gerar um subconjunto de atributos para avaliação */
5    $F^* = gerar(X);$ 
   /* avalia o subconjunto  $F^*$  por meio de  $M$  */
6    $\gamma \leftarrow avalia(F^*, X, M);$ 
7   se  $\gamma > \gamma_{melhor}$  então
8      $\gamma_{melhor} \leftarrow \gamma;$ 
9      $F_{melhor} \leftarrow F^*;$ 
10  fim
11  até  $\delta$  ser atingido;
12  retorna  $F_{melhor}$ 
13 fin
```

$avalia(F^*, X, M)$ (linhas 3 e 6), diferentes técnicas baseadas em filtros, podem ser desenvolvidas. A abordagem baseada em filtro utiliza um critério de avaliação que não faz uso de informações sobre a qualidade da partição, encontrada por um algoritmo de agrupamento de dados, para a seleção dos atributos mais relevantes. Como resultado, não possui nenhum *viés* em relação a uma técnica de agrupamento de dados (Liu & Yu, 2005).

A abordagem *wrapper* é exemplificada no Algoritmo 2.3. Essa abordagem é semelhante a abordagem baseada em filtro, exceto que é utilizado um algoritmo de agrupamento de dados pré-definido A , ao invés de uma medida M , independente de um modelo descritivo, como um agrupamento de dados, para a avaliação do subconjunto de atributos.

Nessa abordagem, cada subconjunto de atributos F^* é avaliado por meio da aplicação de um algoritmo de agrupamento de dados e da avaliação de sua influência na qualidade do agrupamento de dados gerado. Ao variar a estratégia de busca (função $gerar(X)$, linha 5) e o algoritmo de agrupamento A (linhas 3 e 6), é possível gerar diferentes conjuntos de atributos utilizando uma abordagem baseada em *wrapper*. Como algoritmos de agrupamentos são utilizados para controlar a seleção dos subconjuntos de atributos, essa abordagem tende a ser computacionalmente mais custosa do que a abordagem baseada em filtro (Liu & Yu, 2005).

Para aproveitar as vantagens das técnicas baseadas nas abordagens *wrapper* e em filtro, utilizou-se uma abordagem híbrida, generalizada no Algoritmo 2.4. Essa abordagem utiliza uma medida independente e um algoritmo de agrupamento de dados para selecionar um sub-

Algoritmo 2.3: Generalização de uma técnica baseada em *wrapper***Entrada:** Conjunto de dados de treinamento com p atributos: $X = f_0, f_1, \dots, f_p$ **Entrada:** Subconjunto de busca inicial: $F_{inicial}$ **Entrada:** Critério de parada: δ **Saída:** Subconjunto de atributos ótimo: F_{melhor}

```

1 inicio
2    $F_{melhor} \leftarrow F_{inicial};$ 
   /* avalia  $F_{inicial}$  por meio de um algoritmo de agrupamento  $A$  */
3    $\gamma_{melhor} \leftarrow \text{avalia}(F_{inicial}, X, A);$ 
4   repita
   /* Gerar um subconjunto de atributos para avaliação */
5    $F^* = \text{gerar}(X);$ 
   /* avalia o subconjunto  $F^*$  por meio de  $A$  */
6    $\gamma \leftarrow \text{avalia}(F^*, X, A);$ 
7   se  $\gamma > \gamma_{melhor}$  então
8      $\gamma_{melhor} \leftarrow \gamma;$ 
9      $F_{melhor} \leftarrow F^*;$ 
10  fim
11  até  $\delta$  ser atingido;
12  retorna  $F_{melhor}$ 
13 fin

```

conjunto de atributos. Como critério de parada, é utilizada a qualidade da partição gerada por um algoritmo de agrupamento de dados (Liu & Yu, 2005).

Na literatura, é possível encontrar técnicas que se enquadram em um desses modelos. Mirkin (1999) propôs uma versão “dividir para conquistar” do algoritmo *K-means*. Além da partição dos dados, o algoritmo também define a contribuição de cada atributo na geração desses agrupamentos, o que enquadra essa técnica como baseada na abordagem *wrapper*. Essa contribuição é utilizada para definir quais atributos devem ser selecionados. Apesar de apresentar bons resultados, seu desempenho foi avaliado apenas experimentos com conjuntos de dados pequenos: Iris (matriz 150×4), Soybean (matriz 47×35) ambas do repositório Asuncion & Newman (2007) e Disorder (matriz 44×17 , descrita em (Mirkin, 1999)). Por ser uma técnica baseada *wrapper*, existe uma dependência do algoritmo de agrupamento utilizado, nesse caso, uma variante do algoritmo de agrupamento *K-means*.

Dash & Liu (2000) propuseram uma técnica baseada em *wrapper* que utiliza a entropia para identificar atributos de ruído e sem importância em um conjunto de dados. A influência de um dado atributo d é calculada a partir da entropia do conjunto de dados sem o atributo d . Dash & Liu (2000) definem a entropia a partir da similaridade, S_{i_1, i_2} (Equação 2.12), de modo que a similaridade S_{i_1, i_2} entre dois objetos X_1 e X_2 é alta se os objetos estão próximos e baixa quando esses objetos estão distantes. A Entropia E_{i_1, i_2} (Equação 2.13) é baixa se S_{i_1, i_2} for alto ou baixo e E_{i_1, i_2} será alta caso contrário. Isso significa que a entropia apresenta valores baixos quando

Algoritmo 2.4: Generalização de uma técnica de seleção de atributos híbrida

Entrada: Conjunto de dados de treinamento com p atributos: $X = f_0, f_1, \dots, f_p$

Entrada: Subconjunto de busca inicial: $F_{inicial}$

Saída: Subconjunto de atributos ótimo: F_{melhor}

```
1 inicio
2    $F_{melhor} \leftarrow F_{inicial}$ ;
   /* Calcula a cardinalidade de  $F_{inicial}$  */
3    $c_0 = \text{card}(F_{inicial})$ ;
   /* avalia  $F_{inicial}$  por meio de uma medida independente  $M$  */
4    $\gamma_{melhor} \leftarrow \text{avalia}(F_{inicial}, X, M)$ ;
   /* avalia  $F_{inicial}$  por meio de um algoritmo de agrupamento  $A$  */
5    $\theta_{melhor} \leftarrow \text{avalia}(F_{inicial}, X, A)$ ;
6   para  $c \leftarrow c_0 + 1$  até  $p$  fazer
7     para  $i \leftarrow 0$  até  $p - c$  fazer
8       /* gera um subconjunto com cardinalidade  $c$  para
          avaliação */
9        $F^* \leftarrow F_{melhor} \cup \{f_i\}$ ;
       /* avalia o subconjunto  $F^*$  por meio da medida  $M$  */
10       $\gamma \leftarrow \text{avalia}(F^*, X, M)$ ;
11      se  $\gamma > \gamma_{melhor}$  então
12         $\gamma_{melhor} \leftarrow \gamma$ ;
13         $F'_{melhor} \leftarrow F^*$ ;
14      fim
15    fin
16    /* avalia  $F'_{melhor}$  por meio do algoritmo  $A$  */
17     $\theta = \text{avalia}(F'_{melhor}, X, A)$ ;
18    se  $\theta > \theta_{melhor}$  então
19       $\theta_{melhor} \leftarrow \theta$ ;
20       $F'_{melhor} \leftarrow F^*$ ;
21    senão
22      /* Interrompe a execução e retorna  $F_{melhor}$  como
         resultado */
23    retorna  $F_{melhor}$ 
24  fim
```

X_1 e X_2 têm alguma relação e elevados quando esses atributos não estão relacionados.

$$S_{i_1, i_2} = e^{\alpha \times D_{i_1, i_2}} \quad (2.12)$$

$$E_{i_1, i_2} = -S_{i_1, i_2} \log S_{i_1, i_2} - (1 - S_{i_1, i_2} \times \log (1 - S_{i_1, i_2})) \quad (2.13)$$

A similaridade (Equação 2.12) é baseada na distância D_{i_1, i_2} e pode ser aplicada tanto a valores numéricos quanto categóricos. Para dados numéricos, Dash & Liu (2000) utilizam a distância Euclideana. para dados categóricos, a distância de *Hamming* é utilizada. O valor de α é calculado, automaticamente, ao atribuir 0.5 na equação $\bar{S} = e^{-\alpha \times \bar{D}}$, o que significa entropia máxima; tem-se $\alpha = \frac{-\ln 0.5}{\bar{D}}$ em que $\bar{D}$ é a média das distâncias entre os exemplos.

Para definir um *ranking* dos atributos, calcula-se a entropia de cada atributo. Se a remoção do atributo A_1 causar mais desordem que a remoção do atributo A_2 então a $E_{A_1} > E_{A_2}$, em que E_{A_1} e E_{A_2} representam a entropia após a remoção dos atributos A_1 e A_2 , respectivamente. A definição de quantos atributos serão selecionados é feita a partir da verificação da qualidade do resultado do algoritmo *K-means*. O critério de parada para definir quantos atributos devem ser utilizados é definido visualmente, com base na curva gerada pela qualidade da partição gerada pelo *K-means*.

Posteriormente, Dash *et al.* (2002) evoluíram o trabalho de Dash & Liu (2000) e propuseram uma solução baseada em filtros para selecionar os atributos mais relevantes. A entropia é utilizada para identificar grupos nos dados. A entropia é, também, utilizada como uma medida para comparar subconjuntos de atributos, pois seu valor independe da cardinalidade do subconjunto. Nesse caso, baixos valores de entropia representam subconjunto de agrupamentos bem formados e altos valores de entropia, o contrário. A principal desvantagem é a falta de consenso na avaliação dos agrupamentos. Por ser uma técnica baseado em filtros, Dash *et al.* (2002) desenvolveram uma técnica que não necessita de informação *a priori* sobre o conjunto de dados.

Outro método de seleção de atributos para dados não rotulados é o FSSEM (*Feature Subset Selection using EM Clustering*) (Dy & Brodley, 2000a,b, 2004) que faz uso do modelo de mistura Gaussiana para identificar subconjuntos de atributos que melhor descobrem agrupamentos naturais nos dados. O FSSEM é uma técnica baseada em *wrapper* (Dy & Brodley, 2000a). Para a busca no espaço de atributos, Dy & Brodley (2000b) fazem uma busca incremental, iniciando com zero atributos e, sequencialmente, adicionando um atributo de cada vez. O atributo adicionado é aquele que provê o maior valor quando combinado com os atributos previamente escolhidos. A busca é interrompida quando, ao adicionar novos atributos, o critério de seleção não melhora. Por ser baseada em *wrapper*, ela utiliza um algoritmo de agrupamento de dados, nesse caso, o *EM clustering*. Dois critérios de seleção são empregados para o subconjunto de atributos utilizados, são eles: *Scatter Separability* e *Maximum Likelihood*.

- *Scatter Separability*: a separação dos agrupamentos é definida pela *Scatter Separability*.

S_w (Equação 2.14) é a matriz de espalhamento *intra-cluster* e S_b a matriz de espalhamento entre agrupamentos.

$$S_w = \sum_{j=1}^k \pi_j E \left[(X - \mu_j) (X - \mu_j)^T | \omega_j \right] \quad (2.14)$$

$$S_b = \sum_{j=1}^k \pi_j (\mu_j - M_0) (\mu_j - M_0)^T \quad (2.15)$$

$$M_0 = E\{X\} = \sum_{j=1}^k \pi_j \mu_j \quad (2.16)$$

em que π_j é a probabilidade de um exemplo pertencer ao grupo ω_j , X é um vetor de atributos representando um exemplo do conjunto de dados, k é o número de grupos, μ_j é a média dos exemplos do grupo ω_j , M_0 é média de todos os exemplos e $E[\cdot]$ é o valor esperado do operador. Dy & Brodley (2004) escolhem o critério $\text{trace}(S_w^1 S_b)$. $S_w^1 S_b$ é S_b normalizado pela covariância média do grupo. Logo, quanto maior for o valor de $\text{trace}(S_w^1 S_b)$, maior é a distância normalizada entre os grupos, o que resulta em uma melhor discriminação dos grupos.

- *Maximum Likelihood*: Ao utilizar o algoritmo de agrupamento de dados EM, Dy & Brodley (2000b) supõem que cada grupo encontrado é uma Gaussiana. Esse critério descreve o quanto o modelo (uma mistura de Gaussianas) se ajusta aos dados. Os melhores grupos, nesse caso, são os agrupamentos naturais, isto é, os grupos que são Gaussianas.

Ao procurar pelo melhor subconjunto de atributos, o número de grupos k , depende do subconjunto de atributos avaliados (Dy & Brodley, 2004; Li *et al.*, 2007). O FSSEM- k (Dy & Brodley, 2004) procura por um valor de k para um subconjunto de atributos utilizando o método de Bouman (Dy & Brodley, 2004). para mesclar grupos e adicionar um termo de penalização usando o critério de informação *Bayesiana*. A penalização é necessária pois o *Maximum Likelihood* aumenta à medida em que o número de grupos aumenta e procura-se evitar o caso trivial em que cada exemplo é um grupo (Dy & Brodley, 2004).

A técnica SPEC (*Spectral Feature Selection*) (Zhao & Liu, 2007, 2012) pode ser utilizada tanto para problemas supervisionados quanto não-supervisionados. Neste trabalho a técnica SPEC é utilizada como uma abordagem baseada em filtro para seleção de atributos em um contexto não-supervisionado. Ao construir uma matriz de similaridade S a partir do conjunto de dados X , os autores afirmam que a dispersão das instâncias pode ser estudada ao analisar o espectro do grafo G , que é induzido a partir da matriz S .

Ao considerar um contexto não-supervisionado, os rótulos de classe não são utilizados para gerar a matriz S e o grafo G . Entretanto, os rótulos de classe apresentam uma semelhança com a estrutura do grafo G . Um atributo consistente com a estrutura do grafo atribui valores

similares a instâncias que estão próximas no grafo G (Zhao & Liu, 2007). Diversas variantes da técnica SPEC podem ser definidas ao adotar diferentes matrizes de similaridade. Para o caso não-supervisionado, a função de *kernel* gaussiano, definida na Equação 2.17 é utilizada por Zhao & Liu (2007). Nessa equação, x_i e x_j representam o i -ésimo e o j -ésimo exemplo do conjunto de dados X , respectivamente. Como S é simétrico, o grafo G não é direcionado.

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \quad (2.17)$$

Em seguida, calcula-se a matriz de graus D , definida na Equação 2.18. Onde d_i é definido na Equação 2.19 e n é a quantidade de exemplos no conjunto de dados X .

$$D_{i,j} = \begin{cases} d_i & \text{if } i = j \\ 0 & \text{caso contrário.} \end{cases} \quad (2.18)$$

$$\sum_{j=1}^n S_{i,j} \quad (2.19)$$

A partir da matriz de graus D e da matriz de similaridade S , os autores obtiveram a matriz Laplaciana L e a matriz Laplaciana normalizada L^* (Equações 2.20 e 2.21, respectivamente).

$$L = D - S \quad (2.20)$$

$$L^* = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \quad (2.21)$$

Com a matriz Laplaciana normalizada L^* , os autores calcularam a decomposição espectral (λ_i, ξ_i) para definir o valor de *ranking* para o atributo f_i (Equação 2.22). Onde $\xi_0 = \mathcal{D}^{\frac{1}{2}}$ e $\alpha_j = \cos\theta_j$, com θ_j sendo o ângulo entre $\hat{f}_i$ e ξ_j ; e $\hat{f}_i$ vetor de atributos ponderados definido por $\frac{D^{\frac{1}{2}} f_i}{\|D^{\frac{1}{2}} f_i\|}$.

$$\varphi(f_i) = \frac{\sum_{j=1}^{n-1} \alpha_j^2 \lambda_j}{\sum_{j=1}^{n-1} \alpha_j^2} = \frac{\hat{f}_i^T (L^*) \hat{f}_i}{1 - \hat{f}_i^T \xi_0}. \quad (2.22)$$

O Algoritmo 2.5 apresenta um pseudo-código para a criação do ranking de atributos utilizando a técnica SPEC. A quantidade de atributos a serem selecionados é definido pelo usuário a partir do resultado do algoritmo.

Variações da técnica SPEC aqui apresentada podem ser encontradas em (Zhao & Liu, 2012). Essas variações consideram diferentes funções de similaridade e de *ranking*.

Mitra *et al.* (2002) propuseram um método de seleção de atributos baseado filtros. Esse método envolve duas etapas, são elas: particionamento do conjunto de atributos em um número de subconjuntos homogêneos (grupos) e a seleção de atributos representativos de cada grupo.

O particionamento dos atributos é baseado no princípio do k_{NN} , utilizando como medida de similaridade o índice de compressão máxima da informação (λ_2), definido na Equação 2.23,

Algoritmo 2.5: SPEC – Spectral Feature Selection

Entrada: Conjunto de dados X **Entrada:** Número de exemplos: n **Saída:** *Ranking* dos atributos: $Rank_F$

```
1 inicio
2   Gerar matriz de similaridade  $S$  a partir do conjunto de dados  $X$ ;
3   Gerar grafo  $G$  a partir de matriz  $S$ ;
4   Gerar a matriz de graus  $D$  a partir da matriz  $G$ ;
5   Definir  $L$  e  $L^*$  de acordo com as Equações 2.20 e 2.21;
6   para cada vetor de atributos  $f_i$  fazer
7        $\hat{f}_i \leftarrow \frac{D^{\frac{1}{2}} f_i}{\|D^{\frac{1}{2}} f_i\|}$ ;
8        $F_i \leftarrow \varphi(f_i)$ ;
9   fin
10   $Rank_F \leftarrow$  ranking ordenado de maneira decrescente dos valores  $F_i$ ;
11  retorna  $Rank_F$ ;
12 fin
```

onde $var(f_i)$ é a variância de f_i e $\rho(f_i, f_j)$, o é coeficiente de correlação entre f_i e f_j (Equação 2.24) (Mitra *et al.*, 2002).

$$2\lambda_2(f_i, f_j) = (var(f_i) + var(f_j)) - \sqrt{\xi^2 - 4var(f_i)var(f_j)(1 - \rho(f_i, f_j)^2)} \quad (2.23)$$

$$\rho(f_i, f_j) = \frac{cov(f_i, f_j)}{\sqrt{var(f_i)var(f_j)}} \quad (2.24)$$

Considerando a similaridade apresentada, o primeiro passo do método é calcular os k vizinhos mais próximos de cada um dos atributos. Dentre esses, o atributo que contém o subconjunto mais compacto (determinado pela distância do vizinho mais distante) é selecionado e a k vizinhança de atributos é descartada. O processo é repetido para os atributos restantes até que todos sejam selecionados ou descartados (Mitra *et al.*, 2002).

Durante a execução do algoritmo, um limiar de erro constante (ϵ) é definido como a distância do k -ésimo vizinho mais próximo do atributo selecionado na primeira iteração. Esse limiar é utilizado para controlar o valor de k no decorrer das iterações, de modo que se o valor de γ_2 for maior do que ϵ o valor de k é decrementado.

O valor inicial do k_{NN} é informado pelo usuário. Segundo Mitra *et al.* (2002), esse parâmetro pode ser útil para controlar a representação dos dados considerando diferentes níveis de detalhamento. Os resultados apresentados em (Mitra *et al.*, 2002) mostram que esse filtro tem um desempenho superior aos métodos *branch and bound* (Devijver & Kittler, 1982), *sequential floating forward search* (Pudil *et al.*, 1994), *sequential forward search* (Devijver & Kittler, 1982) e *stepwise clustering* (King, 1967), tanto em tempo de execução quanto em termos de qualidade.

No entanto, definir um valor para esse parâmetro pode ser uma tarefa difícil de ser realizada na prática (Covões & Hruschka, 2011), pois resta ao usuário estimar um parâmetro crítico do algoritmo. A complexidade desse método é definida como $O(M^2N)$, onde M e N representam a quantidade de atributos e exemplos, respectivamente, para um determinado valor de k_{NN} (Mitra *et al.*, 2002). Caso o valor de k_{NN} seja desconhecido, esse pode variar no intervalo $[k_{NN_{min}}, k_{NN_{max}}]$. Ao realizar análise considerando $k_{NN_{min}} = 1$ e $k_{NN_{max}} = M - 1$, a complexidade do algoritmo pode chegar a $O(M^2N + M^3)$.

Li *et al.* (2006) propuseram uma abordagem híbrida baseada nos trabalhos de Dash & Liu (2000); Dash *et al.* (2002). Essa abordagem ordena os atributos de acordo com a importância para a partição encontrada por um algoritmo de agrupamento de dados. Para isso, é necessário apenas o cálculo de n valores e a ordenação desses valores. O índice de ordenação utiliza a entropia exponencial. Seja $S_{p,q}$ a similaridade entre dois objetos X_p e X_q , e seja $|D|$ a quantidade de exemplos em que o índice de *ranking* é calculado. Esse índice é definido de acordo com a Equação 2.25.

$$H = \sum_{p=1}^{|D|} \sum_{q=1}^{|D|} [S_{p,q} \times e^{(1-S_{p,q})} + (1 - S_{p,q}) \times e^{S_{p,q}}] \quad (2.25)$$

Em que $S_{p,q}$ admite valores entre $[0.0 - 1.0]$. Quando $S_{p,q} \rightarrow 0(1)$, H diminui. Entretanto, $S_{p,q} \rightarrow 0,5$, H aumenta. O modo de calcular o *ranking* é o mesmo utilizado por Dash & Liu (2000). Para identificar quantos atributos são escolhidos, Li *et al.* (2006) utilizam uma abordagem em dois passos, um utiliza uma técnica baseada em filtro e o outro uma técnica baseada em *wrapper*, o que define a abordagem como híbrida. O primeiro passo é encontrar um conjunto de atributos utilizando o *Fuzzy Feature Evaluation Index* (FFEI). Como um mecanismo para remover atributos potencialmente redundantes não foi definido, o segundo passo consiste em refinar o resultado e utilizar uma abordagem mais sofisticada para procurar um conjunto de atributos mais compacto. Isso é feito utilizando o algoritmo de agrupamento de dados *Fuzzy C-Means* (Bezdek, 1981) e avaliando a qualidade do resultado.

Uma outra abordagem híbrida é a técnica *Simplified Silhouette Filter* (SSF) (Covões *et al.*, 2009; Covões & Hruschka, 2011). Essa técnica particiona o conjunto de atributos F em uma coleção $C^X = \{C_1, C_2, \dots, C_y\}$, formada por y subconjuntos disjuntos de atributos correlacionados C_i de F . Posteriormente, são selecionados atributos de cada um desses grupos.

Assim como no agrupamento de exemplos, atributos em um mesmo grupo são mais semelhantes entre si do que atributos pertencentes a grupos distintos. Para calcular a similaridade (correlação) entre os atributos, os autores utilizaram medidas de correlação definidas por Mitra *et al.* (2002), *Maximal Information Compression Index* (Equação 2.23), e por Au *et al.* (2005), *Interdependence Redundancy Measure*.

A técnica SSF utiliza um procedimento heurístico, baseado no critério da Silhueta Simplificada (Hruschka *et al.*, 2004, 2006), para encontrar o número de grupos e a partição de atributos correspondente.

A Silhueta Simplificada baseia-se no cálculo das distâncias entre atributos e o centroide do grupo, especificamente Covões & Hruschka (2011) utilizam *medoides* ao invés de centroides. Para a Silhueta Simplificada (Hruschka *et al.*, 2004, 2006), a dissimilaridade média $d(f_i, G_j)$ de um atributo f_i é calculada considerando-se a distância entre o atributo f_i e o *medoide* de grupo G_j , Equações 2.26 e 2.27, onde $f(c(f_i, f_j))$ é uma medida de dissimilaridade calculada em função da correlação entre atributos $c(f_i, f_j)$ e η_j é o *medoide* do grupo G_j .

$$d(f_i, G_j) = f(c(f_i, \eta_j)) \quad (2.26)$$

$$\eta_j = \arg \min_{f_v \in G_j} \sum_{f_u \in G_j} f(c(f_v, f_u)) \quad (2.27)$$

Covões & Hruschka (2011) utilizam o algoritmo *k-medoids* (Algoritmo 2.6) para obter as partições que serão avaliadas pela Silhueta Simplificada.

Algoritmo 2.6: Algoritmo *k-medoids*

Entrada: Quantidade de grupos k

Saída: Partição de atributos

1 **inicio**

2 Crie uma partição inicial aleatória de atributos com k grupos não-vazios;

3 **repita**

4 Calcule a *medoide* do grupo, utilizando a equação 2.27, da partição atual;

5 Conecte cada atributo ao grupo com a *medoide* mais próxima;

6 **até não haver alteração nas medoides por duas iterações seguidas;**

7 **retorna** Partição de atributos

8 **fin**

A Silhueta Simplificada é um critério numérico que permite estimar automaticamente o número de grupos, o que possibilita a reduzir o impacto da limitação do *k-medoids*, que é definir *a priori* a quantidade k de grupos. A estratégia de amostragem, descrita no Algoritmo 2.7, apresenta o método utilizado para estimar o valor de k^* de acordo com a Silhueta Simplificada, bem como encontrar a partição de atributos correspondente a esse k^* (Covões *et al.*, 2009; Covões & Hruschka, 2011).

De acordo com os resultados apresentados por Covões & Hruschka (2011), os autores sugerem que dois atributos de cada grupo sejam utilizados. Esses atributos são: o que representa a *medoide* do grupo e o atributo de menor correlação com a *medoide*.

Em seu pior caso, a complexidade do filtro SSF é $O(M^3)$, quando se faz necessária uma busca exploratória considerando todas as possíveis quantidades de grupos de atributos, ou seja, para todo $k \in 2, \dots, M - 1$. Entretanto, a complexidade estimada pode ser reduzida caso algum conhecimento de domínio esteja disponível (Covões & Hruschka, 2011).

Algoritmo 2.7: Estratégia de amostragem para o *k-medoids***Entrada:** Quantidade mínima e máxima de grupos, k_{min} e k_{max} respectivamente.**Entrada:** Quantidade de partições iniciais (n_p) para o *k-medoid***Saída:** SSV e seu correspondente grupo de atributos k^*

```

1  inicio
    /* SSV = Valor da Silhueta Simplificada */
2   $SSV \leftarrow -\text{inf};$ 
3  para cada  $k \in \{k_{min}, k_{min} + 1, \dots, k_{max}\}$  hacer
4      Crie aleatoriamente  $n_p$  partições de atributos iniciais em  $k$  grupos não-vazios;
5      para cada Partição  $a$  criada hacer
6          Execute o k-medoids e calcule a sua silhueta simplificada;
7           $MV \leftarrow$  melhor valor obtido dentre todas as partições;
8      fin
9      se  $MV > SSV$  então
10          $SSV \leftarrow MV;$ 
11          $k^* \leftarrow k;$ 
12         Armazene a partição correspondente para  $k^*$ ;
13     fim
14 fin
15 retorna  $SSV$  e a partição de atributos correspondente  $k^*$ 
16 fin

```

2.5 Considerações Finais

Nesta Tese, a Teoria do Caos serviu de base para a definição da técnica proposta, que utiliza uma modificação do algoritmo de falsos vizinhos mais próximos.

Para comparar a técnica proposta nesta Tese, três técnicas foram selecionadas. São elas: MitraFS(Mitra *et al.*, 2002), FSSEM-k(Dy & Brodley, 2004) e SSF(Covões & Hruschka, 2011), que são abordagens baseada em filtro, *wrapper* e híbrida, respectivamente.

O algoritmo *K-means* e os índices de validação Silhueta e CR são utilizados para avaliar e comparar os resultados das técnicas.

Técnica FQFNN: *Feature Quantification by False Nearest neighbors*

3.1 Considerações Iniciais

A seleção de atributos pode ser formalmente definida da seguinte forma: considere um conjunto de dados X contendo n exemplos, $X = \{x_1, x_2, \dots, x_n\}$. Cada exemplo x_i pode ser descrito como um conjunto $F = \{f_1, f_2, \dots, f_p\}$ de atributos de entrada de tamanho p .

A hipótese principal do processo de seleção de atributos é que existe um subconjunto $F^* \subset F$, onde $|F^*| < |F|$, que descreve X , de tal maneira que o modelo induzido por F^* melhora, ou mantém, a qualidade medida quando comparado ao modelo original (que utiliza F), ao mesmo tempo em que se reduz o tempo de processamento.

Em geral, a seleção de atributos está relacionada com problemas supervisionados. Como esse tipo de problema não é foco deste trabalho, para maiores informações sugere-se a leitura de (Zongker & Jain, 1996; Guyon & Elisseeff, 2003; Liu & Motoda, 2007).

No caso de problemas não-supervisionados, a definição da qualidade do modelo induzido por F^* é dada pela aplicação de uma medida de qualidade ao modelo, por exemplo, a Silhueta (Rousseeuw, 1987; Kaufman & Rousseeuw, 1990).

A seleção de atributos não-supervisionada pode ser dividida em duas etapas, que são bem definidas pelas duas perguntas a seguir:

1. Quais atributos são mais importantes?
2. Quantos atributos devem ser utilizados?

Em geral, as técnicas existentes na literatura procuram resolver as duas questões simultaneamente (Covões & Hruschka, 2011; Li *et al.*, 2007; Mitra *et al.*, 2002; Dy & Brodley, 2004), definindo quais atributos serão selecionados ao mesmo tempo em que definem quantos serão utilizados.

A primeira questão está ligada à definição da relevância do atributo, que varia de acordo com a técnica utilizada. A segunda questão, que diz respeito à quantidade de atributos a selecionar, apresenta uma tendência a ser definida pelo usuário. Podendo ser tanto de maneira direta (Dash & Liu, 2000; Zhao & Liu, 2007; Li *et al.*, 2007), por meio da definição de um limiar ou da intervenção direta do usuário, quanto de maneira indireta (Dy & Brodley, 2004; Mitra *et al.*, 2002; Covões & Hruschka, 2011), por meio de um parâmetro existente na técnica.

Na presente Tese é proposta uma técnica de definição automática da quantidade de atributos selecionados.

3.2 Falsos vizinhos mais próximos aplicado a conjuntos de dados

Técnicas baseadas em teoria do caos dão suporte ao estudo do comportamento de séries temporais. Por exemplo, o método de falsos vizinhos mais próximos (FNN, do inglês *False Nearest Neighbors*) é utilizado para determinar a dimensão embutida, isto é, quantas dimensões ou eixos são necessários para desdobrar o comportamento de uma série, permitindo o seu estudo. De modo complementar, a Auto Informação Mútua calcula a dimensão de separação, permitindo a extração das regras geradoras da série temporal.

Embora os teoremas de imersão de Whitney e Takens sejam geralmente aplicados no contexto de sistemas dinâmicos, com o intuito de estudar séries temporais, órbitas e tendências, esses teoremas são ferramentas matemáticas desenvolvidas para embutir *manifolds* m -dimensionais em espaços de maior dimensão. Ao permitir o desdobramento de atributos, é possível que esses teoremas, especificamente o último, possam ser usados como base para a seleção de atributos não-supervisionado. Para tanto, é investigado o uso da técnica de falsos vizinhos mais próximos para a seleção de atributos.

A dimensão embutida é capaz de identificar o número de dimensões necessárias para desdobrar o comportamento de uma série temporal. Esse desdobramento utiliza o algoritmo de falsos vizinhos mais próximos, que analisa a relação entre dois pontos para diferentes números de dimensões. Para ilustrar o funcionamento desse método, considere um conjunto de dados X , que contém n exemplos, cada um com p atributos. Suponha que $n = \tau$ e os atributos alinhados como mostrado na Figura 3.1 (no exemplo, $m = 2$ e $\tau = n = 3$). Ao utilizar esse atraso e aplicando o teorema de imersão como descrito por Takens (1980) para identificar a dimensão embutida (m), o que se está comparando é a relação entre atributos.

Portanto, é possível utilizar o algoritmo de falsos vizinhos mais próximos, que tem como

$$\begin{array}{ccc}
& & f_{1,1} \\
f_{1,1} & f_{1,2} & \\
f_{2,1} & f_{2,2} & \Rightarrow f_{1,2} \\
f_{3,1} & f_{3,2} & f_{2,1} \\
& & f_{2,2} \\
& & f_{3,1} \\
& & f_{3,2}
\end{array}$$

Figura 3.1: Redefinindo um conjunto de dados como uma série temporal ($n = 2$ and $\tau = m = 3$).

base o trabalho de Kennel *et al.* (1992a) e Takens (1980), para identificar o número de atributos para selecionar.

A técnica FNN produz como resultado a fração de falsos vizinhos, que tem um forte relacionamento com as distâncias entre exemplos reconstruídos no espaço m -dimensional. Quanto menor a fração para um dado m , melhor é a reconstrução em m dimensões. Ao utilizar essa fração, é possível analisar como o conjunto de dados está organizado e o quão informativo ele é, de acordo com número de dimensões (atributos). De acordo com Kennel *et al.* (1992a), quando a fração de falsos vizinhos atinge 0, nenhuma mudança significativa na distância dos pontos foi identificada. Assim, existe informação suficiente para desdobrar e entender o comportamento do que está sendo estudado.

A fração de falsos vizinhos mais próximos (C^d) para uma dada dimensão d é definida pela Equação 3.1, onde τ é o número de atributos (ou o atraso) e R_{tol} é o limiar que define se a posição relativa entre dois exemplos apresentou mudanças significativas. A variação de distância $V_{i,j}^d$, definida na Equação 3.2, é a variação da distância entre dois pontos ou exemplos quando diferentes números de dimensões são usadas. Nessa variação $R_d^2(\cdot)$ e $R_{d+1}^2(\cdot)$ representam a distância entre dois exemplos considerando d e $d + 1$ dimensões, como definido nas Equações 2.3 e 2.4, respectivamente.

$$C^d = \frac{\sum_{i=1}^{\tau-1} \sum_{j>i}^{\tau} \text{sign}[V_{i,j}^d - R_{tol}]}{\frac{\tau(\tau-1)}{2}} \quad (3.1)$$

$$V_{ij}^d = \sqrt{\frac{R_{d+1}^2(x_i, x_j) - R_d^2(x_i, x_j)}{R_d^2(x_i, x_j)}} \quad (3.2)$$

Assim, o FNN supõe inicialmente a dimensão embutida $m = 1$, avaliando quão distantes exemplos, considerando $\mathbb{R}^1$, estão de seus vizinhos ao calcular as distâncias entre $f_{i,1}$ e $f_{j,1} \forall i \neq j$. Quando a dimensão embutida é igual a $m = 2$, o FNN calcula as distâncias entre os pares $(f_{i,1}, f_{i,2})$ e $(f_{j,1}, f_{j,2}) \forall i \neq j$. Generalizando, quando a dimensão embutida é τ' , o FNN calcula as distâncias entre todos os pares $(f_{i,1}, \dots, f_{i,\tau'})$ e $(f_{j,1}, \dots, f_{j,\tau'}) \forall i \neq j$.

Quando comparado com a ideia básica de comparar a distância média entre diferentes números de dimensões, o FNN tem a vantagem de levar em consideração como esses exemplos

Tabela 3.1: Conjunto de dados de exemplo

	F1	F2	F3	F4	F5	F6
A	2	3	26	3	1	12
B	3	5	25	6	2	12
C	5	15	12	13	3	12
D	12	17	8	15	11	12
E	13	16	22	18	12	12
F	15	6	17	6	13	12
G	21	7	1	8	20	12
H	22	8	2	6	21	12

Tabela 3.2: Distâncias – F1

	A	B	C	D	E	F	G	H
A	0,00	1,00	3,00	10,00	11,00	13,00	19,00	20,00
B	1,00	0,00	2,00	9,00	10,00	12,00	18,00	19,00
C	3,00	2,00	0,00	7,00	8,00	10,00	16,00	17,00
D	10,00	9,00	7,00	0,00	1,00	3,00	9,00	10,00
E	11,00	10,00	8,00	1,00	0,00	2,00	8,00	9,00
F	13,00	12,00	10,00	3,00	2,00	0,00	6,00	7,00
G	19,00	18,00	16,00	9,00	8,00	6,00	0,00	1,00
H	20,00	19,00	17,00	10,00	9,00	7,00	1,00	0,00

se relacionam entre si. Essa relação, que não pode ser expressada pela distância média, é o que torna o método FNN uma ferramenta poderosa para identificar o número de atributos para selecionar.

3.3 Análise da interação entre as distâncias

Para ilustrar o funcionamento do método de falsos vizinhos mais próximos, considere a Tabela 3.1. Nessa tabela foi definido um conjunto de dados que contém 6 atributos e 8 exemplos. Os quatro primeiros atributos ($F1, F2, F3$ e $F4$) contêm informação útil. O atributo $F5$ é redundante e seu comportamento é semelhante ao do atributo $F1$. Por último, o atributo $F6$ apresenta um valor constante.

A partir desse conjunto de dados, o método FNN é aplicado para definir os atributos que serão selecionados. Considere $R_{tol} = 2$, conforme indicam Kennel *et al.* (1992a).

Inicialmente, é necessário calcular as distâncias entre os exemplos do conjunto de dados, considerando o atributo $F1$ (Tabela 3.2). Em seguida, o atributo $F2$ também passa a ser considerado. As distâncias entre os exemplos utilizando $F1$ e $F2$ são apresentadas na Tabela 3.3.

De acordo com método FNN, após calcular as distâncias entre exemplos considerando $F1$ e $F1, F2$, se faz necessário avaliar a quantidade de falsos vizinhos por meio da equação 3.1. Para esse cálculo, apenas o vizinho mais próximo de cada exemplo é considerado. Assim, temos que os pares de exemplos (B, C) e (E, F) são falsos vizinhos, pois $V_{BC}^1 = 2,35 > R_{tol}$

Tabela 3.3: Distâncias – F1F2

	A	B	C	D	E	F	G	H
A	0,00	1,12	6,18	8,60	8,51	6,67	9,71	10,31
B	1,12	0,00	5,10	7,50	7,43	6,02	9,06	9,62
C	6,18	5,10	0,00	3,64	4,03	6,73	8,94	9,19
D	8,60	7,50	3,64	0,00	0,71	5,70	6,73	6,73
E	8,51	7,43	4,03	0,71	0,00	5,10	6,02	6,02
F	6,67	6,02	6,73	5,70	5,10	0,00	3,04	3,64
G	9,71	9,06	8,94	6,73	6,02	3,04	0,00	0,71
H	10,31	9,62	9,19	6,73	6,02	3,64	0,71	0,00

Tabela 3.4: Distâncias – F1F2F3

	A	B	C	D	E	F	G	H
A	0,00	0,82	6,23	8,30	5,83	5,36	10,55	10,55
B	0,82	0,00	5,51	7,56	5,06	4,82	10,02	9,99
C	6,23	5,51	0,00	2,77	4,28	4,78	7,00	6,98
D	8,30	7,56	2,77	0,00	4,69	4,84	5,06	4,91
E	5,83	5,06	4,28	4,69	0,00	3,79	8,07	7,78
F	5,36	4,82	4,78	4,84	3,79	0,00	5,71	5,56
G	10,55	10,02	7,00	5,06	8,07	5,71	0,00	0,58
H	10,55	9,99	6,98	4,91	7,78	5,56	0,58	0,00

e $V_{EF}^1 = 2,35 > R_{tol}$, o que significa que a análise deve prosseguir considerando mais um atributo, no caso $F3$.

Na segunda iteração, onde $m = 2$, é necessário calcular as distâncias entre os exemplos considerando 3 atributos (Tabela 3.4).

Analisando os vizinhos mais próximos considerando $F1, F2$ e calculando a variação da distância segundo a Equação 3.1, descobre-se que o par (D, E) é uma falsa vizinhança, pois $V_{DE}^2 = 6,56 > R_{tol}$. Portanto é necessário continuar o processo analisando o atributo $F4$.

Na terceira iteração, $m = 3$, as distâncias entre os exemplos considerando 4 atributos são calculadas (Tabela 3.5). Analisando os pares de vizinhos mais próximos, verificou-se que todos são vizinhos verdadeiros, ou seja $V_{x_i, x_j}^2 < R_{tol}$ em todos os casos. Assim, é possível afirmar que o atributo $F4$ não adiciona novas informações e que apenas $m = 3$ atributos são suficientes para descrever o conjunto de dados.

Considerando o $F6$ como o quarto atributo a ser analisado (Tabela 3.6), o mesmo resultado é obtido, ou seja, todos os pares de vizinhos mais próximos não apresentaram variação na distância relativa.

Entretanto, se a ordem de análise dos atributos não for favorável, o FNN pode apresentar resultados errôneos. Sabe-se que três atributos ($F1, F2, F3$) foram a quantidade ótima a ser considerada, porém se o segundo atributo a ser avaliado for o $F5$, um atributo redundante, o FNN apresenta um resultado errado.

A Tabela 3.7 apresenta as distâncias considerando os atributos $F1$ e $F5$. Avaliando os pares

Tabela 3.5: Distâncias – F1F2F3F4

	A	B	C	D	E	F	G	H
A	0,00	0,97	5,30	6,91	5,76	4,09	8,01	7,95
B	0,97	0,00	4,49	6,10	4,83	3,61	7,53	7,50
C	5,30	4,49	0,00	2,14	3,45	3,99	5,40	5,52
D	6,91	6,10	2,14	0,00	3,60	4,27	4,18	4,32
E	5,76	4,83	3,45	3,60	0,00	4,13	6,55	6,56
F	4,09	3,61	3,99	4,27	4,13	0,00	4,31	4,17
G	8,01	7,53	5,40	4,18	6,55	4,31	0,00	0,66
H	7,95	7,50	5,52	4,32	6,56	4,17	0,66	0,00

Tabela 3.6: Distâncias – F1F2F3F6

	A	B	C	D	E	F	G	H
A	0,00	0,61	4,67	6,22	4,37	4,02	7,91	7,91
B	0,61	0,00	4,13	5,67	3,79	3,61	7,52	7,50
C	4,67	4,13	0,00	2,08	3,21	3,59	5,25	5,23
D	6,22	5,67	2,08	0,00	3,52	3,63	3,79	3,68
E	4,37	3,79	3,21	3,52	0,00	2,84	6,05	5,84
F	4,02	3,61	3,59	3,63	2,84	0,00	4,28	4,17
G	7,91	7,52	5,25	3,79	6,05	4,28	0,00	0,43
H	7,91	7,50	5,23	3,68	5,84	4,17	0,43	0,00

de vizinhos, tem-se que $V_{x_i, x_j}^1 < R_{tol}$ para todos os pares de exemplos. Ou seja, por essa análise, apenas $F1$ é suficiente para descrever o conjunto de dados, o que não é verdade. De maneira similar, considerando $F6$ como o segundo atributo a ser avaliado (distâncias na Tabela 3.8), chega-se à mesma conclusão.

Portanto, se não dispusermos de informações sobre os atributos, é necessário avaliar todas as possíveis permutações do conjunto para que se possa chegar a um resultado válido. No entanto, essa análise tem um alto custo pois apresenta um comportamento exponencial.

Tabela 3.7: Distâncias – F1F5

	A	B	C	D	E	F	G	H
A	0,00	0,71	1,80	7,07	7,78	8,85	13,44	14,14
B	0,71	0,00	1,12	6,36	7,07	8,14	12,73	13,44
C	1,80	1,12	0,00	5,32	6,02	7,07	11,67	12,38
D	7,07	6,36	5,32	0,00	0,71	1,80	6,36	7,07
E	7,78	7,07	6,02	0,71	0,00	1,12	5,66	6,36
F	8,85	8,14	7,07	1,80	1,12	0,00	4,61	5,32
G	13,44	12,73	11,67	6,36	5,66	4,61	0,00	0,71
H	14,14	13,44	12,38	7,07	6,36	5,32	0,71	0,00

Tabela 3.8: Distâncias – F1F6

	A	B	C	D	E	F	G	H
A	0,00	0,50	1,50	5,00	5,50	6,50	9,50	10,00
B	0,50	0,00	1,00	4,50	5,00	6,00	9,00	9,50
C	1,50	1,00	0,00	3,50	4,00	5,00	8,00	8,50
D	5,00	4,50	3,50	0,00	0,50	1,50	4,50	5,00
E	5,50	5,00	4,00	0,50	0,00	1,00	4,00	4,50
F	6,50	6,00	5,00	1,50	1,00	0,00	3,00	3,50
G	9,50	9,00	8,00	4,50	4,00	3,00	0,00	0,50
H	10,00	9,50	8,50	5,00	4,50	3,50	0,50	0,00

3.4 Técnica FQFNN

Conforme mostrado na Seção 3.3, a técnica de falsos vizinhos mais próximos (FNN), quando considerado o problema da seleção de atributos, apresenta uma complexidade exponencial. Esse custo computacional é explicado ao se considerar que todos os arranjos do conjunto de atributos devem ser analisados, para que se possa definir quantos atributos serão selecionados.

Entretanto, este trabalho propõe uma variação da técnica FNN para avaliar um conjunto de atributos ordenados por um critério de relevância. Essa ordenação é realizada por uma técnica de *ranking* de atributos e a definição de relevância depende de qual técnica é utilizada. A Seção 3.5 apresenta os algoritmos utilizados nesta Tese para a confecção do *ranking*.

Ao utilizar um *ranking* de atributos, pode-se afirmar que os mais relevantes estão alocados em melhores posições e são, portanto, analisados primeiro. À medida em que mais atributos são considerados, a quantidade de informação adicionada tende a diminuir. Assim, a distância entre exemplos quando avaliada para diferentes quantidades de atributos (Equação 3.2) tende a se manter estável, diminuindo a fração de falsos vizinhos.

A técnica FNN aplicada à seleção de atributos não-supervisionada é denominada FQFNN – *Feature Quantification by False Nearest Neighbor* (Andrade Filho *et al.*, 2011). A técnica FQFNN é capaz de, com base em um *ranking* de atributos, identificar a quantidade de atributos $p' < p$ que descreve o conjunto de dados original. Essa técnica utiliza uma variação do algoritmo de falsos vizinhos mais próximos para calcular a relação entre exemplos considerando diferentes números de dimensões. O Algoritmo 3.1 apresenta um pseudo-código da técnica FQFNN.

A complexidade estimada da técnica FQFNN é $O(pn^2)$, onde p é a quantidade de atributos e n é a quantidade de exemplos no conjunto de dados. No entanto, quando poucos atributos são necessários para descrever o conjunto de dados, a complexidade estimada se aproxima de $O(n^2)$.

Uma limitação da técnica FQFNN é a dependência de um algoritmo de *ranking*. A qualidade do resultado gerado pela técnica é dependente da ordem com que os atributos são avaliados, como mostrado na Seção 3.3.

Algoritmo 3.1: Algoritmo da técnica FQFNN

Entrada: Ranking de atributos $Rank_F$

Entrada: Tolerância R_{tol}

Saída: Quantidade de atributos a serem selecionados

```
1 inicio
2    $D_{anterior} \leftarrow 0$ ;
3   para  $k \in \{1, 2, \dots, p\}$  fazer
4     /* Quantidade de Falsos vizinhos */
5      $Qtd_{FV} \leftarrow 0$ ;
6     Calcule a matriz de distância  $D_{atual}$ , considerando  $k$  atributos de  $Rank_F$ ;
7     para  $i \in \{x_1, x_2, \dots, x_n\}$  fazer
8        $j \leftarrow$  exemplo mais próximo a  $i$ ;
9        $R_{dist} = \frac{D_{atual}(i,j) - D_{anterior}(i,j)}{D_{anterior}(i,j)}$ ;
10      /* Se houve mudança significativa na distância  $R_{dist}$  */
11      se  $R_{dist} > R_{tol}$  então
12        | incremente  $Qtd_{FV}$ ;
13      fim
14    fin
15     $D_{anterior} \leftarrow D_{atual}$ ;
16    Calcule o percentual  $P_{FV}$  de falsos vizinhos;
17    se  $P_{FV} = 0$  então
18      | /* Adicionar o atributo  $k$  não trouxe novas */
19      | informações
20      | retorna  $k - 1$  atributos a serem considerados
21    senão
22      | Armazene  $k$  e  $P_{FV}$ ;
23    fim
24  fin
25  retorna  $k - 1$ , onde  $k$  apresenta o menor  $P_{FV}$ 
26 fin
```

3.5 Técnicas para *ranking* de atributos

Devido à necessidade de se utilizar uma técnica de *ranking* para que os atributos fossem ordenados de uma maneira significativa, foram escolhidas 4 técnicas para realizar essa tarefa.

A primeira técnica, SPEC(Zhao & Liu, 2007), previamente definida na Seção 2.4 foi selecionada como técnica de *ranking* básica devido à cooperação entre os laboratórios BIOCUM, do ICMC/USP, e DMML, da Arizona State University. Essa técnica é um filtro e tem como resultado o *ranking* de atributos, que é utilizado como entrada para a técnica FQFNN.

A segunda técnica de *ranking* é baseada na Informação Mútua e avalia a correlação entre os valores dos atributos e está definida na Equação 3.3, onde A e B são atributos, $p(a, b)$ é a função de densidade de probabilidade conjunta de A e B e $p_1(a)$ e $p_2(a)$ são as funções de probabilidades marginais de A e B , respectivamente.

$$I(A; B) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p_1(x)p_2(y)} \right) \quad (3.3)$$

Após calcular o valor de informação mútua para cada atributo f_i em relação a f_j , $\forall i \neq j$, o *ranking* dos atributos é calculado pelo valor médio de Informação Mútua de cada atributo.

A terceira abordagem para o cálculo do *ranking* de atributos tem como objetivo reduzir a influência de atributos correlacionados, que pode ocorrer ao utilizar a técnica SPEC. Essa abordagem é denominada SPECMI, cujo pseudo-código é apresentado no Algoritmo 3.2. Nessa técnica, são utilizados como entrada o resultado da técnica SPEC, o par ordenado $R_{SPEC}(f_i, S(f_i))$, onde f_i é a referência do atributo e $S(f_i)$ é o valor definido pelo SPEC para esse atributo; e a matriz de similaridade que é gerada por meio da Informação Mútua $MI(f_i, f_j)$.

A última técnica de *ranking* considerada nesta Tese, é um *ensemble* do resultado de três abordagens: SPEC, Mutual Information e a entropia interna do atributo. A entropia utilizada foi a implementada no pacote *entropy* do *software* R, que corresponde à entropia de Shannon (1948). O *ensemble* utilizado considera a posição de um atributo nos três *rankings* e calcula a nova posição com base na média dessas posições (Zhou, 2012). O desempate é realizado levando-se em consideração a posição do atributo nos *rankings*, por ordem de prioridade, SPEC, Mutual Information e Entropia.

3.6 Considerações Finais

Nesse capítulo foi apresentada a definição a técnica FQFNN. Essa técnica tem como bases ferramentas de teoria do caos, mais especificamente o método de falsos vizinhos mais próximos.

A técnica proposta permite avaliar um conjunto de dados para definir quantos atributos devem ser selecionados e tem uma complexidade estimada em $O(pn^2)$, onde p é a quantidade de atributos e n é a quantidade de exemplos no conjunto de dados. No entanto, a técnica depende da definição de um *ranking* de atributos.

Algoritmo 3.2: Pseudo-código da técnica SPECMI

Entrada: *Ranking* gerado pela técnica SPEC: $R_{SPEC}(f_i, S(f_i))$

Entrada: Matriz de Similaridade: $MI(f_i, f_j)$

Saída: *Ranking*: R_{MI}

```
1 inicio
2    $R_{MI} \leftarrow \emptyset$ ;
3   repita
4     /* Adiciona o atributo mais relevante, de acordo com o
       SPEC, a  $R_{MI}$  */
5      $R_{MI} \leftarrow f_i$  com maior valor de  $S(f_i)$ ;
6     /* Salva a referência do atributo */
7      $f_{atual} \leftarrow f_i$  com maior valor de  $S(f_i)$ ;
8     Remove  $f_{atual}$  de  $R_{SPEC}$ ;
9     para todo  $f_i$  em  $R_{SPEC}$  fazer
10      /* atualiza o valor de  $S(f_i)$  com base na similaridade
        a  $f_{atual}$  */
11       $S(f_i) \leftarrow S(f_i) * (1 - MI(f_{atual}, f_i));$ 
12    fin
13  até  $R_{SPEC} = \emptyset$ ;
14  retorna  $R_{MI}$ 
15 fin
```

Algoritmo 3.3: Pseudo-código da geração do *ensemble*

Entrada: *Ranking* gerado pela técnica SPEC: $pos_{SPEC}(f_i)$

Entrada: *Ranking* gerado pela técnica de Informação Mútua: $pos_{RMI}(f_i)$

Entrada: *Ranking* gerado pelo cálculo da entropia: $pos_{entropy}(f_i)$

Saída: *Ranking*: $R_{ensemble}$

```
1 inicio
2   para todo  $f_i \in F$  fazer
3     /* Calcula o novo valor do ranking */
4      $R_{ensemble}(f_i) \leftarrow \frac{(pos_{SPEC}(f_i) + pos_{RMI}(f_i) + pos_{entropy}(f_i))}{3}$ ;
5   fin
6   Ordena  $R_{ensemble}(f_i)$  de maneira decrescente;
7   Em caso de empate, considere a menor posição no: ranking SPEC, ranking Mutual
  Information, ranking Entropia;
8  retorna  $R_{ensemble}$ 
9 fin
```

Nesse capítulo também foram apresentadas as técnicas de *ranking* que são avaliadas juntamente com a técnica FQFNN no próximo capítulo. As técnicas de *ranking* tem como base o algoritmo SPEC (Zhao & Liu, 2007), esse algoritmo foi escolhido para tirar proveito da interação entre os laboratórios BIOCOT, ICMC/USP, e o DMML (Data Mining and Machine Learning) da Arizona State University, durante o estágio sanduíche.

Experimentos e resultados

4.1 Considerações Iniciais

O presente capítulo apresenta os experimentos realizados e os resultados obtidos. Três diferentes grupos de conjuntos de dados foram utilizados nesses experimentos. O primeiro grupo possui conjuntos de dados selecionados do repositório de dados da UCI (Asuncion & Newman, 2007). Esses conjuntos de dados apresentam poucos atributos e compreendem diversas áreas do conhecimento. O segundo grupo¹ possui conjuntos que apresentam uma maior quantidade de atributos e são relacionados às áreas processamento de imagens e análise de dados de *micro-arrays*. Por último, foi utilizado um conjunto de dados real pertencente à área de botânica. Esse conjunto está relacionado à identificação de espécies endêmicas no estado do Rio de Janeiro.

4.2 Planejamento dos Experimentos

Conforme mencionado, os conjuntos de dados utilizados nos experimentos deste trabalho estão divididos em três grupos. Os dois primeiros grupos são conjuntos de dados públicos. A diferença entre esses grupos está na quantidade de atributos dos conjuntos de dados. O terceiro grupo é composto por um conjunto de dados de espécies endêmicas do Rio de Janeiro. Os principais aspectos desse conjunto de dados serão apresentados na Seção 4.5.

O modo como os experimentos utilizando os conjuntos de dados dos grupos *I* e *II* foram realizados está resumido na Figura 4.1. Duas etapas foram definidas, a primeira é a aplicação da técnica de seleção de atributos aos conjuntos de dados. As seguintes técnicas foram investigadas: SSF (Covões & Hruschka, 2011); MitraFS (Mitra *et al.*, 2002); FSSEM-k (Dy & Brodley, 2004); FQFNN + SPEC (Zhao & Liu, 2007); FQFNN + MI; FQFNN + ensemble

¹Disponível em: <http://featureselection.asu.edu>

(SPEC/MI/Entropia); FQFNN + SPECMI. A aplicação dessas técnicas aos conjuntos de dados tem como resultado um subconjunto de atributos. Esse subconjunto, assim como o conjunto original de atributos, são utilizados como entrada para a etapa seguinte. A Tabela 4.1 apresenta os parâmetros utilizados na aplicação das técnicas neste trabalho, onde k representa o número real de grupos. Os parâmetros foram definidos de acordo com as indicações dos autores de cada técnica. No caso da técnica FQFNN os parâmetros utilizados são baseados nos resultados de Kennel *et al.* (1992a).

Tabela 4.1: Principais parâmetros utilizados nos experimentos

Técnica	Parâmetros
SSF	$\gamma = 2$
MitraFS	Análise exploratória de k_m^2
FSSEM-k	$k_{max} = 2k$
FQFNN+SPEC	$R_{tol} = 2$
FQFNN+MI	$R_{tol} = 2$
FQFNN+SPECMI	$R_{tol} = 2$
FQFNN+ensamble	$R_{tol} = 2$

Figura 4.1: Planejamento dos experimentos

A segunda etapa avalia os atributos selecionados, comparando as diferentes técnicas. Essa comparação é realizada em conjunto com o algoritmo de agrupamento de dados *K-means*. Para realizar uma comparação mais justa, foi criado um conjunto de centroides pré-definidos, de modo que os mesmos centroides fossem utilizados por técnicas diferentes. Dessa maneira, se duas técnicas utilizarem o mesmo atributo, os valores iniciais do centroide para esse atributo serão iguais. Como atributos diferentes podem ser selecionados pelas técnicas, essa foi a melhor maneira encontrada para tornar a comparação justa.

Em seguida, o algoritmo de agrupamento *K-means* foi executado, considerando os centroides iniciais definidos anteriormente; Para cada conjunto de dados, o valor de K foi definido como igual à sua quantidade de classes. As partições resultantes da execução do *K-means* são

avaliadas por meio de duas medidas de qualidade. A primeira delas, o Rand Corrigido (CR), é utilizada como uma métrica externa e compara a partição obtida pelo algoritmo de agrupamento com a partição real. A segunda, Silhueta, é utilizada para avaliar como os grupos gerados se apresentam.

Para cada conjunto de dados e técnica, a etapa de avaliação é repetida pelo menos 30 vezes. Mais iterações são realizadas caso não se atinja um intervalo de confiança de 95%, considerando-se o resultado das medidas de qualidade CR e Silhueta. Em seguida, os resultados obtidos pelas diversas técnicas são comparados por meio de um teste de hipótese.

O grupo de conjuntos de dados III, formado pelo conjunto de dados de espécies endêmicas do Rio de Janeiro, é avaliado de acordo com a quantidade de atributos selecionados, nesse conjunto, a qualidade da partição gerada não é avaliada. Assim, apenas a etapa de aplicação da técnica é realizada. Mais detalhes sobre o conjunto de dados e os experimentos realizados são abordados na seção 4.5.

4.3 Grupo I: conjuntos de dados da UCI

Do grupo I fazem parte os conjuntos de dados do repositório UCI³ (Asuncion & Newman, 2007) descritos na tabela 4.2

Tabela 4.2: Conjuntos de dados do Grupo I

Dataset	# Classes	# Atributos	# Exemplos
Breast tissue	6	9	106
Cardiotocography	3	21	2126
Dermatology	6	34	358
Ecoli	8	7	336
Glass	7	9	214
Heart	2	13	270
Iris	3	4	150
Lung Cancer	3	56	27
Soybean	4	35	47
Swissbank	2	6	200
Wdbc	2	30	569
Yeast	4	20	205

Nos experimentos realizados para esse grupo de dados, a etapa de aplicação da técnica resultou em 96 subconjuntos de atributos a serem analisados (subconjuntos resultantes da aplicação de 7 técnicas a 12 conjuntos de dados, mais o conjunto total de atributos para os 12 conjuntos de dados). Cada um desses 96 subconjunto de atributos foi avaliado e alguns dos resultados obtidos são detalhados a seguir.

A Figura 4.2 apresenta os resultados da obtidos a partir da aplicação das técnicas ao conjunto de dados *Iris*. Para esse conjunto, os subconjuntos gerados foram idênticos para as 7 técnicas

³<http://archive.ics.uci.edu/ml/>

utilizadas. Foram selecionados os mesmos dois atributos por todas as técnicas e, por causa do modo como os centroides foram definidos, isso permitiu que os resultados fossem idênticos. A seleção de dois atributos permitiu melhorar a qualidade da partição, quando se consideram os valores das medidas CR e Silhueta.

Figura 4.2: Conjunto de dados Iris

No caso do conjunto de dados *Breast Tissue*, a utilização da técnica FQFNN permitiu a seleção de menos atributos (Figura 4.3 (a)) do que as demais técnicas avaliadas. Entretanto, não houve diferença na avaliação das métricas (b), mesmo quando menos atributos foram considerados.

Figura 4.3: Conjunto de dados Breast Tissue

No conjunto de dados *Soybean* (Figura 4.5) observou-se que todas as técnicas consideradas selecionaram menos de 50% dos atributos e, mesmo assim, a qualidade medida por meio do CR apresentou melhoria. Além disso, é notável a melhoria da Silhueta quando consideradas as técnicas SSF, FQFNN+SPEC, FQFNN+SPECMI e FQFNN+ensamble.

Figura 4.4: Conjunto de dados Soybean

Por outro lado, nos conjuntos de dados Yeast, Ecoli e Glass (Figuras 4.5, 4.6 e 4.7, respectivamente) não houve melhoria da qualidade da partição resultante do uso das diferentes técnicas de seleção. Nesse caso, a vantagem do uso de técnicas de seleção de atributos se dá apenas pela redução da quantidade de informação a ser processada.

Figura 4.5: Conjunto de dados Yeast

O único caso em que utilização do FQFNN, juntamente com as técnicas de *ranking*, obteve um resultado superior foi na análise do conjunto de dados *Swissbank*, Figura 4.8. No entanto esse resultado só foi obtido porque o FQFNN indicou mais atributos do que as outras técnicas.

Os piores desempenhos da técnica FQFNN, com relação à medida CR, foram observados nos conjunto de dados *Lungcancer* e *WDBC*, Figuras 4.9 e 4.10, respectivamente.

De forma resumida, os resultados obtidos nos experimentos desse grupo estão divididos nas tabelas 4.3, 4.4, 4.5, 4.6. A Tabela 4.3 apresenta os valores da medida CR obtidas pelas técnicas utilizadas na comparação. De maneira complementar, a Tabela 4.4 apresenta os resultados da medida CR quando utilizada a técnica proposta com os diversos rankings. As tabelas 4.5 e

Figura 4.6: Conjunto de dados Ecoli

Figura 4.7: Conjunto de dados Glass

Figura 4.8: Conjunto de dados Swissbank

Figura 4.9: Conjunto de dados Lung cancer

Figura 4.10: Conjunto de dados WDBC

4.6 apresentam os resultados da Silhueta para a base de comparação e para a técnica FQFNN, respectivamente.

De acordo com as Tabelas 4.3 e 4.4, é possível ver que a técnica FQFNN seleciona um número menor de atributos em 10 dos 12 conjuntos desse grupo, quando consideramos a utilização da técnica SPEC como algoritmo de ranking. No entanto, se considerado o ranking de Mutual Information, em 8 conjuntos, a técnica FQFNN seleciona menos atributos. Essa diferença é dada pela formação do *ranking* e evidencia a importância da seleção de um bom algoritmo para essa tarefa. Nesse grupo de experimento, as técnicas SPEC e SPECMI obtiveram resultados semelhantes.

Além da quantidade de atributos selecionados, é importante verificar a qualidade das partições formadas ao utilizar esses atributos. Em relação à métrica CR (Tabelas 4.3 e 4.4), que avalia a correspondência da partição gerada com a partição real, em 4 casos houve uma perda de qualidade ao se comparar à utilização de todos os atributos, enquanto que em 5 casos houve

Tabela 4.3: Resultados CR – Base comparativa – Grupo I

Dataset	SSF		MitraFS		FSSEM-k		Completo	
	F	CR	F	CR	F	CR	F	CR
Breast tissue	6	0.36 ± 0.02	6	0.36 ± 0.02	6	0.36 ± 0.02	9	0.18 ± 0.01
Cardiotocography	16	0.18 ± 0.06	19	0.17 ± 0.05	15	0.12 ± 0.03	21	0.04 ± 0.01
Dermatology	13	0.22 ± 0.04	18	0.26 ± 0.06	14	0.23 ± 0.05	34	0.37 ± 0.08
Ecoli	5	0.25 ± 0.04	5	0.25 ± 0.04	5	0.25 ± 0.04	7	0.41 ± 0.05
Glass	4	0.34 ± 0.02	4	0.31 ± 0.01	5	0.35 ± 0.01	9	0.45 ± 0.03
Heart	7	0.12 ± 0.08	8	0.09 ± 0.05	5	0.14 ± 0.09	13	0.03 ± 0.01
Iris	2	0.88 ± 0.02	2	0.88 ± 0.02	2	0.88 ± 0.02	4	0.68 ± 0.12
Lung Cancer	15	0.29 ± 0.04	21	0.31 ± 0.06	12	0.28 ± 0.04	56	0.20 ± 0.09
Soybean	12	0.55 ± 0.04	16	0.58 ± 0.05	15	0.57 ± 0.08	35	0.28 ± 0.08
Swissbank	4	0.93 ± 0.05	4	0.90 ± 0.02	4	0.93 ± 0.05	6	1.00 ± 0.01
Wdbc	17	0.63 ± 0.05	18	0.45 ± 0.02	17	0.61 ± 0.04	30	0.49 ± 0.01
Yeast	3	0.75 ± 0.01	9	0.72 ± 0.01	6	0.71 ± 0.02	20	0.92 ± 0.08

Tabela 4.4: Resultados CR – FQFNN e rankings – Grupo I

Dataset	FQFNN + SPEC		FQFNN + MI		FQFNN + SPECMI		FQFNN + ensemble	
	F	CR	F	CR	F	CR	F	CR
Breast tissue	4	0.32 ± 0.02	5	0.34 ± 0.02	4	0.32 ± 0.02	6	0.36 ± 0.02
Cardiotocography	18	0.13 ± 0.04	17	0.15 ± 0.06	15	0.12 ± 0.07	16	0.17 ± 0.08
Dermatology	11	0.20 ± 0.07	13	0.21 ± 0.05	10	0.20 ± 0.04	11	0.21 ± 0.06
Ecoli	5	0.21 ± 0.02	6	0.24 ± 0.03	5	0.21 ± 0.02	5	0.21 ± 0.01
Glass	4	0.34 ± 0.02	5	0.35 ± 0.01	4	0.34 ± 0.02	5	0.35 ± 0.01
Heart	4	0.17 ± 0.12	8	0.08 ± 0.06	5	0.14 ± 0.09	5	0.14 ± 0.09
Iris	2	0.88 ± 0.02	2	0.88 ± 0.02	2	0.88 ± 0.02	4	0.68 ± 0.12
Lung Cancer	10	0.17 ± 0.04	9	0.16 ± 0.03	8	0.16 ± 0.05	11	0.23 ± 0.03
Soybean	9	0.52 ± 0.09	16	0.57 ± 0.04	10	0.53 ± 0.07	12	0.54 ± 0.05
Swissbank	5	0.98 ± 0.02	5	0.98 ± 0.02	5	0.98 ± 0.02	5	0.98 ± 0.02
Wdbc	12	0.21 ± 0.02	13	0.23 ± 0.05	11	0.19 ± 0.02	13	0.23 ± 0.03
Yeast	5	0.73 ± 0.01	9	0.73 ± 0.07	6	0.70 ± 0.04	7	0.72 ± 0.03

melhora na qualidade da partição. Em comparação às outras técnicas, a qualidade foi similar em 10 casos e pior em 2.

Para comparar as diversas técnicas, foi escolhido o teste de Friedman para múltiplas comparações (Demsär, 2006). A hipótese nula foi definida como a igualdade no desempenho entre as técnicas comparadas. No caso de rejeição dessa hipótese, o teste de Nemenyi foi utilizado para identificar quais técnicas têm resultados diferentes. Em todas comparações, o nível de confiança no teste foi definido em 95%.

Os resultados da avaliação obtidos pelo teste de hipótese são apresentados na Tabela 4.7. Nessa tabela cada célula é formada por 8 símbolos, cada símbolo representa um diferente resultado do teste de hipótese. Cada célula compara a técnica da respectiva coluna com as outras técnicas (a ordem é a mesma da linha de título). Por exemplo, para o conjunto de dados Breasttissue, considerando a técnica SSF, temos a seguinte sequência: “○=====▲”, o primeiro

Tabela 4.5: Resultados Silhueta – Base comparativa – Grupo I

Dataset	SSF		MitraFS		FSSEM-k		Completo	
	F	Sil	F	Sil	F	Sil	F	Sil
Breast tissue	6	0.59 ± 0.01	6	0.59 ± 0.01	6	0.59 ± 0.01	9	0.59 ± 0.01
Cardiotocography	16	0.29 ± 0.03	19	0.25 ± 0.01	15	0.26 ± 0.02	21	0.15 ± 0.01
Dermatology	13	0.40 ± 0.06	18	0.31 ± 0.02	14	0.34 ± 0.04	34	0.28 ± 0.01
Ecoli	5	0.28 ± 0.01	5	0.28 ± 0.01	5	0.28 ± 0.01	7	0.23 ± 0.02
Glass	4	0.39 ± 0.03	4	0.37 ± 0.02	5	0.35 ± 0.04	9	0.37 ± 0.08
Heart	7	0.48 ± 0.01	8	0.44 ± 0.04	5	0.49 ± 0.11	9	0.37 ± 0.00
Iris	2	0.66 ± 0.00	2	0.66 ± 0.00	2	0.66 ± 0.00	4	0.54 ± 0.01
Lung Cancer	15	0.11 ± 0.06	21	0.08 ± 0.02	12	0.12 ± 0.01	56	0.08 ± 0.01
Soybean	12	0.73 ± 0.08	16	0.23 ± 0.02	15	0.24 ± 0.03	35	0.15 ± 0.01
Swissbank	4	0.48 ± 0.01	4	0.34 ± 0.00	4	0.48 ± 0.01	6	0.51 ± 0.00
Wdbc	17	0.53 ± 0.00	18	0.50 ± 0.01	17	0.53 ± 0.00	30	0.69 ± 0.00
Yeast	7	0.40 ± 0.05	9	0.41 ± 0.05	6	0.41 ± 0.01	20	0.52 ± 0.10

Tabela 4.6: Resultados Silhueta – FQFNN e rankings – Grupo I

Dataset	FQFNN + SPEC		FQFNN + MI		FQFNN + SPECMI		FQFNN + ensemble	
	F	Sil	F	Sil	F	Sil	F	Sil
Breast tissue	4	0.51 ± 0.03	5	0.60 ± 0.03	4	0.51 ± 0.03	6	0.59 ± 0.01
Cardiotocography	18	0.28 ± 0.01	17	0.28 ± 0.01	15	0.27 ± 0.02	16	0.28 ± 0.02
Dermatology	11	0.46 ± 0.08	13	0.37 ± 0.05	10	0.63 ± 0.05	11	0.45 ± 0.06
Ecoli	5	0.27 ± 0.02	6	0.27 ± 0.02	5	0.27 ± 0.02	5	0.27 ± 0.02
Glass	4	0.39 ± 0.03	5	0.34 ± 0.04	4	0.39 ± 0.03	5	0.34 ± 0.04
Heart	4	0.57 ± 0.12	8	0.42 ± 0.05	5	0.50 ± 0.13	5	0.50 ± 0.13
Iris	2	0.66 ± 0.00	2	0.66 ± 0.00	2	0.66 ± 0.00	2	0.66 ± 0.00
Lung Cancer	10	0.15 ± 0.01	9	0.18 ± 0.01	8	0.22 ± 0.02	11	0.13 ± 0.01
Soybean	9	0.97 ± 0.01	16	0.20 ± 0.01	10	0.97 ± 0.01	12	0.70 ± 0.10
Swissbank	5	0.50 ± 0.00	5	0.50 ± 0.00	5	0.50 ± 0.00	5	0.50 ± 0.00
Wdbc	12	0.54 ± 0.00	13	0.54 ± 0.00	11	0.57 ± 0.00	13	0.54 ± 0.00
Yeast	5	0.42 ± 0.00	9	0.40 ± 0.06	6	0.41 ± 0.01	7	0.39 ± 0.06

símbolo, no caso $\circ$, representa a comparação da técnica SSF com ela mesma (no caso, o teste é ignorado); o segundo símbolo, $=$, é a comparação da técnica SSF com a técnica MitraFS e assim por diante. O último símbolo, $\blacktriangle$, é a comparação da técnica SSF com o conjunto completo de atributos.

O mesmo teste de hipótese foi utilizado para a avaliação da estrutura dos grupos formados por meio da medida de Silhueta. A Tabela 4.8 apresenta essa comparação. Nessa medida, a técnica FQFNN teve um desempenho pelo menos igual em 10 conjuntos de dados desse grupo, quando comparado com as outras técnicas.

É importante notar que a técnica FQFNN seleciona, em média, menos do que 50% dos atributos dos conjuntos de dados desse grupo. Utilizar menos atributos significa utilizar menos informações sobre o conjunto de dados, o que pode explicar a perda de qualidade na partição gerada (como os conjuntos *Wdbc* and *Ecoli*). Por outro lado, ao realizar a seleção, pode-se

Tabela 4.7: Resultado da avaliação do teste de hipótese – CR – Grupo I

Dataset	SSF	MitraFS	FSSEM-k	FQFNN SPEC	FQFNN MI	FQFNN SPECMI	FQFNN ensemble	Completo
Breast tissue	○=====▲	○=====▲	○=====▲	○=====▲	○=====▲	○=====▲	○=====▲	▼▼▼▼▼○
Cardiotocography	○=▲▲=====	○=▲▲=====	▼▼○=====	▼▼○=====	○=====	○=====	○=====	▼▼▼▼▼○
Dermatology	○=====	○=====	○=====	○=====	○=====	○=====	○=====	=====○
Ecoli	○=====▼	○=====▼	○=====▼	○=====▼	○=====▼	○=====▼	○=====▼	▲▲▲▲▲○
Glass	○▲=====▼	▼○▼▼▼▼▼	○=====▼	○=====▼	○=====▼	○=====▼	○=====▼	▼▼▼▼▼○
Heart	○=====▲	○=====▲	○=====▲	○=====▲	○=====▲	○=====▲	○=====▲	▼▼▼▼▼○
Iris	○=====	○=====	○=====	○=====	○=====	○=====	○=====	▼▼▼▼▼○
Lung Cancer	○=▲▲▲▲	○=▲▲▲▲	○=▲▲▲▲	▼▼○=====	▼▼○=====	▼▼○=====	▼▼▲▲▲○	▼▼=====
Soybean	○=====	○=====	○=====	○=====	○=====	○=====	○=====	▼▼▼▼▼○
Swissbank	○=====	○=▼▼▼▼▼	○=====	○=====	○=====	○=====	○=====	○=====
Wdbc	○▲=▲▲▲▲	▼○▼▲▲▲=	○=▲▲▲▲	▼▼○=====	▼▼○=====	▼▼○=====	▼▼=====	▼=▼▲▲▲○
Yeast	○=====▼	○=====▼	○=====▼	○=====▼	○=====▼	○=====▼	○=====▼	▲▲▲▲▲○

○: Teste ignorado, mesma técnica da coluna.
= : Nenhuma diferença estatística encontrada.
▲: Técnica da coluna apresenta melhor resultado.
▼: Técnica da coluna apresenta pior resultado.

Tabela 4.8: Resultado da avaliação do teste de hipótese – Silhueta – Grupo I

Dataset	SSF	MitraFS	FSSEM-k	FQFNN SPEC	FQFNN MI	FQFNN SPECMI	FQFNN ensemble	Completo
Breast tissue	○=▲=▲=	○=▲=▲=	○=○=	▼▼○▼=▼▼	○=▲=	▼▼=▼○▼▼	○=▲=	○=▲=
Cardiotocography	○▲▲=====	▼○▲=====	▼▼○=====	▼▼○=====	○=====	○=====	○=====	▼▼▼▼▼○
Dermatology	○▲=====	▼○=▼=▼	○=▼=▼	○=▲=	○=▼=	▲▲▲▲○▲	○=▲=	▼▼▼▼▼○
Ecoli	○=====	○=====	○=====	○=====	○=====	○=====	○=====	○=====
Glass	○=====	○=====	○=====	○=====	○=====	○=====	○=====	○=====
Heart	○=====	○=▼=	○=====	○=○=	○=▼=	○=====	○=====	▼▼▼▼▼○
Iris	○=====	○=====	○=====	○=====	○=====	○=====	○=====	▼▼▼▼▼○
Lung Cancer	○=▼=	○=▼=	○=▼=	○=○=	○=○=	○=○=	○=▼=	○=▼=
Soybean	○▲▼▲=	▼○=▼=	▼○=▼=	○=○=	○=○=	○=○=	○=○=	○=○=
Swissbank	○▲=====	▼○▼▼▼▼	○=	○=	○=	○=	○=	○=
Wdbc	○▲=▼=	▼○▲▲▲=	○=	○=	○=	○=	○=	○=
Yeast	○=====	○=====	○=====	○=====	○=====	○=====	○=====	○=====

○: Teste ignorado, mesma técnica da coluna.
= : Nenhuma diferença estatística encontrada.
▲: Técnica da coluna apresenta melhor resultado.
▼: Técnica da coluna apresenta pior resultado.

remover atributos que que dificultam o processo de agrupamento (atributos de ruído), o que leva à melhora da qualidade da partição obtida. A Tabela 4.9 apresenta a diferença entre a quantidade de atributos selecionados pela técnica FQFNN+SPEC e pelas outras técnicas avaliadas.

Para comparar os atributos selecionados, verificamos o grau de similaridade entre os sub-conjuntos selecionados pelas diferentes técnicas. A similaridade SI_{ab} é definida na Equação 4.1, em que F_x é o conjunto de atributos gerados por uma técnica x. Os resultados estão no intervalo $[0, 1]$. Quando a similaridade SI_{ab} é 0, não existe concordância entre os atributos selecionados pelas técnicas, por outro lado, um resultado igual a 1 significa que as técnicas selecionaram os mesmos atributos.

$$SI_{ab} = \frac{|F_a \cap F_b|}{|F_a \cup F_b|} \quad (4.1)$$

De acordo com os resultados mostrados na Tabela 4.10, a similaridade média da técnica FQFNN+SPEC quando comparada às técnicas estudadas é de 0.66. Isso significa que existe uma sobreposição dos atributos selecionados pelas técnicas. Apenas no conjunto de dado *Iris*, as técnicas selecionaram os mesmos atributos, ou seja, $SI = 1.0$. Por outro lado, no conjunto de dados *Lung Cancer*, foi encontrada uma baixa similaridade entre os atributos selecionados. Nesse caso, a seleção feita pela técnica FQFNN+SPEC apresentou um baixo valor de CR , ou seja, uma baixa correspondência com a partição real, o que indica que os atributos selecionados

Tabela 4.9: Diferença entre a quantidade de atributos selecionados pela técnica FQFNN+SPEC – Grupo I

Dataset	SSF	MitraFS	FSSEM-k
Breast tissue	−2	−2	−2
Cardiotocography	+2	−1	+3
Dermatology	−2	−7	−3
Ecoli	0	0	0
Glass	0	0	−1
Heart	−3	−4	−1
Iris	0	0	0
Lung Cancer	−3	−11	−2
Soybean	−2	−7	−6
Swissbank	+1	+1	+1
Wdbc	−5	−6	−5
Yeast	−2	−4	−1
+a : FQFNN+SPEC selecionou <i>a</i> atributos a mais			
−a : FQFNN+SPEC selecionou <i>a</i> atributos a menos			

não foram bons para gerar grupos próximos dos grupos verdadeiros.

4.4 Grupo II: Conjunto de dados de imagens e *microarray*

Diferente do Grupo I, os conjunto de dados do Grupo II ⁴ apresentam um grande número de dimensões. Os três primeiros são conjuntos de análise de imagem e os outros conjuntos possuem valores de expressão gênica obtidos por experimentos com *microarray*. Os detalhes desses conjuntos estão na tabela 4.11.

Nesse grupo de experimentos, a etapa de aplicação da técnica resultou em 48 subconjuntos de atributos a serem analisados (aplicadas 7 técnicas de seleção aos 6 conjuntos de dados, mais o conjunto total de atributos). Cada um desses 48 subconjuntos de atributos foi avaliado e os resultados obtidos são detalhados a seguir.

Para o conjunto de dados *AR10P*, Figura 4.11, observou-se que todas as técnicas utilizadas selecionaram em torno de 25% dos atributos. Essa seleção permitiu melhorar a média do desempenho da medida de *CR* do valor anterior, 0.02, utilizando todos os atributos, para uma média de 0.14 ao utilizar os atributos selecionados. Entretanto, não foi possível observar uma melhora no resultado da medida Silhueta, que permaneceu no mesmo patamar.

Para o conjunto de dados *PIE10P*, Figura 4.12, foi selecionado um conjunto maior de atributos, aproximadamente 40% do conjunto completo. Ao analisar esse conjunto, também foi possível verificar uma melhora no desempenho da métrica *CR*, ou seja, uma melhor correspondência com a partição real. Diferente do conjunto de dados *AR10P*, a análise da medida Silhueta mostrou uma melhora no seu valor para algumas técnicas de seleção.

⁴disponível em: <http://featureselecion.asu.edu>

Tabela 4.10: Similaridade entre os atributos selecionados pela técnica FQFNN+SPEC – Grupo I

Conjunto	SSF	MitraFS	FSSEM-k
Breast tissue	0.66	0.66	0.66
Cardiotocography	0.82	0.95	0.73
Dermatology	0.92	0.52	0.78
Ecoli	0.66	0.66	0.66
Glass	1.00	0.60	0.80
Heart	0.45	0.25	0.50
Iris	1.00	1.00	1.00
Lung Cancer	0.21	0.16	0.37
Soybean	0.81	0.47	0.60
Swissbank	0.80	0.80	0.80
Wdbc	0.61	0.57	0.61
Yeast	0.65	0.55	0.83
Média	0.71 ± 0.22	0.58 ± 0.24	0.69 ± 0.16

Tabela 4.11: Conjuntos de dados do Grupo II

Dataset	# Classes	# Atributos	# Exemplos
AR10P	10	2400	130
ORL10P	10	10304	100
PIE10P	10	2420	210
CLL-SUB-111	3	11340	111
GLI-85	2	22283	85
TOX-171	4	5748	171

Figura 4.11: Conjunto de dados AR10P

Com relação ao conjunto *CLL-SUB-111*, Figura 4.13, foram selecionadas em torno de 75% dos atributos originais. Foi possível perceber que, considerando a medida *CR*, utilizar o *MI* como algoritmo de *ranking* causa uma degradação na qualidade do resultado. As técnicas que

Figura 4.12: Conjunto de dados PIE10P

utilizam o *MI* (*SPECMI*, *MI* e o *ensemble*) tiveram uma qualidade inferior às outras técnicas. No entanto, a técnica *SPECMI* e o *ensemble* obtiveram resultados similares à utilização de todos os atributos. A análise da medida de Silhueta mostrou que o desempenho de todas as técnicas avaliadas é similar, porém, apresentam um desempenho inferior ao valor obtido quando todos os atributos são utilizados.

Figura 4.13: Conjunto de dados CLL-SUB-111

O conjunto de dados GLI-85 mostrou-se difícil ser analisado, os valores do índice CR estão próximos de 0 e apenas com a técnica SSF e a combinação FQFNN+ensemble o valor da Silhueta apresentou melhorias em comparação ao uso de todos os atributos 4.14. É importante ressaltar que a técnica MitraFS não foi utilizada para a avaliação desse conjunto de dados, devido ao alto custo computacional envolvido na análise exploratória para a definição do parâmetro k_m .

Todos os resultados para os conjuntos de dados do Grupo II estão sumarizados nas Tabelas 4.12, 4.13, 4.14, 4.15. A organização das tabelas é mesma utilizada na Seção 4.3.

É importante ressaltar que, à exceção de 1 caso, a quantidade de atributos selecionados

Figura 4.14: Conjunto de dados GLI-85

Tabela 4.12: Resultados CR – Base comparativa – Grupo II

Dataset	SSF		MitraFS		FSSEM-k		Completo	
	F	CR	F	CR	F	CR	F	CR
AR10P	571	0.15 ± 0.03	588	0.13 ± 0.03	593	0.11 ± 0.01	2400	0.02 ± 0.02
ORL10P	5939	0.57 ± 0.09	7077	0.54 ± 0.06	7088	0.55 ± 0.06	10304	0.61 ± 0.05
PIE10P	796	0.12 ± 0.01	981	0.11 ± 0.01	998	0.11 ± 0.01	2420	0.06 ± 0.01
CLL-SUB-111	7378	0.13 ± 0.03	8888	0.16 ± 0.05	8881	0.11 ± 0.07	11340	0.08 ± 0.00
GLI-85	20257	0.03 ± 0.00	12247	0.04 ± 0.01	12266	0.02 ± 0.02	22283	0.08 ± 0.03
TOX-171	5704	0.21 ± 0.01	4976	0.07 ± 0.04	4984	0.05 ± 0.04	5748	0.10 ± 0.05

ao utilizar a técnica *FQFNN* foi pelo menos igual às quantidades selecionadas pelas outras técnicas, selecionando uma menor quantidade de atributos na maioria dos casos. Além disso, ao utilizar o algoritmo SPEC como técnica de *ranking*, foi possível verificar uma melhora na qualidade da medida *CR*, em comparação ao valor obtido quando utilizados todos os atributos, em 3 conjuntos de dados (Tabela 4.16). Ao comparar com as outras técnicas, a combinação *FQFNN+SPEC* tem desempenho inferior à técnica *SSF* em 2 casos.

Quando observado o valor da medida de qualidade Silhueta, o teste de hipótese mostrou que em 3 casos a combinação *FQFNN+SPEC* teve uma perda de qualidade ao ser comparada com o uso de todos os atributos (Tabela 4.17). Na comparação entre as técnicas, não há uma, dentre as avaliadas, que possa ser definida como a melhor, visto que, na maioria dos casos, o desempenho dessas técnicas é similar.

A Tabela 4.18 apresenta a diferença na quantidade de atributos selecionados. Como base de comparação, novamente é utilizada a técnica *FQFNN+SPEC*. Essa técnica seleciona menos atributos do que as demais técnicas avaliadas.

É possível afirmar que existe uma boa relação entre os atributos selecionados pelo *FQFNN+SPEC* e pelas outras técnicas, como pode ser observando na Tabela 4.19, que apresenta a similaridade entre os atributos selecionados de acordo com a Equação 4.1. Isso se dá pelo fato de que quase

Tabela 4.13: Resultados CR – FQFNN e rankings – Grupo II

Dataset	FQFNN + SPEC		FQFNN + MI		FQFNN + SPECMI		FQFNN + ensemble	
	F	CR	F	CR	F	CR	F	CR
AR10P	458	0.15 ± 0.01	460	0.15 ± 0.01	454	0.14 ± 0.02	462	0.14 ± 0.03
ORL10P	5075	0.51 ± 0.10	5078	0.47 ± 0.12	5072	0.61 ± 0.07	5072	0.61 ± 0.07
PIE10P	791	0.12 ± 0.01	804	0.12 ± 0.01	781	0.12 ± 0.01	789	0.13 ± 0.01
CLL-SUB-111	7290	0.10 ± 0.06	7661	0.02 ± 0.08	7467	0.06 ± 0.02	7418	0.07 ± 0.01
GLI-85	8695	0.05 ± 0.04	8591	0.01 ± 0.01	8598	0.03 ± 0.03	8750	0.04 ± 0.02
TOX-171	3366	0.04 ± 0.04	3372	0.02 ± 0.01	3359	0.03 ± 0.02	3364	0.01 ± 0.01

Tabela 4.14: Resultados Silhueta – Base comparativa – Grupo II

Dataset	SSF		MitraFS		FSSEM-k		Completo	
	F	CR	F	CR	F	CR	F	CR
AR10P	571	0.11 ± 0.01	588	0.10 ± 0.02	593	0.09 ± 0.03	2400	0.13 ± 0.01
ORL10P	5939	0.13 ± 0.01	7077	0.14 ± 0.02	7088	0.14 ± 0.01	10304	0.17 ± 0.01
PIE10P	796	0.23 ± 0.04	981	0.20 ± 0.02	998	0.20 ± 0.04	2420	0.17 ± 0.02
CLL-SUB-111	7378	0.14 ± 0.02	8888	0.13 ± 0.01	8881	0.11 ± 0.03	11340	0.49 ± 0.00
GLI-85	20257	0.15 ± 0.00	12247	0.18 ± 0.04	12266	0.08 ± 0.02	22283	0.10 ± 0.01
TOX-171	5704	0.08 ± 0.01	4976	0.07 ± 0.01	4984	0.07 ± 0.00	5748	0.12 ± 0.01

todos os atributos selecionados pela técnica *FQFNN+SPEC* foram também selecionados pelas outras técnicas.

4.5 Grupo III: Conjunto de dados de Espécies Endêmicas

O conjunto de dados de espécies endêmicas é formado por registros de ocorrência de indivíduos pertencentes a diferentes espécies de plantas da região do Estado do Rio de Janeiro. Nesse conjunto existe um forte indício de endemismo, ou seja, com ocorrência restrita a essa região. A Tabela 4.20 apresenta as informações gerais sobre o conjunto de dados. Os atributos considerados no contexto deste trabalho foram as variáveis ambientais preditoras (temperatura, precipitação, altitude, evapotranspiração, etc). Além desses atributos, a base de dados contém informações sobre longitude, latitude e o nome da espécie, essa última é identificada como a classe dos exemplos.

A lista das espécies foi levantada por meio da Lista de Espécie da Flora do Brasil ⁵ e os respectivos registros de ocorrências são oriundos de diversos herbários, através da rede *speciesLink* ⁶. As variáveis ambientais são originadas de várias fontes, sendo a principal delas o *CliMond: Global Climatologies for bioclimatic modelling* ⁷. Essa base de dados foi compilada pela Dra. Marinez Ferreira de Siqueira, pesquisadora do Jardim Botânico do Rio de Janeiro, dentro do escopo do projeto Padrões de Endemismo de Plantas ocorrentes no Estado do Rio de Janeiro.

⁵Disponível em: <http://floradobrasil.jbrj.gov.br/jabot/listaBrasil/ConsultaPublicaUC/ConsultaPublicaUC.do>

⁶<http://splink.cria.org.br/>

⁷<https://www.climond.org/>

Tabela 4.15: Resultados Silhueta – FQFNN e rankings – Grupo II

Dataset	FQFNN + SPEC		FQFNN + MI		FQFNN + SPECMI		FQFNN + ensemble	
	F	CR	F	CR	F	CR	F	CR
AR10P	458	0.12 ± 0.01	460	0.11 ± 0.00	454	0.11 ± 0.01	462	0.11 ± 0.01
ORL10P	5075	0.13 ± 0.01	5078	0.11 ± 0.03	5072	0.14 ± 0.01	5072	0.14 ± 0.01
PIE10P	791	0.25 ± 0.07	804	0.25 ± 0.01	781	0.24 ± 0.03	789	0.21 ± 0.02
CLL-SUB-111	7290	0.10 ± 0.02	7661	0.10 ± 0.02	7467	0.10 ± 0.04	7418	0.10 ± 0.03
GLI-85	8695	0.06 ± 0.3	8591	0.03 ± 0.02	8698	0.04 ± 0.01	8750	0.18 ± 0.09
TOX-171	3366	0.05 ± 0.01	3372	0.04 ± 0.02	3359	0.06 ± 0.01	3364	0.07 ± 0.01

Tabela 4.16: Resultado da avaliação do teste de hipótese – CR – Grupo II

Dataset	SSF	MitraFS	FSSEM-k	FQFNN SPEC	FQFNN MI	FQFNN SPECMI	FQFNN ensemble	Completo
AR10P	○=▲=====▲	=○=====▲	▼=○▼▼=▲	=▲○=====▲	=▲=○=▲=▲	=====○=▲	=====○▲	▼▼▼▼▼▼○
ORL10P	○=====▲▼▼▼	=○=====▲▼▼▼	=○=▲▼▼▼	=○=○=▼▼▼	▼▼▼=○▼▼▼	▲▲▲▲○=	▲▲▲▲=○=	▲▲▲▲=○=
PIE10P	○=====▲	=○=====▲	=○=○=▲	=○=○=▲	=○=○=▲	=====○=▲	=====○▲	▼▼▼▼▼▼○
CLL-SUB-111	○=====▲▲▲▲	=○=====▲▲▲▲	=○=○=▲	=○=○=▲	▼▼▼▼○▼▼▼	▼▼=▲=○=	▼▼=▲=○=	▼▼=▲=○=
GLI-85	○=▼▲=====▼	=○=▼=====▼	=○=▼=====▼	▲=○▲=▼	▼=○=○=▼	=====○=	=-=====○▼	▲-▲▲▲▲○
TOX-171	○▲▲▲▲▲▲	▼○=▲=▲=	▼=○=▲=▲=	▼=○=○=▼	▼=○=○=▼	▼▼▼=○=▼	▼▼▼=○=▼	▼=▲▲▲▲○

○: Teste ignorado, mesma técnica da coluna.
 =: Nenhuma diferença estatística encontrada.
 ▲: Técnica da coluna apresenta melhor resultado.
 ▼: Técnica da coluna apresenta pior resultado.

As metas que levaram ao desenvolvimento desse conjunto de dados são:

1. Verificar o padrão de distribuição geográfica dessas espécies endêmicas na região da Baía do Atlântico Sudeste (onde está localizado o estado do Rio de Janeiro);
2. Estabelecer quais variáveis ambientais explicam melhor essa distribuição;
3. Levantar hipóteses para explicar esse padrão de distribuição.

O principal desafio desse conjunto de dados é que, por se tratarem de espécies endêmicas, uma parte considerável dessas espécies tem distribuição geográfica bastante restrita, sendo comum espécies representadas por um ou poucos registros de ocorrência. Essa amostragem extremamente baixa coloca um desafio grande para se analisar esses dados e encontrar resultados satisfatórios. Por outro lado, espécies endêmicas e/ou raras tem um apelo conservacionista muito grande, o que torna este conjunto de dados biologicamente e ecologicamente muito interessante.

No contexto desta Tese, esse conjunto de dados é avaliado para verificar quais variáveis ambientais explicam o padrão de distribuição geográfica das espécies endêmicas. Esse é um problema de seleção de atributos e a base de dados considerada não é rotulada, assim sua análise se dá por técnicas não supervisionadas. Desse modo, aplicamos as técnicas previamente descritas e os resultados da quantidade de atributos selecionados são mostrados na Tabela 4.22.

Como o objetivo da análise realizada nesse conjunto de dados é a seleção de atributos que auxiliem na identificação de espécies endêmicas, o algoritmo *K-means* foi executado com o valor de $k = 2$, de modo a agrupar os exemplos como sendo espécie endêmica ou não-endêmica.

Tabela 4.17: Resultado da avaliação do teste de hipótese – Silhueta – Grupo II

Dataset	SSF	MitraFS	FSSEM-k	FQFNN SPEC	FQFNN MI	FQFNN SPECMI	FQFNN ensemble	Completo
AR10P	o=====▼	=o=====▼	==o=====▼	====o=====	====o=====▼	====o=====▼	====o=====▼	▲▲▲=▲▲▲o
ORL10P	o=====▼	=o=====▼	==o=====▼	====o=====▼	====o=====▼	====o=====▼	====o=====▼	▲▲▲▲▲▲o
PIE10P	o=====▲	=o=====▲	==o=====▲	====o=====▲	====o=====▲	====o=====▲	====o=====▲	▼▼▼▼▼o
CLL-SUB-111	o=▲▲▲▲▼	=o=▲▲▲▲▼	▼=o=====▼	▼▼=o=====▼	▼▼=o=====▼	▼▼=o=====▼	▼▼=o=====▼	▲▲▲▲▲o
GLI-85	o=▲▲▲▲▼	=o=▲▲▲▲▼	▼=o=====▼	▼▼=o=====▼	▼▼=o=====▼	▼▼=o=====▼	▼▼=o=====▼	▼=▲▲▲▲o
TOX-171	o=▲▲▲=▼	=o=▲▲▲=▼	==o=====▼	▼▼▼o=====▼	▼▼▼o=====▼	▼▼▼o=====▼	▼▼▼o=====▼	▲▲▲▲▲o

o: Teste ignorado, mesma técnica da coluna.
 =: Nenhuma diferença estatística encontrada.
 ▲: Técnica da coluna apresenta melhor resultado.
 ▼: Técnica da coluna apresenta pior resultado.

Tabela 4.18: Diferença entre a quantidade de atributos selecionados com FQFNN+SPEC – Grupo II

Dataset	SSF	MitraFS	FSSEM-k
AR10P	−113	−130	−135
ORL10P	−864	−2002	−2013
PIE10P	−5	−190	−207
CLL-SUB-111	−88	−1598	−1591
GLI-85	−11562	—	−3571
TOX-171	−2338	−1610	−1618

+a : FQFNN+SPEC selecionou a atributos a mais
 −a : FQFNN+SPEC selecionou a atributos a menos

Entretanto, para esse valor de k , foi obtida uma distribuição dos exemplos de aproximadamente 60/40, o que não condiz com a proposta do conjunto de dados. Isso se deve ao fato de as espécies endêmicas se apresentarem como *outliers* e aparecerem em poucas ocorrências. Assim, o algoritmo *K-means* agrupa espécies endêmicas com outras espécies. A Tabela ?? apresenta os resultados obtidos da avaliação da silhueta considerando os atributos selecionados considerando $k = 2$. Se for considerado o valor de k como igual ao número de classes, a avaliação da qualidade se dará em relação à partição real, o que não é o objetivo do conjunto de dados.

Nos experimentos para seleção de atributos, as técnicas SSF, MitraFS e FSSEM-k selecionam poucos atributos, pois, esses atributos são suficientes para agrupar os exemplos. No entanto, ao utilizar a técnica FQFNN para definir a quantidade de atributos (com base em um ranking), atributos são adicionados até que a distância entre exemplos não apresente variação, isto é, o percentual de falsos vizinhos seja igual a zero.

Para esse conjunto de dados, a seleção de poucas características não é adequada. Isso se deve ao fato de que espécies endêmicas podem se assemelhar a outras espécies, tendo poucos atributos que apresentam variações.

A definição dos atributos importantes para a identificação de espécies endêmicas é um problema complexo e ainda está em estudo. Selecionar poucos atributos pode significar a não identificação de uma espécie como endêmica. Por outro lado, a utilização de muitos atributos aumenta a carga de trabalho dos profissionais que coletam informações sobre as espécies, bem como dificulta a identificação de uma espécie como sendo endêmica.

Tabela 4.19: Similaridade entre os atributos selecionados com FQFNN+SPEC – Grupo II

Conjunto	SSF	MitraFS	FSSEM-k
AR10P	0.77	0.74	0.72
ORL10P	0.85	0.71	0.71
PIE10P	0.98	0.80	0.79
CLL-SUB-111	0.98	0.82	0.82
GLI-85	0.42	—	0.70
TOX-171	0.58	0.65	0.67
Média	0.76 ± 0.22	0.73 ± 0.06	0.73 ± 0.05

Tabela 4.20: Conjuntos de dados do Grupo III

Dataset	# Classes	# Atributos	# Exemplos
Espécies Endêmicas RJ	609	45	2402

4.6 Considerações Finais

Nesse capítulo foram apresentados os experimentos realizados nesta Tese. Esses experimentos permitiram a comparação de técnicas existentes na literatura com a combinação da técnica FQFNN e 4 algoritmos de *ranking*. O objetivo da técnica proposta é a definição da quantidade de atributos a serem selecionados, tendo como base um *ranking* de atributos. Assim, verificou-se que a quantidade de atributos sugeridos pela técnica FQFNN é compatível com as outras técnicas da literatura.

Considerando-se conjuntos de dados com poucos atributos (Grupo I, Seção 4.3), houve pouca variação na quantidade de atributos selecionados quando são comparadas as indicações da técnica FQFNN e as outras técnicas da literatura. Ao considerar a qualidade da partição gerada pelas técnicas, na maioria dos casos, o resultado das medidas de qualidade mostrou que a técnica FQFNN e os algoritmos de *ranking* tem desempenho semelhante às outras técnicas.

Os resultados obtidos com os conjuntos de dados do Grupo II (Seção 4.4, conjuntos de dados de alta dimensionalidade) mostraram que a quantidade de atributos indicados pela técnica FQFNN foi menor do que a quantidade selecionada pelas outras técnicas. Considerando a qualidade que foi medida utilizando as técnicas Silhueta e CR, os resultados são compatíveis quando comparados às outras técnicas.

Ao analisar os resultados do Grupo III (Seção 4.5), observou-se que a quantidade de atributos indicados pela técnica FQFNN apresentou melhores resultados do que as outras técnicas. Para o conjunto de dados de espécies endêmicas do Rio de Janeiro, o melhor desempenho da técnica FQFNN pode ser explicado pelo modo como a técnica proposta indica a quantidade de atributos a serem utilizados.

Considerando os resultados dos três grupos, é possível afirmar que a técnica FQFNN pode ser utilizada para indicar a quantidade de atributos a serem utilizados, tendo como base um

Tabela 4.21: Resultados para o conjunto de dados de espécies endêmicas

Técnica	# Selecionados
SSF	7
MitraFS	3
FSSEM-k	3
FQFNN+SPEC	26
FQFNN+MI	28
FQFNN+SPECMI	29
FQFNN+ensemble	23

Tabela 4.22: Avaliação de Silhueta para o conjunto de dados de espécies endêmicas – 2 grupos

Técnica	Silhueta
SSF	0.19 ± 0.02
MitraFS	0.18 ± 0.01
FSSEM-k	0.21 ± 0.05
FQFNN+SPEC	0.22 ± 0.04
FQFNN+MI	0.22 ± 0.05
FQFNN+SPECMI	0.20 ± 0.06
FQFNN+ensemble	0.21 ± 0.01

ranking de atributos.

Conclusão

5.1 Considerações

Com o crescente aumento na quantidade de dados disponíveis em diversas áreas, a identificação de quais dados são úteis e quais não agregam informação se torna cada vez mais importante. Áreas como análise de expressão gênica por meio de *microarrays* e análise de imagens são bons exemplos desse problema. Conjuntos de dados dessas áreas têm como característica uma grande quantidade de informação, na forma de atributos, sobre poucas observações (exemplos). Nesses conjuntos de dados, a identificação de atributos irrelevantes e de ruído é importante, pois ao reduzir a quantidade de atributos considerados, o efeito da maldição da dimensionalidade é reduzido. Além disso, a redução da quantidade de atributos proporciona uma melhora no desempenho da técnica de aprendizado (agrupamento de dados, no contexto desta Tese) e uma redução no tempo necessário para a aplicação da técnica. Esses benefícios também são importantes em conjuntos de dados menores.

A seleção de atributos no contexto não-supervisionado apresenta um desafio maior. Por não existir informação sobre o atributo alvo (rótulo de classes), não existe um guia claro para definir quais são os atributos mais importantes. Diversas técnicas foram apresentadas na literatura para tratar do problema da seleção de atributos não-supervisionados. O que se pode observar é que, em vários casos, as técnicas apresentam como resultado um *ranking* e selecionam a quantidade de atributos a utilizar por meio de um limiar. Outras, utilizam o resultado de um algoritmo de agrupamento para definir quantos atributos devem ser selecionados. Poucas técnicas verificam a relação entre os atributos para então definir quais devem ser selecionados.

A proposta desta Tese é uma técnica que, ao avaliar a variação da distância entre exemplos quando consideradas quantidades diferentes de atributos, define um ponto em que essa distância não apresenta alteração significativa. Assim, ao comparar a distância relativa considerando m

e $m + 1$ atributos, a distância não apresenta variação significativa, ou seja, pouca informação é adicionada ao considerar o atributo adicionado. A técnica proposta recebe o nome de FQFNN, *Feature Quantification by False Nearest Neighbors*, e tem como base técnicas de teoria do caos.

A premissa básica de comparar quantidades diferentes de atributos leva a um problema que é exponencial, pois, todos os possíveis arranjos devem ser considerados para definir quantos atributos devem ser selecionados. No entanto, a premissa básica também permite dizer quais os atributos devem ser considerados. Essa análise será investigada em trabalhos futuros. Para minimizar esse problema, nesta Tese é considerada a ordem de relevância dos atributos, ou seja, um *ranking* de atributos. Ao utilizar um *ranking*, supõe-se que a quantidade de informação relevante diminui à medida que mais atributos são adicionados. São consideradas 4 medidas para a definição de *rankings*, são elas: SPEC, Mutual Information, SPECMI e um ensemble. A qualidade do *ranking* utilizado como entrada para a técnica FQFNN é importante para o resultado final.

Nos experimentos realizados foi possível verificar que a técnica proposta indica uma quantidade de atributos menor do que as outras técnicas, na maioria dos casos. Ao considerar os conjuntos de dados do Grupo I (conjuntos de dados pequenos), observou-se que a quantidade de atributos selecionados apresenta pouca diferença quando comparadas às outras técnicas, enquanto que a qualidade medida por meio das medidas CR e Silhueta se manteve semelhante. No entanto, em nenhum caso, a técnica FQFNN apresentou melhores resultados, o que pode ser explicado por: 1) ao selecionar menos atributos, menos informações sobre os dados estão disponíveis, diminuindo, assim, a qualidade da partição gerada (*no free lunch*); 2) o *ranking* utilizado pelo FQFNN não estava bem formado. Quando analisados os conjuntos do Grupo II (conjuntos de dados com muitos atributos), verifica-se um resultado semelhante, a técnica FQFNN indica menos atributos do que as técnicas comparadas e os resultados das medidas de qualidade apresentam resultados semelhantes às outras técnicas na maioria dos conjuntos. Novamente, no Grupo II, ao utilizar o FQFNN, em nenhum caso um resultado melhor do que as outras técnicas foi obtido.

Por último, considerando o problema de identificação de espécies endêmicas no Rio de Janeiro (Grupo III). As técnicas existentes selecionaram poucos atributos que, de acordo com especialistas, não apresentam informações suficientes para distinguir espécies endêmicas e não-endêmicas. Pode-se observar que a análise das distâncias relativas apresentou resultados mais significativos.

Portanto, considerando os três grupos de experimentos, é possível concluir que a técnica FQFNN é útil para a identificação da quantidade de atributos a ser considerados. Essa técnica é utilizada em conjunto com um método que forneça um *ranking* de atributos.

5.2 Contribuições

As principais contribuições desta Tese estão listadas a seguir:

- Análise da variação de distância entre exemplos ao considerar quantidades diferentes de atributos;
- Proposta de uma técnica que, a partir da variação na distância do vizinho mais próximo, indica uma quantidade de atributos a serem utilizados;
- Implementação da técnica de *ranking* SPECMI, que adiciona a análise de similaridade entre atributos, por meio da técnica de *Mutual Information*, à técnica SPEC;
- Análise do conjunto de dados reais de espécies endêmicas do Rio de Janeiro.

5.3 Publicações

Durante o desenvolvimento deste trabalho, as seguintes publicações foram geradas:

- *Using False Nearest Neighbors to Quantify Features in Unsupervised Problems*. José A. Andrade Filho, André C. P. L. F. de Carvalho, Huan Liu. Aceito para publicação na revista *Intelligent Data Analysis*, 2013;
- *Quantifying Features Using False Nearest Neighbors: An Unsupervised Approach*. José A. Andrade Filho, Andre C. P. L. F. de Carvalho, Rodrigo F. Mello, Salem Alelyani, Huan Liu. Publicado em: 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, Florida, USA.

Além dessas publicações, outras estão em processo de submissão. A primeira, a ser submetida a um periódico, contempla a comparação entre os diversos tipos de *ranking* utilizados em conjunto com o FQFNN e o detalhamento da importância da etapa de *ranking* para o FQFNN. A segunda, ainda sem detalhes de submissão, introduz a técnica de *ranking* SPECMI. Essa técnica ajusta o *ranking* gerado pela técnica SPEC (Zhao & Liu, 2012), de acordo com a similaridade entre atributos baseando-se na análise de informação mútua.

5.4 Trabalhos futuros

A técnica FQFNN, proposta nesta tese, tem como objetivo definir a quantidade de atributos a serem utilizados. Durante a realização da pesquisa, foram verificados dois aspectos da técnica FQFNN que podem ser investigados futuramente. O primeiro diz respeito à análise da influência de agrupamentos de sub-espaco no resultado do FQFNN. A existência desses agrupamentos pode refletir em variações na fração de falsos vizinhos, que é calculada pela técnica FQFNN.

O segundo aspecto refere-se à capacidade de utilizar o algoritmo de falsos vizinhos mais próximos para selecionar os atributos, sem a necessidade de algoritmos auxiliares. Para tanto, é necessário investigar abordagens que levem em consideração a natureza exponencial do conjunto de soluções possíveis, como uma abordagem gulosa.

Por último, outro trabalho que deve ser continuado é a colaboração para a análise do conjunto de dados de espécies endêmicas. A utilização do conhecimento de domínio pode melhorar a qualidade dos resultados obtidos bem como reduzir a carga de trabalho dos pesquisadores que coletam as informações em campo. Além disso, as variáveis de localização (latitude e longitude) não foram utilizadas neste trabalho. No entanto, a implementação de uma técnica que leve essas variáveis em consideração pode auxiliar na identificação das espécies endêmicas.

Referências

- Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. (2001). *On the surprising behavior of distance metrics in high dimensional space*. Springer.
- Alligood, K. T.; Sauer, T. D.; Yorke, J. A. (1997). *Chaos: An Introduction to Dynamical Systems*. Springer.
- Andrade Filho, J. A.; Carvalho, A. C. P. L. F.; de Mello, R. F.; Alelyani, S.; Liu, H. (2011). Quantifying features using false nearest neighbors: An unsupervised approach. *Proceedings of the IEEE 23rd International Conference on Tools with Artificial Intelligence, ICTAI 2011, Boca Raton, FL, USA, November 7-9, 2011*, p. 994–997.
- Asuncion, A.; Newman, D. (2007). UCI machine learning repository.
- Au, W.-H.; Chan, K.; Wong, A. K. C.; Wang, Y. (2005). Attribute clustering for grouping, selection, and classification of gene expression data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, v.2, n.2, p.83–101.
- Barbara, D. (2000). An introduction to cluster analysis for data mining – available at: www.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.pdf. access : January, 2010.
- Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer US.
- Boutsidis, C.; Mahoney, M. W.; Drineas, P. (2009). Unsupervised feature selection for the k-means clustering problem. *Advances in Neural Information Processing Systems (NIPS)*.
- Covões, T.; Hruschka, E.; Castro, L.; Santos, Á. (2009). A cluster-based feature selection approach. Corchado, E.; Wu, X.; Oja, E.; Herrero, Á.; Baroque, B., editores, *Hybrid Artificial Intelligence Systems*, v. 5572 de *Lecture Notes in Computer Science*, p. 169–176. Springer Berlin Heidelberg.
- Covões, T. F.; Hruschka, E. R. (2011). Towards improving cluster-based feature selection with a simplified silhouette filter. *Information Sciences*, v.181, n.18, p.3766 – 3782.

-
- Dash, M.; Choi, K.; Scheuermann, P.; Liu, H. (2002). Feature selection for clustering - a filter solution. *Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on*, p. 115–122.
- Dash, M.; Liu, H. (2000). Feature selection for clustering. *Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference*.
- Demsăr, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, v.7, p.1–30.
- Devaney, M.; Ram, A. (1997). Efficient feature selection in conceptual clustering. *International Conference on Machine Learning*, p. 92–97.
- Devijver, P. A.; Kittler, J. (1982). *Pattern recognition: A statistical approach*. Prentice Hall.
- Dy, J. G.; Brodley, C. E. (2000a). Feature subset selection and order identification for unsupervised learning. *Proceedings of the 17th International Conference on Machine Learning*.
- Dy, J. G.; Brodley, C. E. (2000b). Visualization and interactive feature selection for unsupervised data. *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 360–364, New York, NY, USA. ACM.
- Dy, J. G.; Brodley, C. E. (2004). Feature selection for unsupervised learning. *J. Mach. Learn. Res.*, v.5, p.845–889.
- Faceli, K. (2006). *Um framework para análise de agrupamento baseado na combinação multi-objetivo de algoritmos de agrupamento*. Tese (Doutorado), ICMC/USP.
- Faceli, K.; Lorena, A. C.; Gama, J.; de Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. LTC.
- Fraser, A. M.; Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, v.33, n.2, p.1134–1140.
- Guyon, I.; Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, v.3, p.1157–1182.
- Hruschka, E.; de Castro, L.; Campello, R. J. G. B. (2004). Evolutionary algorithms for clustering gene-expression data. *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, p. 403–406.
- Hruschka, E. R.; Campello, R. J.; de Castro, L. N. (2006). Evolving clusters in gene-expression data. *Information Sciences*, v.176, n.13, p.1898 – 1927.
- Hubert, L.; Arabie, P. (1985). Comparing partitions. *Journal of Classification*, p. 193–218.

- Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, v.31, n.8, p.651 – 666.
- Jain, A. K.; Dubes (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- Jain, A. K.; Murty, M. N.; Flynn, P. J. (1999). Data clustering: A review. *ACM Comput. Surv.*, v.31, n.3, p.264–323.
- Jiang, D.; Tang, C.; Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *Knowledge and Data Engineering, IEEE Transactions on*, v.16, n.11, p.1370–1386.
- Kaufman, L.; Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 9th edição.
- Kennel, M. (2002). The multiple-dimensions mutual information program – available at: <http://www-ncsl.postech.ac.kr/en/software/archives/mmi.tar.z>. access: December, 2009.
- Kennel, M.; Brown, R.; Abarbanel, H. (1992a). Determining embedding dimension for phase space reconstruction using the method of false nearest neighbors. Relatório técnico, Institute for Nonlinear Science and Department of Physics, University of California, San Diego, Mail Code R-002, La Jolla, CA, EUA.
- Kennel, M. B.; Brown, R.; Abarbanel, H. D. I. (1992b). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, v.45, n.6, p.3403–3411.
- King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, v.62, n.317, p.pp. 86–101.
- Kononenko, I.; Kukar, M. (2007). *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited.
- Li, Y.; Dong, M.; Hua, J. (2007). Localized feature selection for clustering and its application in image grouping. *Multimedia and Expo, 2007 IEEE International Conference on*, p. 651–654.
- Li, Y.; Lu, B.-L.; Wu, Z.-F. (2006). A hybrid method of unsupervised feature selection based on ranking. *icpr*, v.2, p.687–690.
- Liu, H.; Motoda, H. (2007). *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC.
- Liu, H.; Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, v.17, n.4, p.491–502.

-
- Lorenz, E. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Science*, v.20, p.130–141.
- Mañé, R. (1980). *On the dimension of the compact invariant sets of certain nonlinear maps*. Springer.
- Medio, A.; Gallo, G. (1993). *Chaotic Dynamics: Theory and Applications to Economics*. Cambridge University Press.
- Mello, R. F. (2009). *Sistemas Dinâmicos e Técnicas Inteligentes para a Predição de Comportamento de Processos: Uma Abordagem para Otimização de Escalonamento em Grades Computacionais*. Tese (Doutorado), ICMC/USP.
- Mirkin, B. (1999). Concept Learning and Feature Selection Based on Square-Error Clustering. *Machine Learning*, v.35, n.1, p.25–39.
- Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.
- Mitra, P.; Murthy, C.; Pal, S. (2002). Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v.24, n.3, p.301–312.
- Palis Jr., J.; Melo, W. (1978). *Introdução aos Sistemas Dinâmicos*. Edgard Blücher, 1 edição.
- Pudil, P.; Novovicova, J.; Kittler, J. (1994). Floating search methods in feature-selection. *PRL*, v.15, n.11, p.1119–1125.
- Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, v.20, n.1, p.53–65.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, p. 379–423.
- Simovici, D. A. (2007). Data mining algorithms I: Clustering. *Handbook of Applied Algorithms – Capítulo 6*, p. 177–218.
- Takens, F. (1980). Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence*, p. 366–381. Springer.
- Whitney, H. (1936). Differentiable manifolds. *The Annals of Mathematics*, v.37, n.3, p.645–680.
- Zhao, Z.; Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. *Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, p. 1151–1157.
- Zhao, Z. A.; Liu, H. (2012). *Spectral Feature Selection for Data Mining*. CRC Press.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edição.

Zongker, D.; Jain, A. (1996). Algorithms for feature selection: An evaluation. *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, v.2, p.18–22 vol.2.