Universidade Federal do Rio Grande - FURG
Centro de Ciências Computacionais - C3
Programa de Pós-Graduação em Computação
Mestrado em Engenharia de Computação

Caroline Tomasini

Seleção automática de índices internos de validação de agrupamento.

Rio Grande/RS
26 de agosto de 2015

Caroline Tomasini

Seleção automática de índices internos de validação de agrupamento.

Dissertacão de Mestrado apresentada ao Programa de Pós-Graduação em Computaçãoda Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtencão do grau de Mestre em Engenheira de Computação.

Orientador(a): Prof. Dra. Karina dos Santos Machado

Co-orientador(a): Prof. Dr. Eduardo Nunes Borges

Rio Grande/RS

26 de agosto de 2015

Ficha catalográfica

T655s Tomasini, Caroline.
Seleção automática de índices internos de validação de agrupamento / Caroline Tomasini.– 2015.
69 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-graduação em Engenharia de Computação, Rio Grande/RS, 2015.
Orientadora: Drª. Karina dos Santos Machado.
Coorientador: Dr. Eduardo Nunes Borges.

1. Agrupamento de dados 2. Validação 3. Regressão linear 4. Aprendizado de máquina I. Machado, Karina dos Santos II. Borges, Eduardo Nunes III. Título.

CDU 004.42

Catalogação na Fonte: Bibliotecário Me. João Paulo Borges da Silveira CRB 10/2130

Caroline Tomasini

Seleção automática de índices internos de validação de agrupamento.

Dissertacão de Mestrado apresentada ao Programa de Pós-Graduação em Computaçãoda Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtencão do grau de Mestre em Engenheira de Computação.

Aprovado em: 14 / 05 / 2015

BANCA EXAMINADORA

Prof. Dra. Karina dos Santos Machado (Orientadora)
Centro de Ciências Computacionais – FURG

Prof. Dra. Ana Trindade Winck
Universidade Federal de Santa Maria – UFSM

Prof. Dr. Adriano Velasque Werhli
Centro de Ciências Computacionais – FURG

Prof. Dra. Cleo Zanella Billa
Centro de Ciências Computacionais – FURG

Dedico mais esta conquista a minha família: Aos meus pais Ronaldo e Carina Tomasini; Ao meu noivo Thyago Salvá; A minha avó, tios e primos; E as estrelas que guiam meu caminho, avôs, vó e tio Edu (in memorian).

Agradecimentos

Inicialmente agradeço aos meus pais Ronaldo e Carina pelo incentivo, amor, carinho, compreensão nos momentos de ausência e pelo apoio incondicional em todas as minhas escolhas e decisões. E a toda minha família que sempre esteve presente me apoiando e incentivando.

Ao meu noivo Thyago Salvá pelo amor, paciência, por ter permanecido incansavelmente ao meu lado, incentivando, compartilhando angústias e dúvidas. E a sua família por ter me recebido desde o início como uma filha.

Aos professores Leonardo Ramos Emmendorfer e Adriano Velasque Werhli pelas ideias, sugestões e discussões. Assim como aos demais professores do Centro de Ciências Computacionais que de alguma forma contribuíram para a minha formação.

Agradecimentos especiais aos meus orientadores, Professora Karina dos Santos Machado e Professor Eduardo Nunes Borges pela paciência, ideias, discussões, conhecimento transmitido e pela confiança depositada em mim, fundamental para o desenvolvimento deste trabalho, além de Mestres são amigos.

E aos demais que de alguma forma contribuíram na elaboração desta monografia.

"When face to face with all our fears
Learned our lessons through the tears
Made memories we knew would never fade"

AVICII

Resumo

A validação dos resultados de agrupamento é uma questão importante na área de aprendizado de máquina e é essencial para o sucesso das aplicações relacionadas a agrupamento de dados. No entanto, escolher o índice de validação adequado para avaliar os resultados de um algoritmo de agrupamento específico continua sendo um desafio. A qualidade das partições geradas por diferentes algoritmos de agrupamento pode ser avaliada utilizando diferentes índices com base em critérios externos ou internos. Um critério externo requer que o particionamento ideal seja conhecido *a priori* para a comparação com os resultados de agrupamento. Já o critério interno avalia os resultados de agrupamento considerando apenas as propriedades do conjunto de dados. Neste trabalho, é proposta uma metodologia para a escolha do índice interno de validação de agrupamento mais adequado, relacionando critérios externos e internos através de um modelo de regressão linear aplicado sobre os resultados de algoritmos de agrupamento particionais e baseados em densidade. Cada algoritmo foi aplicado sobre conjuntos de dados sintéticos que foram gerados para este fim, usando diferentes configurações. Os resultados de agrupamento foram avaliados por diferentes índices com base em critérios internos e externos que geraram a entrada para os modelos de regressão. A análise destes modelos permitiu a inferência do índice interno mais adequado para cada método de algoritmo de agrupamento. Por fim, foi realizada uma validação dos modelos encontrados utilizando conjuntos de dados reais e sintéticos utilizados em outros trabalhos da literatura.

Palavras-chave: Avaliação de Agrupamentos, Critérios de Validação, Regressão Linear

Abstract

Validation of clustering results is an important issue in the context of machine learning research and it is essential for the success of clustering applications. Choosing the appropriate validation index for evaluating the results of a particular clustering algorithm remains a challenge. The quality of partitions generated by different clustering algorithms can be evaluated using different indices based on external or internal criteria. An external criterion requires a partitioning of the data defined a priori for comparison with the clustering results while an internal criterion evaluates clustering results considering only the data properties. In this paper, we have proposed a methodology for selecting the most suitable cluster validation internal index, relating external and internal criteria through a linear regression model applied on the results of partitioning and density-based clustering algorithms. Each algorithm was run over synthetic datasets generated for this purpose, using different configurations. Clustering results were evaluated by different indices based on internal and external criteria generating the input for regression models. The analysis of these models allowed the inference of the most suitable internal index for each method of clustering algorithm. Finally was performed a validation of the found models using real datasets.

Keywords: Cluster Evaluation, Validation Criteria, Linear Regression.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

C	*C-index*
DBI	*Davies Bouldin Index*
D	*Dunn Index*
FM	*Fowlkes-Mallows Index*
Γ	*Gamma Index*
J	*Jaccard Index*
LM	*Linear Model*
S	*Silhouette Index*
R	*Rand Index*
TDIDT	*Top Down Induction Decision Trees*

Sumário

Capítulo 1

Introdução

Agrupamento é uma tarefa de mineração de dados não supervisionada baseada na similaridade entre as instâncias [TSK+06]. Segundo Chen *et al.* [CHY96], o processo de agrupar instâncias físicas ou abstratas em classes de instâncias similares é chamado agrupamento (*clustering*). Um grupo (*cluster*) é um subconjunto de instâncias que podem ser tratadas coletivamente [HKP06]. Um algoritmo de agrupamento tem como objetivo maximizar a similaridade intra grupo e minimizar a similaridade entre instâncias inter grupo. Hoje em dia, o agrupamento de dados é amplamente usado em diversas aplicações científicas ou organizacionais, tais como análise de dados complexos, pesquisa de mercado, processamento de imagem, teste de hipóteses, recuperação de informação, biologia, mineração de texto, marketing e descoberta de perfis [XW09].

Vários algoritmos de agrupamento foram propostos nas últimas décadas [XW+05, Ber06]. Eles podem ser classificados de acordo com o método utilizado para agrupar as instâncias. Em geral, segundo Han e Kamber [HKP06], os métodos podem ser classificados em: particionais, baseados em densidade, hierárquicos e baseados em grade.

O método particional é baseado em centroide e busca encontrar a melhor partição das n instâncias em k grupos. Esta característica exige que o usuário defina *a priori* o número de grupos (k). O algoritmo de agrupamento mais comum que utiliza esse método é o *k-means* [HW79] que não é adequado para a descoberta de agrupamentos com formas não convexas ou agrupamentos que resultem em grupos muito diferentes. Outros exemplos de métodos particionais são *k-medoids* [KR87] e CLARANS [NH02].

O método baseado em densidade permite a identificação de grupos de formatos ar-

bitrários. Mais especificamente, esses métodos classificam como grupos as regiões onde há o maior número de elementos (instâncias) no espaço de dados que são, naturalmente, separados pelas áreas de baixa densidade, conhecidas como ruídos [HKP06]. Entre os algoritmos de agrupamento baseados em densidade destaca-se o DBSCAN [EKSX96]. Este algoritmo requer dois parâmetros atribuídos pelo usuário: uma vizinhança que é definida pelo raio ε e o número mínimo de pontos $MinPoints$ nesta vizinhança.

No método hierárquico as instâncias são decompostas na forma de árvore (dendograma), dividindo-a recursivamente em conjuntos menores de instâncias. Esta divisão pode ser feita de maneira aglomerativa (*bottom-up*) onde, cada instância inicia em um grupo, para depois serem unidos até que todos estejam em um único grupo ou que uma determinada condição seja satisfeita. Ou então, pode ser feita de maneira divisiva (*top-dow*) que ao contrário da aglomerativa, as instâncias iniciam todas em um único grupo e a cada iteração do algoritmo o grupo é divido em um grupo menor, até que cada instância esteja em um único grupo ou uma condição seja satisfeita. DIANA [KR09] e ROCK [GRS99] são exemplos de algoritmos hierárquicos [HKP06].

O método baseado em grade quantifica o espaço das instâncias em um número finito de células que formam uma estrutura de grade em que todas as operações de agrupamento são realizadas. Um resumo dos vários outros algoritmos de agrupamento e suas características podem ser encontrados nos seguintes *surveys* [XW+05, Ber06].

A validação dos resultados é essencial para o sucesso de aplicações de agrupamento [HBV01]. A qualidade das partições geradas por diferentes algoritmos pode ser avaliada utilizando inspeção visual e diferentes índices baseados em critérios internos ou externos. Devido a alta dimensionalidade e cardinalidade dos conjuntos de dados, geralmente a inspeção visual se torna impraticável. Critérios externos exigem conhecimento prévio das classes de dados para comparação com o particionamento resultante após a aplicação de um algoritmo [XW09]. No entanto, a grande maioria dos problemas que exigem a utilização de técnicas de agrupamento não tem os dados rotulados *a priori*. Portanto, uma forma usual de avaliar os resultados de agrupamento é utilizar índices baseados em critérios internos, que consideram apenas as propriedades dos dados, procurando agrupamentos com grupos compactos e bem separados [Fac11].

Diversos índices de validação de agrupamento têm sido propostos na literatura [XW09,

Ran71, FM83, DB79, Dun74, Rou87, BH75, VCH10, HL76]. Cada índice concentra-se em uma propriedade particular das partições, e muitos deles são influenciados pelo impacto do ruído, pela variação da densidade, ou pela presença de subgrupos. Não é possível apontar um índice universalmente mais confiável [LLX+10, VCH10]. Por este motivo, permanece o desafio de escolher os índices apropriados para avaliar os resultados de um algoritmo de agrupamento em especial [LLX+10].

Sendo assim, para ajudar neste processo de selecionar o índice interno de validação de agrupamento mais adequado, é proposta uma metodologia baseada na indução de modelos estatísticos, que mostra a relação entre os índices internos e externos. Foram adotados modelos de regressão, onde cada índice externo é definido como o atributo alvo a ser previsto, e os índices internos são os atributos da regressão. Cada algoritmo foi executado sobre conjuntos de dados sintéticos gerados para este fim, usando diferentes configurações. Os resultados do agrupamento foram avaliados por diferentes índices internos e externos gerando a entrada para os modelos de regressão. Os resultados dos experimentos mostram que alguns índices internos têm uma relação direta com índices externos e eles quantificam essa relação, como mostram os resultados da regressão. A análise dos modelos de regressão permitiu a inferência dos índices internos mais adequados para cada algoritmo de agrupamento.

1.1 Motivação

O principal problema da maioria das abordagens de agrupamento é que a partir de um conjunto de dados elas podem produzir diversos agrupamentos diferentes [zeng2002adaptive]. Com base nesta afirmativa surgem algumas questões tais como: Qual resultado é o melhor? Quanto se pode confiar no resultado gerado? Existe um resultado melhor do que o resultado encontrado? É possível combinar todos os resultados disponíveis para um melhor entendimento dos dados?

Além disso, há vários índices de validação de agrupamento e definir qual índice é mais apropriado para avaliar qual tipo de agrupamento não é uma tarefa trivial. Dessa forma, nesta dissertação de mestrado é abordada uma metodologia para investigar as questões supracitadas.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral desta dissertação é investigar a relação entre critérios de validação de agrupamentos externos e internos. Desta maneira, é proposta uma metodologia baseada na indução de modelos estatísticos que mostram esta relação.

1.2.2 Objetivos Específicos

- Gerar *datasets* com característica de densidade e compacidade;
- Quantificar a relação entre os índices externos e internos;
- Inferir qual(is) índice(s) interno(s) é/são mais adequado(s) para cada algoritmo de agrupamento.

1.3 Justificativa

Atualmente existem diversos índices de validação de agrupamento consolidados na literatura. Estes índices são muito úteis na prática como uma medida quantitativa para avaliar a qualidade dos agrupamentos [VCH10]. Estes índices possuem características particulares que podem fazer com que cada um deles seja capaz de superar outros em classes específicas de problemas. Desta maneira, não é possível apontar um índice universalmente mais confiável [LLX+10, VCH10]. Por este motivo, escolher os índices mais apropriados para avaliar os resultados de um algoritmo de agrupamento em especial continua sendo um desafio [LLX+10].

1.4 Organização da Dissertação

O restante do texto esta organizado da seguinte forma: o capítulo 2 apresenta a fundamentação teórica sobre agrupamento, critérios de validação e regressão; o capítulo 3 relaciona alguns trabalhos já publicados com o conteúdo desta dissertação; a metodologia, objetivo deste trabalho, é apresentada no capítulo 4; no capítulo 5 são apresentadas as

avaliações experimentais realizadas com base na metodologia; a validação da metodologia proposta são apresentadas no capítulo 6; e por fim o capítulo 7 apresenta as considerações finais desta dissertação, com sugestões para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo são abordados conceitos relacionados ao desenvolvimento deste trabalho a fim de facilitar a compreensão das características contempladas pelas técnicas da literatura e pelo trabalho proposto. A primeira seção apresenta o conceito de algoritmos de agrupamento, sua classificação e os algoritmos utilizados neste trabalho. Uma série de medidas que avaliam qualidade de um agrupamento são apresentadas na seção 2.2. Por fim, é apresentado o conceito de regressão descrevendo a técnica de regressão linear e o algoritmo M5 que são utilizados neste trabalho.

2.1 Algoritmos de Agrupamento

Os algoritmos de agrupamento existentes são muitos e cada algoritmo utiliza uma determinada estratégia para agrupar as instâncias. Esses algoritmos podem ser classificados por meio de diferentes aspectos. A classificação mais usual foi proposta por Jain *et al.* [JMF99], Os algoritmos são classificados de acordo com o método adotado para definir os grupos, ou seja, algoritmos hierárquicos, particionais, baseados em grid ou baseados em densidade.

2.1.1 K-Means

O algoritmo *K-Means* [HW79] é classificado como particional. Tem como parâmetro de entrada o número de grupos k, que deve ser definido pelo usuário, o algoritmo particiona

o conjunto de dados de n pontos em k grupos, formados de acordo com alguma medida de distância, tais como: Manhattan, Euclidiana, Chebyshev, etc [HBV01].

O algoritmo inicia definindo um conjunto de k centróides para cada grupo. Esta seleção pode ser feita aleatoriamente ou através de uma heurística. Cada objeto é atribuído ao centróide mais próximo formando assim o conjunto de grupos inicial. O centroide de cada grupo é recalculado e este processo é repetido até que os grupos estabilizem, ou seja, não haja mais alterações no grupo a que cada instância pertence. O objetivo deste algoritmo é minimizar a distância entre cada objeto e o centroide do grupo ao qual pertence [HBV01]. A descrição do *K-Means* é apresentada no algoritmo 1. E a figura 2.1 demonstra as etapas realizadas por este algoritmo.

Algorithm 1 K-Means

Entrada: Um conjunto de instâncias, Número de grupos k

Saida: Uma partição de X em k grupos

Escolher aleatoriamente k instâncias como centroides dos grupos

repita

para cada instância $x_i \in X$ e grupos C_j, j = 1,...,k *faça*

Calcular a distância $d(x_i,\vec{j})$ entre x_i e o centroide $\bar{x}^j$ do grupo

fim

para cada instância x_i *faça*

Associar x_i ao *cluster* com centróide mais próximo

fim

para cada grupo C_j, j = 1,...,k *faça*

Recalcular o centróide

fim

até não haver mais alteração na associação dos objetos aos clusters;

A maior desvantagem deste algoritmo é a necessidade de se selecionar um número de grupos k previamente, o que exige que se saiba *a priori* quantos grupos tem o *dataset* ou então executar o algoritmo diversas vezes variando o valor de k até encontrar uma partição ideal. Desta maneira entende-se que o valor de k é extremamente importante e depende diretamente do algoritmo [BL97].

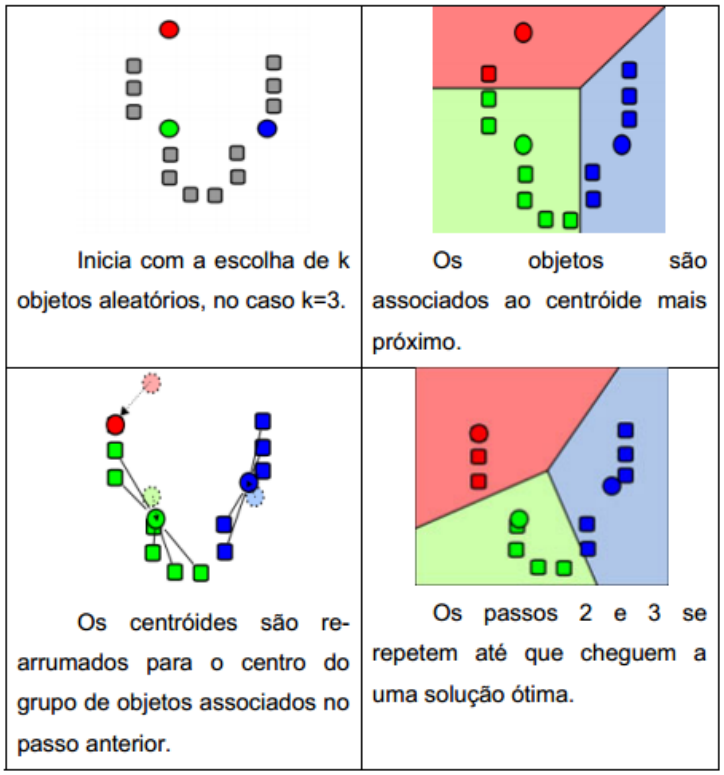

Figura 2.1: Exemplo da execução algoritmo K-means. (Extraída de [Cas09])

2.1.2 DBSCAN

Os algoritmos de agrupamento baseados em densidade têm como objetivo determinar grupos com alta densidade de objetos separados por regiões de baixa densidade. Entre os algoritmos baseados em densidade o DBSCAN [SEKX98] é um dos mais conhecidos na literatura possuindo uma complexidade $O(n^2)$ [EKSX96].

O DBSCAN utiliza o conceito de densidade baseada em centro, ou seja, a densidade de um objeto x_i é a quantidade mínima de objetos em uma vizinhança *MinPoints* determinada por um raio Eps (ε) incluindo o próprio objeto.

Deste modo a densidade de um objeto tem forte dependência com o valor do (ε) definido como parâmetro. Isso mostra a importância da escolha de bons valores para parametrizar o algoritmo, o que na maioria das vezes não é uma tarefa trivial.

A abordagem baseada em centro classifica um ponto como: centro, limite ou borda e ruído [TSK+06]. Um objeto é central se o número de vizinhos dentro de sua vizinhança, conforme uma função de distância e Eps, ultrapassar um limite (*MinPoints*). Um objeto

classificado como limite quando este não é um objeto central, mas fica dentro da vizinhança de um objeto central. E um objeto é classificado como ruído quando não é nem objeto central nem objeto limite.

Para aplicação do DBSCAN devem ser seguidos os seguintes passos conforme o algoritmo 2 [TSK+06]. A figura 2.2 demonstra um exemplo de aplicação do algoritmo DBSCAN.

Algorithm 2 DBSCAN

1 - Classificar todas as instâncias como objetos do tipo: centro, limite ou ruído;

2 - Eliminar os objetos rotulados como ruído;

3 - Colocar uma aresta entre todos os objetos de centro que estejam dentro do Eps uns dos outros;

4 - Tornar cada grupo de Objetos de centro um grupo separado;

5 - Atribuir cada objeto limite a um dos grupos dos seus objetos de centro associados;

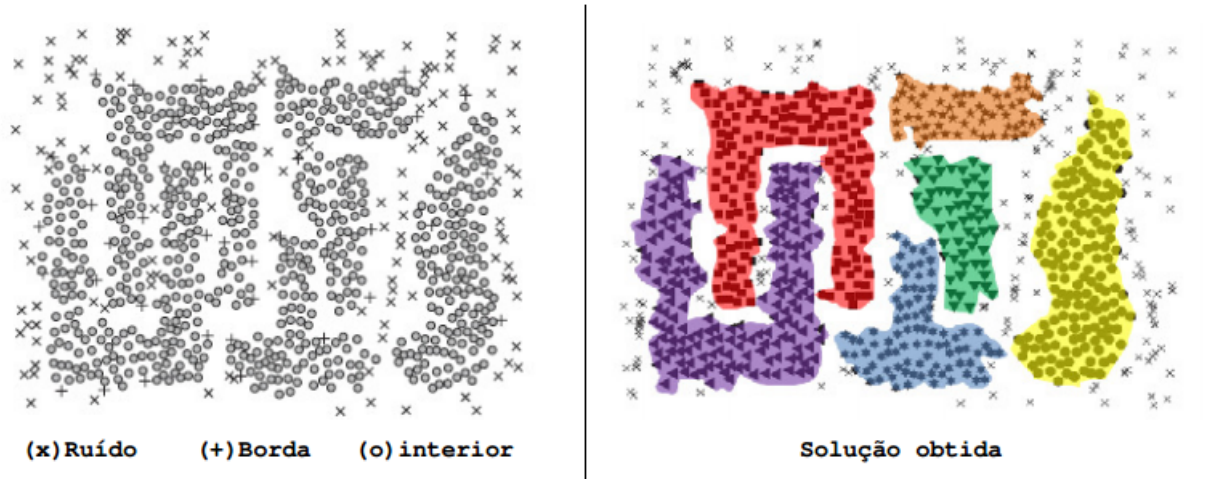

Figura 2.2: Classificação de 3000 objetos de duas dimensões pelo DBSCAN. (Extraída de [KST09])

2.2 Critérios de Validação

Uma das questões mais importantes na análise de agrupamento é a avaliação dos resultados para encontrar o particionamento que melhor se ajusta aos dados subjacentes

[LLX+10]. Este procedimento é conhecido pelo termo de validação de agrupamentos [TSK+06] ou validade de grupos [HBV01].

Dois critérios foram propostos para a validação de agrupamento e seleção de um esquema ideal de agrupamento [BL96]: (i) compacidade - os membros de cada grupo devem estar o mais próximo possível uns dos outros; (ii) separação - os grupos devem ser amplamente espaçados. Um índice de compacidade comum é a variância [HKP06], a qual deve ser minimizada. A separação pode ser medida por meio da distância entre os centros dos grupos ou entre os mais próximos ou mais distantes membros. Existem vários índices de validação de agrupamento diferentes que são muito úteis como medidas quantitativas para avaliar a qualidade da partições dos dados.

Apesar de terem sido propostos vários índices, cada um é focado em um determinado tipo de agrupamento, e eles não podem lidar com alguns aspectos, tais como a variação de densidade ou ruído. Estas propriedades ou aspectos transformam cada índice aptos a superar outros em classes específicas de problemas. Com base nos argumentos acima, escolher o índice de validação mais adequado para avaliar os resultados de um agrupamento continua sendo um desafio [VCH10].

2.2.1 Critérios Externos

Os índices externos são normalmente utilizados para comparar os resultados de um agrupamento com uma partição conhecida previamente [XW09]. Esta partição pode refletir a nossa intuição sobre a estrutura dos dados, ser sugerida por um especialista ou então ser definida com base em uma correspondência entre os grupos encontrados e os rótulos já conhecidos.

Considere P uma partição previamente conhecida de um conjunto de dados X com n pontos. C é o resultado de um algoritmo de agrupamento aplicado sobre X. A avaliação de C por um índice externo é obtida comparando C e P. Considerando-se um par de pontos $(x_i, x_j) \in \{X \times X\} | 1 \leq i \leq n, 1 \leq j \leq n$, pode-se calcular a frequências a, b, c e d, que se referem a quatro casos diferentes com base em como x_i e x_j são arranjados em C e P [XW09].

- a: frequência de pares x_i e x_j que pertencem ao mesmo grupo em C e mesma

categoria em P;

- b: frequência de pares x_i e x_j que pertencem ao mesmo grupo em C mas diferentes categorias em P;
- c: frequência de pares x_i e x_j e pertencem a grupos diferentes em C mas a mesma categoria em P;
- d: frequência de pares x_i e x_j que pertencem a grupos diferentes em C e diferentes categorias em P;

Neste trabalho, foram utilizados os índices externos *Jaccard* [XW09], *Rand* [Ran71] e *Fowlkes-Mallows* [FM83]. Estes critérios são baseados na frequência de pares de instâncias correta e incorretamente agrupados conforme os casos apresentados anteriormente.

A Fig. 2.3 apresenta um exemplo de uma partição P em que todas as instâncias pertencem a mesma categoria (a) e o resultado de um agrupamento C contendo 3 grupos distintos (b), identificados pela cor dos pontos. As frequências para cada um dos quatro casos são: $a = 15$, $b = 0$, $c = 40$ e $d = 0$, conforme Tabela 2.1 que apresenta o cálculo das frequências. As próximas sub-seções descrevem o conjunto de índices utilizados neste trabalho.

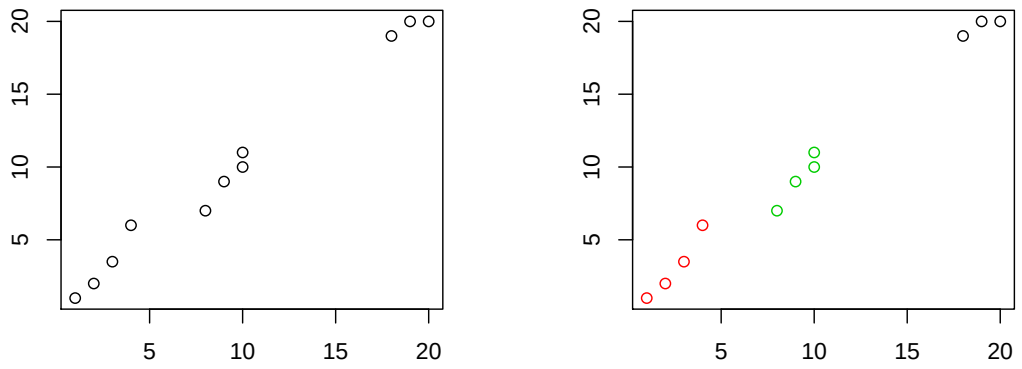

Figura 2.3: (a) Partição P conhecida *a priori*. (b) Partição C resultado do agrupamento.

Jaccard

O índice de *Jaccard* [XW09] J ou coeficiente de similaridade de *Jaccard* é uma medida estatística utilizada para comparar a similaridade e a diversidade entre conjuntos de dados.

Tabela 2.1: Cálculo das frequências dos casos para o exemplo da figura 2.3

Caso	**Par de Pontos**	**Total**
a	(x1 e x2),(x1 e x3), (x1 e x4), (x2 e x3), (x3 e x4), (x9 e x10), (x9 e x11), (x10 e x11)	15
b		0
c	(x1 e x5), (x1 e x6), (x1 e x7), (x1 e x8), (x2 e x5), (x2 e x6), (x2 e x7), (x2 e x8), (x3 e x5), (x3 e x6), (x3 e x7), (x3 e x8), (x4 e x5), (x4 e x6), (x4 e x 7), (x4 e x8)	40
d	(x5 e x9), (x5 e x10), (x5 e x11), (x6 e x9), (x6 e x10), (x6 e x11), (x7 e x9), (x7 e x10), (x7 e x11), (x8 e x9), (x8 e x10), (x8 e x11)	0

Este índice resulta em valores que variam no intervalo fechado $[0, 1]$. J retorna um valor perto de 0 quando aplicado em partições diferentes e perto de 1 quando aplicado em partições similares. O índice *Jaccard* é definido pela equação 2.1.

$$J = \frac{a}{a + b + c} \tag{2.1}$$

Para o exemplo da figura 2.3 o valor do índice de *Jaccard* é $0,27$, conforme a equação 2.2, o que significa que há uma baixa similaridade entre a partição P que originalmente possuía apenas 1 grupo e o agrupamento C que resultou em 3 grupos.

$$J = \frac{15}{15 + 0 + 40} = 0,27 \tag{2.2}$$

Rand

Assim como J, o índice *Rand* [Ran71] R mede a similaridade entre duas partições P e C. Este índice também resulta em valores no intervalo $[0, 1]$, sendo que 0 indica que C e P são muito diferentes e 1 indica que os conjunto de dados são fortemente similares. *Rand*

é definido pela Eq. 2.3 e o valor apresentado foi calculado com base no exemplo da Fig. 2.3.

$$R = \frac{a+d}{a+b+c+d} \tag{2.3}$$

Para o exemplo da figura 2.3 o valor do índice de *Rand* é apresentado na equação 2.4.

$$R = \frac{15+0}{15+0+40+0} = 0,27 \tag{2.4}$$

Fowlkes And Mallows

O valor deste índice está diretamente relacionado com a similaridade entre C e P, o que significa que quanto maior o valor obtido, maior é a similaridade entre as partições analisados. O índice *Fowlkes-mallows* [FM83] FM é definido pela equação 2.5.

$$FM = \sqrt{\frac{a}{a+b}\,\frac{a}{a+c}} \tag{2.5}$$

Para o exemplo da figura 2.3 o valor do índice de *Fowlkes And Mallows* é apresentado na equação 2.6.

$$FM = \sqrt{\frac{15}{15+0}\,\frac{15}{15+40}} = 0,52 \tag{2.6}$$

2.2.2 Critérios Internos

Na prática, informações externas, como o rótulos das classes, muitas vezes não estão disponíveis. Portanto, na situação em que não existam informações externas disponíveis, os índices internos de validade são a única maneira de avaliar um agrupamento [LLX+10]. Normalmente, estes índices são capazes de quantificar a qualidade do resultado do agrupamento usando somente frequências e propriedades inerentes do agrupamento [HBV01] como, por exemplo, considerando apenas a matriz de proximidade.

Neste trabalho, foram utilizados os índices internos *DBI* [DB79], *Dunn* [Dun74], *Gamma* [BH75, VCH10], *C-index* [HL76, VCH10] e *Silhouette* [Rou87].

Os índices internos foram aplicados em diferentes configurações dos algoritmos de agrupamento, a fim de compreender melhor o comportamento e as relações entre estes

métodos de validação de agrupamento. Os índices internos foram utilizado como critérios relativos. De acordo com Xu et al. [XW09], os critérios relativos comparam resultados de agrupamento gerados por diferentes algoritmos ou então pelo mesmo algoritmo, mas com diferentes parâmetros de entrada, ou seja, eles avaliam o resultado do agrupamento comparando-o com outros agrupamentos [HBV01]. As próximas sub-seções descrevem o conjunto de índices citados anteriormente.

DBI

O índice *Davies-Bouldin* (*DBI*) [DB79] é calculado em função da razão entre a soma da dispersão interna dos agrupamentos e a distância entre eles, e é definido pela Eq. 2.7

$$DBI = \frac{1}{n}\sum_{i=1}^{n} max_{j:i\neq j}\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \tag{2.7}$$

onde n é o número de grupos, σ_i é a distância média entre todos os pontos do grupo i e seu centroide c_i, σ_j é a distância média entre todos os pontos do grupo j e seu centroide c_j e $d(c_i, c_j)$ é a distância entre os centroides c_i e c_j.

Fica claro para a definição acima que *DBI* é a similaridade média entre cada cluster e seu correspondente mais semelhante [HBV01]. É desejável que os agrupamentos tenham o mínimo de similaridade possível entre si. Assim, menores valores de *DBI* correspondem a agrupamentos compactos com centroids distantes uns dos outros.

A tabela 2.2 apresenta a distribuição dos pontos referente ao gráfico da figura 2.3(b) que será utilizada como base para exemplificar o cálculo do índice *DBI*.

O primeiro passo é calcular o centróide para cada grupo do *dataset* (tabela 2.3) e a distância do centroide de um determinado grupo em relação aos outros (tabela 2.4). Neste exemplo a distância adotada foi a euclidiana.

O segundo passo é calcular, para cada instância, a distância em relação ao centroide do grupo o qual pertence. Ao final calcula-se o valor da distância média de cada grupo. Este processo é demonstrado nas tabelas 2.5 - 2.7.

No terceiro passo, para cada par de grupo, a distância média é somada e dividida pela distância entre os centroides correspondentes, conforme tabela 2.8. Após efetuado o cálculo, é realizado o somatório dos valores máximos entre os pares dos grupos, conforme tabela 2.9.

Tabela 2.2: Distribuição dos pontos do *dataset* apresentado na figura 2.3(b)

Instância	**Atributo x**	**Atributo y**
0	1	1
1	2	2
2	4	6
3	3	3,5
4	10	11
5	10	10
6	8	7
7	9	9
8	20	20
9	18	19
10	19	20

Tabela 2.3: Centróides dos grupos do *dataset* apresentado na figura 2.3(b)

Centroide	**x**	**y**
Grupo 1	2,5	3,125
Grupo 2	9,25	9,25
Grupo 3	19	6

Tabela 2.4: Distância entre os centróides dos grupos *dataset* apresentado na figura 2.3(b)

Centroides	**Distância**
Grupo 1 - Grupo 2	9,12
Grupo 1 - Grupo 3	23,37
Grupo 2 - Grupo 3	19

Tabela 2.5: Distância de todas as instâncias do grupo 1 em relação ao seu centróide.

Instância	**Distância**
0	2,601
1	1,24
2	3,25
3	0,625
Dist Média	1,924

Tabela 2.6: Distância de todas as instâncias do grupo 2 em relação ao seu centróide.

Instância	**Distância**
4	1,90
5	1,06
6	2,58
7	0,36
Dist Média	1,48

Tabela 2.7: Distância de todas as instâncias do grupo 3 em relação ao seu centróide.

Instância	**Distância**
8	1,05
9	1,20
10	0,34
Dist Média	0,87

O resultado final do índice *DBI* é dado pelo somatório obtido na tabela 2.9 dividido pela quantidade de grupos encontrados. Para este exemplo o resultado final seria $0,9093 \div 3 = 0,3031$.

Tabela 2.8: Cálculo da distância média entre os pares de grupos dividida pela distância de seu respectivo centroide.

Grupos	$(\sigma_i + \sigma_j) \div d(c_i, c_j)$
1 - 2	0,3728
1 - 3	0,1193
2 - 3	0,1637

Tabela 2.9: Máximo valor das distâncias entre os grupos.

Grupos	**Max**
1 - 2	0,3728
1 - 3	0,1193
2 - 3	0,1637
Somatório	0,9093

Dunn

O índice de *Dunn* [Dun74] D é calculado a partir da razão entre a menor distância intergrupo e a maior distância intragrupo. Seu valor varia no intervalo $[0, \infty)$, significando que quanto maior o valor obtido, mais compactos e bem separados são os grupos. Este índice tenta identificar aglomerados compactos e bem separados [HBV01].

Seja $D(C_i, C_j)$ a distância entre dois grupos C_i e C_j calculada como a menor distância entre um par de pontos $x \in C_i$ e $y \in C_j$ (Eq. 2.8) e $diam(C_i)$ o diâmetro do grupo C_i igual a máxima distância entre dois de seus componentes (Eq. 2.9). O índice *Dunn* é formalmente definido pela Eq 2.10, onde k é o número de grupos.

$$D(C_i, C_j) = \min_{x \in C_i, y \in C_j} (D(x, y)) \tag{2.8}$$

$$diam(C_i) = \max_{x,y \in C_i} (D(x, y)) \tag{2.9}$$

$$D(k) = \min_{i=1,\ldots,k} \left(\min_{j+1,\ldots,k} \left(\frac{D(C_i, C_j)}{\max_{l=1,\ldots,K} (diam(C_l))} \right) \right) \tag{2.10}$$

A principal desvantagem em relação aos outros critérios é a complexidade computacional quadrática. Além disso, este critério é bastante sensível a ruído.

Para exemplificar o cálculo do índice *Dunn* será utilizado o mesmo conjunto de dados referente ao exemplo do *DBI* que foi apresentado na figura 2.3(b).

Inicialmente é necessário calcular a distância entre os pontos de uma mesma classe e a maior distância de cada classe, conforme exemplificado nas tabelas 2.10 - 2.12. Em seguida, é calculada a distância entre os pontos que não estão presentes nas tabelas 2.10 - 2.12 e a menor distância deste conjunto.

Tabela 2.10: Cálculo da distância dos pontos do grupo 1.

Pontos	**Distâncias**
0 - 1	1,411
0 - 2	5,830
0 - 3	3,201
1 - 2	4,472
1 - 3	1,802
2 - 3	2,692
Máximo	5,830

Para finalizar o cálculo é necessário dividir a distância mínima pela distância máxima. Ou seja, o valor final do índice *Dunn* é $4,123 \div 5,830 = 0,707$

Silhouette

O índice *Silhouette* [Rou87] define a qualidade dos agrupamentos com base na proximidade entre os objetos de um determinado grupo e na proximidade desses objetos ao grupo mais próximo. *Silhouette* é definido formalmente pela Eq. 2.11 que calcula o valor do índice para uma única instância x que pertence ao grupo j, onde $d(x, C_j)$ é a dissimilaridade média de x em relação a todos os pontos de j, e h é o grupo mais próximo da instância x.

Tabela 2.11: Cálculo da distância dos pontos do grupo 2.

Pontos	**Distâncias**
4 - 5	1
4 - 6	4,472
4 - 7	2,236
5 - 6	3,605
5 - 7	1,414
6 - 7	2,236
Máximo	4,472

Tabela 2.12: Cálculo da distância dos pontos do grupo 3.

Pontos	**Distâncias**
8 - 9	2,236
8 - 10	1
9 - 10	1,414
Máximo	2,236

$$s(x) = \frac{d(x, C_h) - d(x, C_j)}{\max(d(x, C_h), d(x, C_j))} \tag{2.11}$$

$s(x)$ varia entre o intervalo $[-1, 1]$. Quanto mais próximo de 1 melhor a alocação do objeto no grupo. Após calcular o valor para todos os pontos do agrupamento, deve ser calculada a média para o grupo (S_j) e em seguida para o agrupamento como um todo (GS), conforme Eq. 2.12 e 2.13, onde N_j é o número de pontos do grupo j e K é o número de grupos.

$$S_j = \frac{\sum_{i=1}^{N_j} s(x_i)}{N_j} \tag{2.12}$$

$$GS = \frac{\sum_{j=1}^{K} S_j}{K} \tag{2.13}$$

Gamma

O índice *Gamma* [BH75, VCH10] Γ calcula o número de pares concordantes de objetos S_+, que é o número de vezes que a distância entre um par de objetos do mesmo grupo é menor do que a distância entre um par de objetos de um grupo diferente. Esse índice também calcula o número de pares discordantes S_-, que é o número de vezes que a distância entre um par de objetos do mesmo grupo é maior do que a distância entre um par de objetos de grupos diferentes. Γ é definido pela equação 2.14.

$$\Gamma = \frac{S_+ - S_-}{S_+ + S_-} \tag{2.14}$$

Este índice varia no intervalo $[-1, 1]$. Melhores partições devem ter valores mais altos de S_+, baixos valores de S_- e, consequentemente, terão altos valores de Γ.

C-Index

C-index [HL76, VCH10] é definido como:

$$C = \frac{d_w - min(d_w)}{max(d_w) - min(d_w)} \tag{2.15}$$

onde d_w é a soma das distâncias de todos os pares de instâncias de um mesmo grupo. Seja j o número de pares de instâncias no mesmo grupo, $max(d_w)$ e $min(d_w)$ são a soma das j maiores e menores distâncias, respectivamente, considerando todos os pares de distâncias. Assim, este índice deve ser minimizado e varia no intervalo entre $[0, 1]$.

2.3 Tarefa de Regressão

Regressão é um tipo de modelagem preditiva onde a variável dependente é contínua. Esta técnica é utilizada em diversas áreas como economia, administração, engenharias, etc. Segundo Hines [Hin06] a análise de regressão investiga a relação entre duas ou mais variáveis (variáveis preditoras) e uma variável dependente. Sendo que, quando os problemas envolvem mais de uma variável preditora, esses modelos são chamados de modelos de regressão múltipla.

2.3.1 Regressão Linear

A relação entre as variáveis explanatórias e a variável dependente é representada pela equação de regressão múltipla 2.16.

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon \tag{2.16}$$

As variáveis explanatórias são representadas pelos termos x_i, ao passo que a variável dependente é representada por y. A relação entre as variáveis é demonstrada por β_i, que indica quantitativamente como as variáveis explanatórias determinam a variáveis dependente e o ϵ é o erro aleatório em y [Har01]. Para estimar os coeficientes de regressão β_i, é utilizado o método dos mínimos quadrados [Wei05].

Segundo Han *et al.* [HKP06] existem 2 tipos principais de árvores para predição numérica: árvores de regressão e árvores modelo. Na árvore de regressão cada nó folha guarda um valor contínuo. Na árvore modelo cada nó folha contém um modelo de regressão que representa uma equação com múltiplas variáveis para a predição do atributo.

2.3.2 Árvore de Regressão M5

O algoritmo M5 foi desenvolvido por [Q+92] para tratar atributos e classes contínuas. M5 é um processo *Top Down Induction Decision Trees (TDIDT)*, mas com funções de regressão linear nos nós folhas, ao invés de um valor categórico predizendo a classe.

Os resultados da execução deste algoritmo são chamados de árvore modelo, conforme exemplo apresentado na figura 2.4. Após a árvore ser obtida, o algoritmo possui uma fase de particionamento, que divide o conjunto de dados; uma fase de poda, para reduzir o número de nós da árvore obtida; e uma fase adicional denominada *smoothing*, que tem como objetivo reduzir a grande diferença dos valores preditos entre os nós-folha [WW96]. No modelo linear conforme figura 2.4 (direita), cada parte da equação corresponde a um dos atributos do *dataset Wine* [CCA+09] (por exemplo, calorias, densidade, pH, etc...) multiplicados uma constante que quantifica sua contribuição no valor final do atributo alvo.

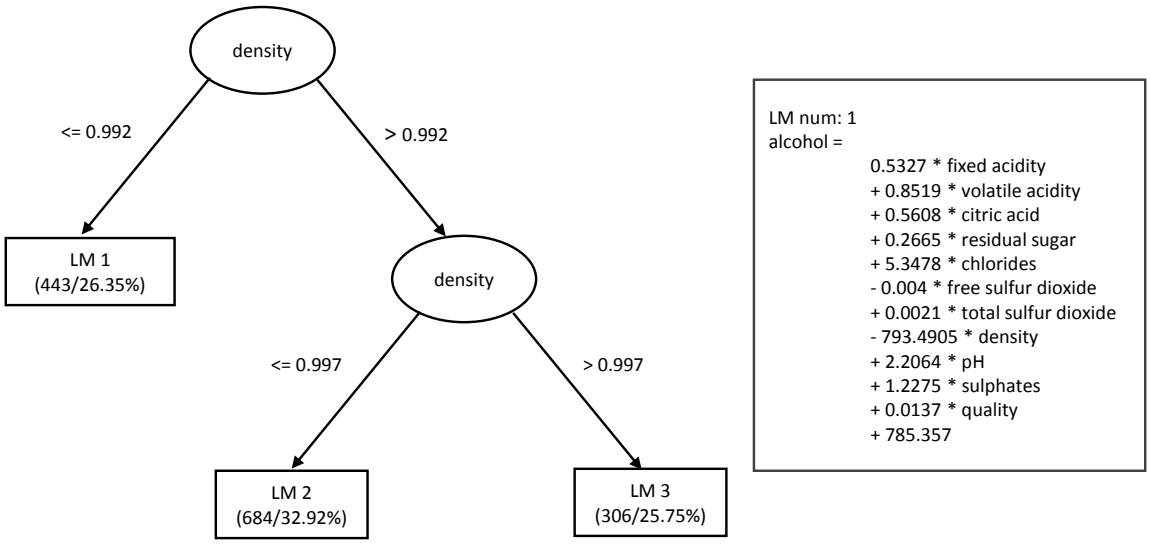

Figura 2.4: Exemplo de árvore modelo e modelo linear para o *dataset* Wine.

Capítulo 3

Trabalhos Relacionados

Neste capítulo são apresentados três trabalhos que possuem relação com esta dissertação.

3.1 Relative Clustering Validity Criteria: A Comparative Overview

[VCH10] propõem uma comparação de critérios de validação de agrupamentos utilizando um método estatisticamente mais robusto do que o tradicionalmente usado na literatura [MC85]. Os autores calculam a correlação de Pearson entre critérios internos e um critério externo com o objetivo de identificar relações entre os valores.

A metodologia utilizada no artigo pode ser explicada utilizando a tabela 3.1. Cada linha da tabela representa um *dataset*, as colunas 1 e 2 apresentam o valor de um índice interno e a terceira coluna o valor de um índice externo. Calculando a correlação de Pearson entre o primeiro critério interno e o critério externo foi obtido o valor de 0,9627. Este elevado valor reflete o fato do Critério Interno 1 particionar os dados da mesma forma que o Critério Externo os classifica. Então pode-se concluir que o Critério Interno 1 tem correlação com o critério externo. Os *datasets* utilizados pelos autores para realizar os experimentos foram os mesmos descritos em [MC85, Mil81].

Na metodologia proposta nesta dissertação de mestrado, apresentada na Seção 4, a relação entre critérios internos e externos é capturada diretamente através de um modelo de aprendizado baseado em regressão linear e árvores de regressão. Ao contrário de

[VCH10], foi apresentado a estratégia de geração dos *datasets* sintéticos e a avaliação dos modelos aprendidos.

Partição	Crit. Interno. 1	Crit. Interno 2	Crit. Ext.
1	0.75	0.92	0.82
2	0.55	0.22	0.49
3	0.20	0.56	0.31
4	0.95	0.63	0.89
5	0.60	0.25	0.67

Tabela 3.1: Exemplo de critérios internos e externos de avaliação de cinco partições de um conjunto de dados usando Correlação de Pearson. Adaptado de [Vendramin et al. 2010]

3.2 A Combination Approach to Cluster Validation Based on Statistical Quantiles

Um problema conhecido e muito discutido na literatura comumente associado ao uso de algoritmos de agrupamento é estimar o número de grupos existentes em um conjunto de dados. Grande parte dos algoritmos de agrupamento necessitam deste número previamente definido, pois utilizam como parâmetro de entrada. Uma possível solução para este problema é avaliar a qualidade de cada agrupamento proposto para um determinado conjunto de dados.

No trabalho proposto por [AS09] foram analisadas diferentes técnicas para detectar o número de grupos mais adequado para um conjunto de dados e também foi proposto um novo algoritmo baseado na combinação de vários índices de validação. Os índices de validação estudados e os algoritmos foram testados em dados sintéticos com distribuição gaussiana e no *dataset* real Iris. A técnica proposta baseia-se no cálculo das estatísticas de quantis das curvas de validação.

3.3 Understanding of Internal Clustering Validation Measures

Conforme discutido anteriormente, a validação de agrupamento pode ser dividida basicamente em dois tipos: a validação externa e interna. Na literatura foram propostos diversos índices de validação tanto internos quanto externos, no entanto estes índices podem ser afetados por diversas características dos dados como, por exemplo, o ruído dos dados ou então a densidade dos agrupamentos. Estas características podem ter um impacto significativo sobre o desempenho de uma medida de validação.

Levando isso em consideração, Liu *et al.* [LLX+10] apresentam um estudo detalhado de onze índices internos de validação de agrupamento e investigam as propriedades de validação dos índices em diferentes aspectos, tais como: impacto da monotonicidade, ruído, densidade, sub-grupos e distribuição heterogênea.

Para realizar os experimentos os autores geraram dados sintéticos com as características descritas acima, o resultado pode ser observado na figura 3.1.

Os experimentos realizados mostram que a maioria dos índices têm certas limitações

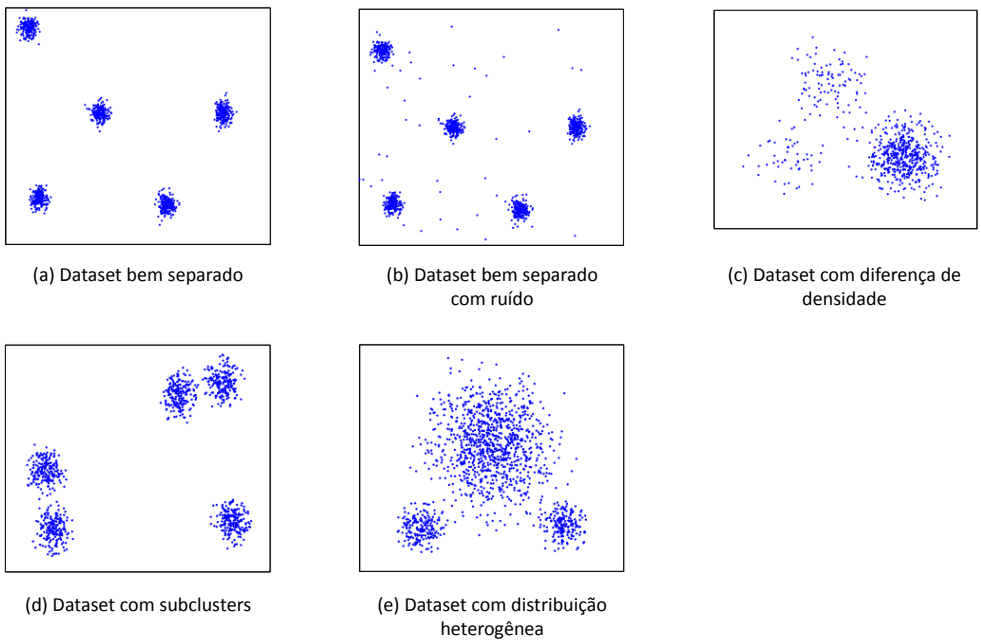

Figura 3.1: Exemplos de datasets com diferentes características.

em diferentes cenários de aplicação, principalmente quando os aspectos são ruído e sub grupos. O único índice que apresentou bons resultados em todos os aspectos testados foi o *S_Dbw* [HBV01].

Para o aspecto de monotonicidade, os melhores índices foram: *CH* [CH74], *I* [MB02], *Dunn*, *Silhouette*, *SD* [HVB00] e *XB* [XB91]; Para o aspecto de ruído, os melhores índices foram: *I*, *Silhouette*, *DBI*, *SD*, *XB*; Para o aspecto de densidade, os melhores índices foram: *CH*, *Dunn*, *Silhouette*, *DBI*, *SD*, *XB*; Para o aspecto de sub-grupos os melhores índices foram: *I* e *CH*; E para o aspecto de distribuição heterogênea, os melhores índices foram: *I*, *Dunn*, *Silhouette*, *DBI*, *SD*, *XB*;

Com base nos principais trabalhos relacionados estudados, não foi encontrado um trabalho que estabelecesse e quantificasse a relação entre critérios internos e externos com base em regressão e árvores modelo. Deste modo, nos próximos capítulos é apresentada a metodologia proposta para esse trabalho, um conjunto de experimentos realizados e a validação da metodologia proposta.

Capítulo 4

Metodologia

Este capítulo apresenta em detalhe a metodologia proposta nesta dissertação para investigar a relação dos critérios de validação de agrupamento externos e internos. A Fig. 4.1 apresenta uma visão geral da metodologia que é dividida em 5 etapas: (1) seleção ou de geração de conjuntos de dados, (2) agrupamento, (3) avaliação dos agrupamentos, (4) transformação de dados e (5) mineração de dados.

A primeira etapa é responsável pela seleção ou geração de um conjunto de *datasets*. Na segunda etapa é aplicado um algoritmo de agrupamento sobre os *datasets* gerados anteriormente. Com os dados agrupados, são calculados os critérios de validação externos e internos. A quarta fase organiza os índices calculados na validação dos grupos para que, por fim, seja aprendido um modelo preditivo baseado na relação entre os critérios externos e internos.

4.1 Geração Datasets

Esta etapa define os conjuntos de dados que serão utilizados. As instâncias neste conjunto de dados devem ser classificadas como partições previamente definidas para que possam ser utilizadas nos critérios de validação externos. É importante ressaltar que pode ser utilizado tanto conjuntos de dados reais quanto conjuntos sintéticos.

A utilização de conjuntos de dados sintéticos permite a variação das propriedades, tais como número de instâncias, características e grupos, distribuição de densidade, ruído, e assim por diante. O número de *datasets* a ser utilizado deve ser suficiente para gerar as

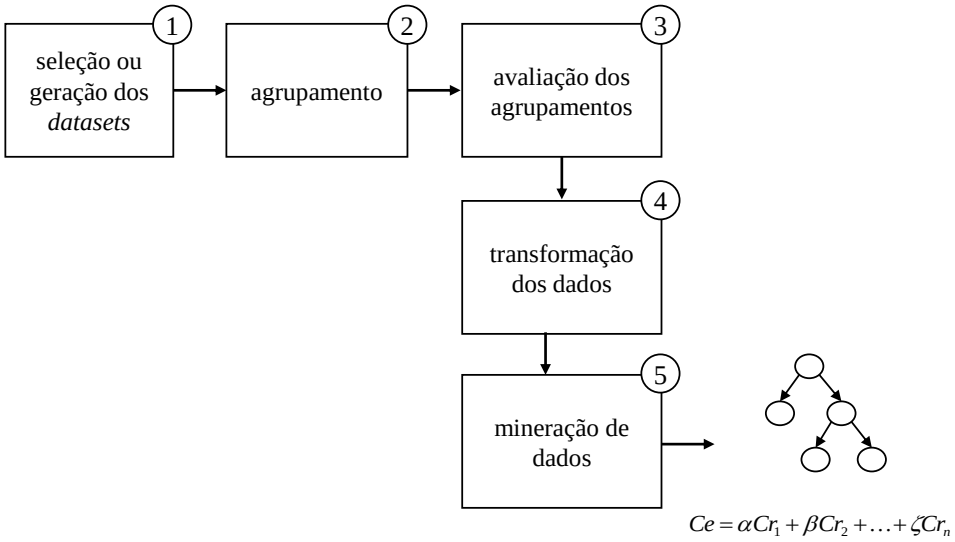

Figura 4.1: Metodologia proposta para selecionar de índices interno de validação de agrupamento mais adequado.

entradas para a tarefa de mineração seguinte. Por essa razão, o uso de *datasets* sintéticos se mostra mais apropriado que dados reais.

4.2 Agrupamento

A segunda etapa da metodologia consiste em selecionar e aplicar um algoritmo de agrupamento no conjuntos de dados gerado ou selecionado no passo anterior. Esta escolha depende das propriedades dos *datasets* da etapa anterior. Por exemplo, se os conjuntos de dados escolhidos forem muito grandes e apresentarem grupos de tamanhos diferentes com formas não convexas e múltiplas densidades, é recomendado que nesta etapa seja utilizado um algoritmo baseado em densidade com baixa complexidade computacional. Pois esse algoritmo de agrupamento escolhido deve ser executado várias vezes, variando os parâmetros de entrada para procurar por partições diferentes.

A saída gerada por esta etapa é um conjunto de partições, uma por *dataset*, que foram agrupadas seguindo as regras do algoritmo aplicado.

4.3 Avaliação dos Grupos

A terceira etapa calcula um conjunto de índices de validação de agrupamento internos e externos para cada resultado de agrupamento obtido na etapa anterior. Estes índices quantificam a qualidade dos resultados de agrupamento. Neste trabalho estão sendo aplicados os índices descritos no capítulo 2, *Jaccard*, *Rand* e *Fowlkes-Mallows* como índices externos e *DBI*, *Dunn*, *Gamma*, *C-index* e *Silhouette* como índices relativos.

É importante ressaltar que novos índices podem ser incorporados, assim como alguns podem não ser considerados dependendo do tipo de *dataset* ou algoritmo escolhido nas etapas anteriores.

4.4 Transformação

Os valores dos índices previamente calculados são transformados para serem utilizados como entrada para mineração de dados. Os critérios de minimização, tais como, DBI e C-index são invertidos, pois os demais índices são critérios de maximização. Assim, com os índices transformados, quanto maiores forem os valores dos índices, mais compactos e bem separados estão os grupos.

A transformação de um índice t_{ip} é definida pela equação 4.1, onde v_{ip} é o valor de um índice i calculado sobre a partição p e $\bar{v}_i$ é a média dos valores do índices i considerando todos os *datasets*.

Todos os valores transformados são normalizados. Pois conforme explicado no capítulo 2, os índices *DBI*, *Dunn* apresentam valores que podem ser de $[0, \infty)$, *Silhouette*, Γ apresentam valores que podem ser de $[-1, 1]$. Dessa forma, esses índices são normalizados para que todos os índices fiquem em intervalos de valores de $[0, 1]$.

$$t_{ip} = \begin{cases} 2\,\bar{v}_i - v_{ip}, & \text{se } i \text{ é um critério de minimização} \\ v_{ip}, & \text{caso contrário} \end{cases} \tag{4.1}$$

4.5 Mineração dos Dados

A tarefa de mineração de dados usa um conjunto de treinamento em que cada atributo é um índice transformado e cada instância representa um *dataset* gerado na primeira etapa

Tabela 4.1: Um exemplo de conjunto de treinamento com *Jaccard* (J) como atributo alvo.

	DBI	*S*	*D*	Γ	*C*	*J*
1	0.75994	0.79304	0.35734	0.72317	0.48791	**0.41040**
2	0.76354	0.69743	0.31469	0.62481	0.21182	**0.41048**
...						
n	0.87119	0.75364	0.24754	0.67538	0.40145	**0.41707**

da metodologia.

A Tabela 4.1 apresenta um exemplo de conjunto de treinamento onde *Davies-Bouldin*, *Silhouette*, *Dunn*, *Gamma* e *C-index* foram definidos como atributos preditivos e o índice externo *Jaccard* foi definido como atributo alvo. Este conjunto de treinamento é utilizado como entrada na tarefa de regressão para estimar os valores de índice externo baseado nos índices internos. A análise dos modelos obtidos na tarefa de regressão permite verificar qual(is) índice(s) interno é/são mais adequado(s) para avaliar os conjuntos de dados gerados ou selecionados na primeira etapa.

Capítulo 5

Avaliação Experimental

Este capítulo descreve os experimentos realizados, a fim de validar empiricamente a metodologia proposta no capítulo 4 e avaliar a qualidade dos modelos de regressão.

Duzentos conjuntos de dados sintéticos foram gerados durante a avaliação experimental. Estes conjuntos de dados contêm múltiplas distribuições de pontos em um espaço bi-dimensional, o que permite verificar visualmente a validade dos resultados [HBV01]. Para os experimentos realizados foi utilizado o software de computação estatística R[1] nos passos 1 a 4 da metodologia. E na última etapa foi utilizado o software Weka[2].

Os experimentos foram divididos em dois estudos de casos: (i) no primeiro estudo de caso a metodologia proposta é aplicada em *datasets* com características de compacidade e um algoritmo de agrupamento particional; (ii) no segundo estudo de caso foram utilizado *datasets* com múltiplas densidades e um algoritmo de agrupamento baseado em densidade. As seguintes métricas foram utilizados para avaliar os modelos de regressão obtidos: coeficiente de correlação e raiz do erro quadrático relativo (*root relative squared error*) [HKP06].

[1] `http://www.r-project.org/`

[2] `http://www.cs.waikato.ac.nz/ml/weka/`

5.1 Estudo de Caso 1: *datasets* com característica de compacidade e algoritmo de agrupamento particional

No primeiro estudo de caso, foram utilizados *datasets* com características de compacidade [HKK05]. A figura 5.1 mostra a metodologia descrita no capítulo 4 aplicada ao primeiro estudo de caso.

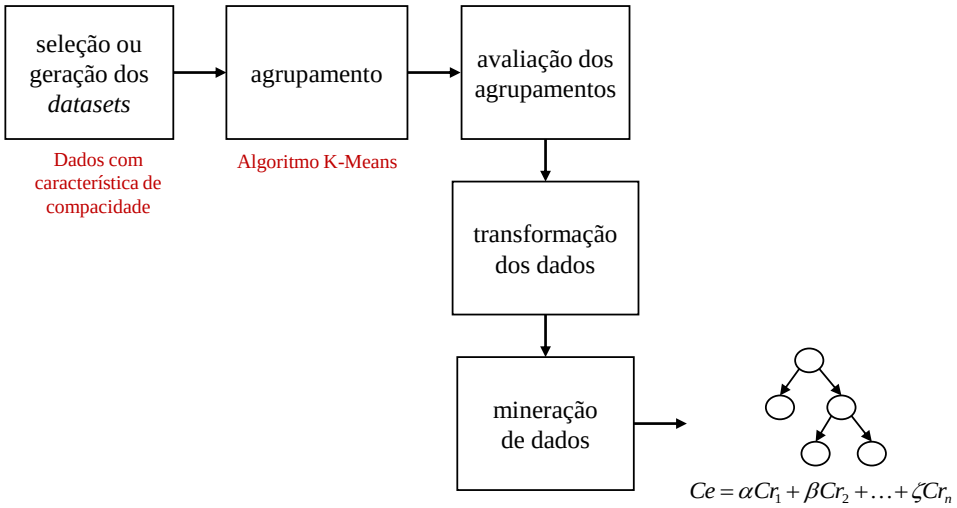

Figura 5.1: Metodologia proposta para selecionar os índices internos de validação de agrupamento mais adequados aplicada ao estudo de caso 1 considerando *datasets* com características de compacidade.

Na primeira etapa foram gerados 100 *datasets* nos quais foram distribuídos 150 pontos em um espaço bi-dimensional. O número de classes n_c variou de 2 a 5 sendo que o número de instâncias de cada grupo foi gerado aleatoriamente sempre totalizando 150 instâncias, ou seja, foram gerados 25 *datasets* com 2 classes, 25 *datasets* com 3 classes e assim sucessivamente. Metade dos pontos seguem uma Distribuição Gaussiana em que o raio r foi definido entre 1 e 10 e o desvio padrão $sd = 0.33r$. E a outra metade uma Distribuição Uniforme, simulando assim um ruído. A Figura 5.2 (esquerda) demonstra um exemplo de *dataset* com 3 classes distintas identificadas pela cor das instâncias. Este *dataset* foi gerado utilizando os parâmetros: $n_c = 3$, $r = 6.37$, $sd = 2.34$.

Após a geração dos *datasets*, na segunda etapa foi aplicado o algoritmo de agrupamento *k-Means* com $k = n_c$. Na terceira etapa os resultados de agrupamento foram avaliados utilizando os índices *Jaccard*, *Rand*, *Fowlkes-Mallows*, *DBI*, *Dunn*, *Gamma*, *C-index* e *Silhouette* que foram descritos na seção 2.2. Na etapa 4, para cada partição, os índices de validação de agrupamento foram invertidos e normalizados, conforme explicado no capítulo 4. A Figura 5.2 (direita) mostra o agrupamento realizado pelo *k-means* com $k = 3$ quando aplicado ao *dataset* gerado (à esquerda). Para este agrupamento, os valores dos índices calculados foram: $DBI = 0.77701$, $S = 0.46821$, $D = 0.05092$, $\Gamma = 0.83456$, $C = 0.29191$, $J = 0.28979$, $R = 0.43660$ e $FM = 0.48004$. E os resultados após transformação foram: $DBI = 0.31690$, $S = 1$, $\Gamma = 0.99682$, $C = 0.31690$, $D = 0.55269$, $J = 0.28979$, $R = 0.43660$ e $FM = 0.48004$.

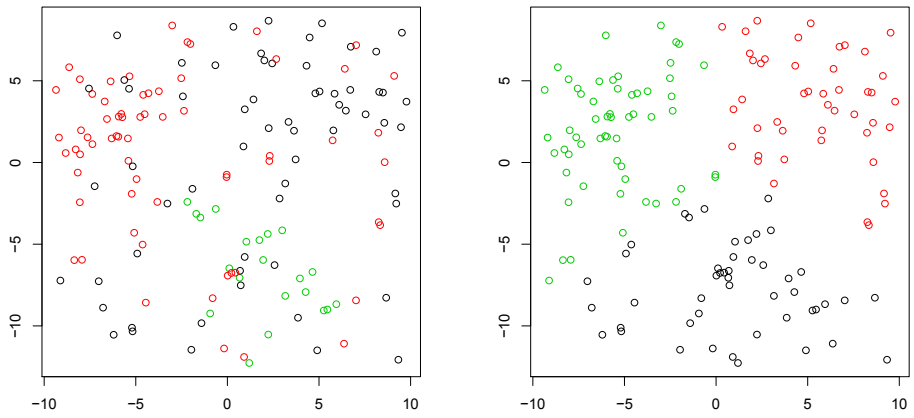

Figura 5.2: Um exemplo de *dataset* gerado (esquerda) e o resultado de agrupamento gerado pelo *k-Means* (direta).

Na etapa 5, foi gerado um modelo de regressão linear para cada índice externo de validação usando os internos. Os parâmetros padrão do Weka foram mantidos para todos os experimentos. As equações 5.1 - 5.3 apresentam os modelos de regressão linear obtidos no primeiro estudo de caso.

$$J = 0,4691\ S - 0,2766\ \Gamma - 0,2447\ C + 0,363 \tag{5.1}$$

$$R = -0,1746\ S + 0,2492\ \Gamma - 0,0543\ C + 0,5078 \tag{5.2}$$

$$FM = 0,581\ S - 0,3891\ \Gamma - 0,2563\ C + 0,5333 \tag{5.3}$$

Observando os modelos de regressão linear pode-se notar que *DBI* e *Dunn* foram irrelevantes para estimar os índices externos, pois não foram utilizados em nenhuma das equações de regressão linear obtidas. Já o índice *Silhouette* teve um impacto positivo no *Jaccard* e *Fowlkes-Mallows* mas negativo no *Rand*. O índice *Gamma* teve um comportamento oposto ao *Silhouette*, pois teve impacto positivo apenas sob o índice *Rand*. E, por fim, o *C-index* teve influência negativa em todos os índices externos.

Também foram aprendidos modelos baseados em árvores de regressão utilizando o algoritmo M5 [Q+92] para estimar os mesmos três índices externos. Foi definido um número mínimo de instâncias por folha $m = 4$, sendo m um parâmetro do algoritmo M5. A Fig. 5.3 apresenta a árvore modelo do índice *Jaccard*. Cada folha é um modelo de regressão linear distinto (LM) construído utilizando um subconjunto de instâncias, que foram selecionados utilizando a regra definida pelo caminho entre a raiz e a folha. Estes nós apresentam o número de casos e o erro de predição de cada LM. Para a árvore modelo do índice *Jaccard* obtida no primeiro estudo de caso, o índice *Gamma* foi o atributo mais discriminatório (raiz da árvore), seguido do índice *Silhouette*. Outros índices ajudaram a construir a árvore, mas todos LMs apresentaram uma influência mínima do *DBI*, além dos índices citados anteriormente. Para todos LMs, o índice *Silhouette* impactou positivamente na predição do *Jaccard* enquanto o *Gamma* e *DBI* tiveram um impacto negativo.

Em relação ao uso do *DBI* como um nó interno, a árvore de regressão do índice *Rand* não produziu diferenças significativas nos dez LMs (nós folhas), os quais são compostos pelos mesmos índices incluindo *Silhouette* e *Gamma*. Para o índice *Fowlkes-Mallows*, *M5* produziu uma árvore de um único nó com o mesmo LM apresentado na equação 5.3.

A Tabela 5.1 compara os modelos obtidos com regressão linear e árvore modelo. Para cada modelo aprendido, são apresentadas duas medidas de qualidade: coeficiente de correlação e raiz do erro quadrático relativo [HKP06]. Ao analisar o coeficiente de correlação, é possível observar que o *M5* teve desempenho melhor ou igual ao modelo de regressão linear.

Ao realizar a análise dos modelos é possível verificar que os índices de validação de

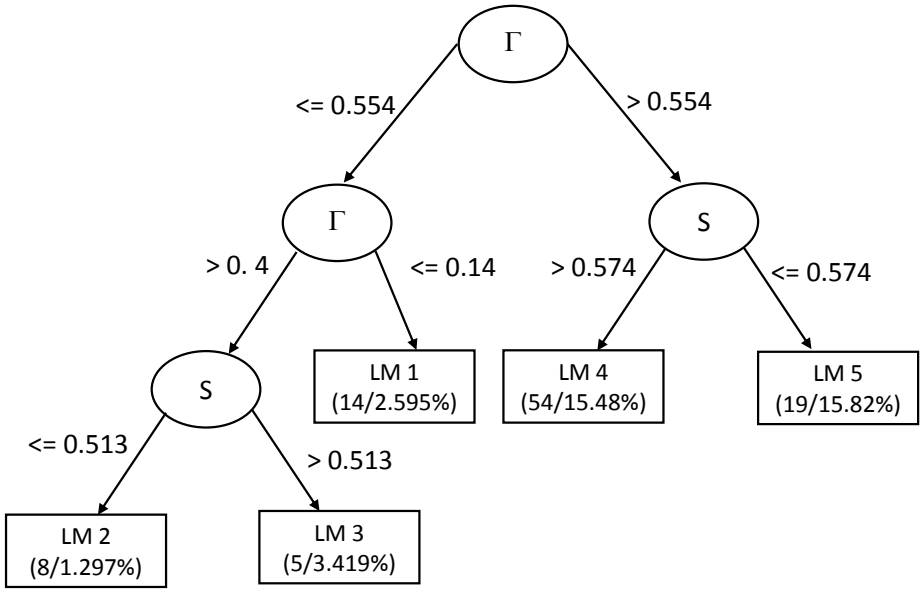

Figura 5.3: Árvore modelo usando o índice J como classe.

Tabela 5.1: Resultado das métricas de avaliação para o algoritmo *k-means*.

	$J(\%)$		$R(\%)$		$FM(\%)$	
	Reg. Linear	M5	Reg. Linear	M5	Reg. Linear	M5
Coeficiente de Correlação	97.82	98.75	91.82	96.60	98.01	98.01
Raiz do erro quadrático relativo	20.78	15.83	39.61	25.91	19.86	19.86

agrupamento internos mais adequados para avaliar os conjuntos de dados gerados usando o algoritmo *k-means* foram *Silhouette* e *Gamma*. Estes índices são capazes de avaliar a grande maioria das partições de um modo semelhante, quando comparado com os índices externos.

5.2 Estudo de Caso 2: *datasets* com densidade múltipla e algoritmo de agrupamento baseado em densidade

Neste segundo estudo de caso, a metodologia proposta foi aplicada utilizando vários conjuntos de dados com múltiplas densidades [HKK05] e um algoritmo de agrupamento baseado em densidade conforme mostra a figura 5.4.

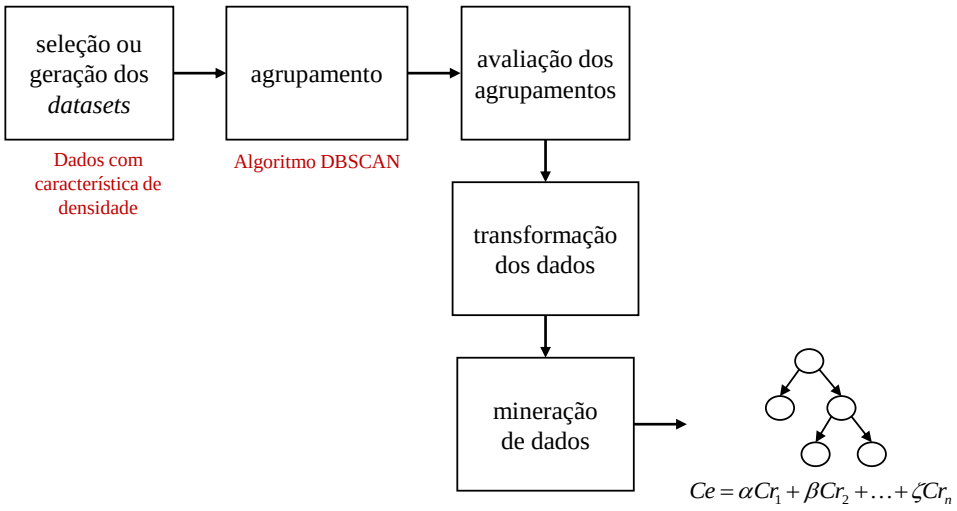

Figura 5.4: Metodologia proposta para selecionar os índices internos de validação de agrupamento mais adequados aplicada ao estudo de caso 2 considerando *datasets* com características de densidade.

Na primeira etapa, 100 conjuntos de dados sintéticos distintos foram gerados com 150 instâncias em cada um, distribuídos em um espaço bidimensional. Para estes conjuntos de dados o número de classes variou da seguinte forma: 25 *datasets* com 2 classes, 25 *datasets* com 3 classes e assim por diante. Os 150 pontos de cada *dataset* foram gerados utilizando uma distribuição gaussiana bivariada [Goo63] definida pela equação 5.4.

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right) \tag{5.4}$$

onde μ_x é a média de x, μ_y é a média de y, σ_x é o desvio padrão de x, σ_y é o desvio padrão de y e ρ é a correlação.

O objetivo de usar uma distribuição gaussiana bivariada para geração dos *datasets* é gerar conjuntos de dados com diferentes formas. Para cada grupo, σ_x e σ_y foram definidos aleatoriamente de 0,5 a 1,0. Isto assegura uma variação aleatória para espalhar os pontos em torno de x e y em cada grupo.

A variação da densidade foi obtida considerando um número aleatório de pontos em cada grupo. Foi estabelecido que cada grupo deve ter pelo menos 2 pontos. Considerando um exemplo de um conjunto de dados com dois grupos, temos $Size_{C1} = random[2, 148]$, $Size_{C2} = 150 - Size_{C1}$.

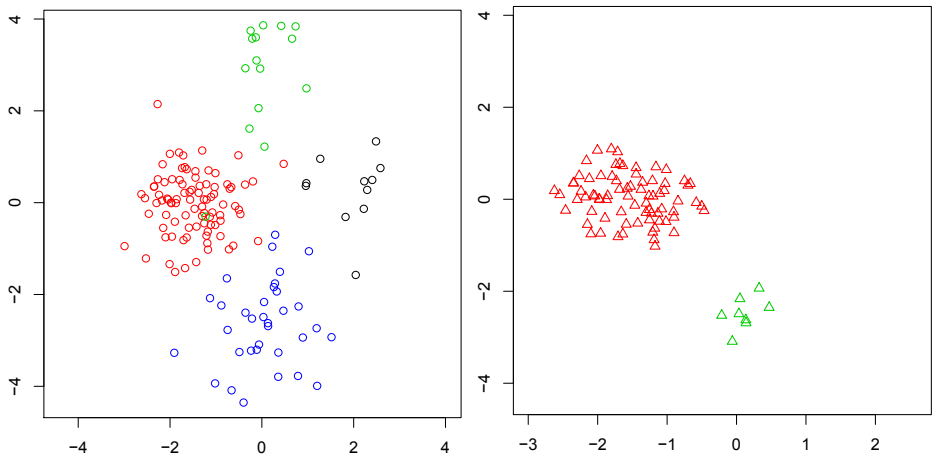

Figura 5.5: Exemplo de um conjunto de dados gerados com 4 classes (esquerda) e o agrupamento obtido após aplicação do DBSCAN (direita) com 3 grupos.

Na figura 5.5 (à esquerda) é apresentado um exemplo de um conjunto de dados com 4 grupos. Pode-se notar que esses grupos têm diferentes densidade e forma.

A segunda etapa da metodologia consiste em aplicar o algoritmo DBSCAN para cada conjunto de dados gerados. No entanto, um problema clássico em algoritmos de agrupamento baseado em densidade é definir os valores apropriados para o raio (ε) e número mínimo de pontos (*MinPoints*) para gerar um bom resultado de agrupamento. Muitas vezes, este problema é resolvido por tentativa e erro [CAC10]. Como foram gerados 100 *datasets* se tornaria inviável setar os parâmetros por tentativa e erro, desta maneira, foi adotada a estratégia descrita em [ZWL12], que propõem que esses parâmetros sejam

adaptáveis e auto-ajustáveis, sem qualquer intervenção manual.

A abordagem para determinar os parâmetros ε e *MinPoints* para cada conjunto de dados é dividida em quatro etapas [ZWL12]:

1. Calcular a matriz de distâncias para todos os pontos e ordenar os valores desta matriz de maneira ascendente linha por linha. Calcular a média de cada coluna (ε_vector). Então, $\varepsilon\_vector[i]$ é a distância média entre um ponto o i-ésimo ponto mais próximo;

2. Calcular o *MinPoints* usando ε_vector. Inicialmente é verificado quantos pontos existem dentro de ε considerando-se os valores da matriz de distância. Por exemplo, na primeira linha da matriz de distâncias é verificado quantas distâncias são menores do que o $\varepsilon\_vector[1]$. A mesma verificação é feita para a segunda linha e assim sucessivamente. Esses valores irão compor o $MinPoints\_vector$;

3. Aplicar o algoritmo *DBSCAN* usando ε_vector e $MinPoints\_vector$ como parâmetros. Para cada par de $\varepsilon\_vector[i]$, $MinPoints\_vector[i]$ o *DBSCAN* é aplicado e é armazenado o número de grupos resultante. Estes valores correspondem a uma matriz de parâmetros exemplificada na figura 5.6. A coluna estabilidade representa a diferença entre a quantidade de grupos;

4. Usando a matriz de parâmetros podemos descobrir quando o número de grupos estabiliza e localizar os valores ótimos de ε e *MinPoints*;

Seguindo a metodologia proposta, na segunda etapa é aplicado o algoritmo *DBSCAN* para os 100 conjuntos de dados gerados, considerando os parâmetros ε e *MinPoints* calculados para cada conjunto de dados, conforme explicado anteriormente. Em seguida, na terceira etapa, os resultados de agrupamento são avaliados utilizando todos os índices de validação de agrupamento descritos na seção 2.2.

Na quarta etapa, os índices foram transformados e normalizados para serem utilizados como entrada para a etapa de mineração. A figura 5.5 (à direita) mostra o agrupamento realizado pelo *DBSCAN* sobre o conjunto de dados original (à esquerda), considerando $\varepsilon = 0,493$ e $MinPoints = 6$. Os valores dos índices calculados foram: $DBI = 0,68797$, $S = 0,54938$, $\Gamma = 0,91735$, $C = 0,25997$, $J = 0,64087$, $R = 0,81476$ e $FM = 0,78114$.

k	EPS	MinPts	Qtd_Grupos	Estabilidade
1	0	-1	150	-
2	0,234192	-1	81	69
3	0,342545	1	49	32
4	0,423771	3	10	39
5	0,496816	4	6	4
6	0,550628	6	5	1
7	0,602069	7	3	2
8	0,648298	8	4	-1
9	0,694340	10	3	1
10	0,732673	11	3	0
11	0,770796	12	3	0
12	0,808510	13	3	0
13	0,843694	15	3	0

Quantidade de Grupos estabilizou

Figura 5.6: Exemplo da matriz de parâmetros gerada para obter os valores de ε e *MinPoints* para um *dataset*.

Os valores após transformação e normalização foram: $DBI = 0,91569$, $S = 0,81860$, $\Gamma = 0,92748$, $C = 0,83713$, $J = 0,64087$, $R = 0,81476$ e $FM = 0,78114$. Ó índice *Dunn* não foi considerado neste estudo de caso.

Na etapa 5, foi aplicado o algoritmo de regressão linear, disponível no *Weka*, para cada índice externo usando os índices internos como atributo alvo. Os parâmetros utilizados foram os mesmos do estudo de caso 1. Os modelos obtidos são descritos nas equações 5.5 - 5.7.

$$J = 0.1927\ S + 0.1712\ \Gamma + 0.0268\ DBI + 0.353 \tag{5.5}$$

$$R = 0.5798\ \Gamma + 0.2868 \tag{5.6}$$

$$FM = 0.2585\ S + 0.6064 \tag{5.7}$$

Analisando os resultados das equações 5.5 - 5.7 do segundo estudo de caso é possível observar que o índice *Silhouette* tem influência positiva sobre os índices externos *Jaccard* e *Fowlkes-Mallows*. O índice *Gamma* também é um índice importante para algoritmos baseados em densidade, uma vez que tem influência positiva sobre os índices *Jaccard* e

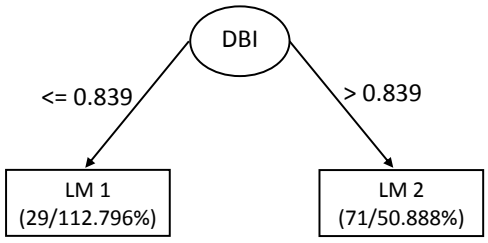

Figura 5.7: Árvore modelo usando *Jaccard* como classe.

Rand. Já o impacto do *C-index* é insignificante, uma vez que não aparece nos modelos de regressão linear, assim como o *DBI* que aparece mas como uma constante muito pequena.

Assim como no primeiro estudo de caso, também foi aplicado o algoritmo *M5* a fim de obter as árvores modelo para estimar os índices *Jaccard*, *Rand* e *Fowlkes-Mallows*. Para os índices *Rand* e *Fowlkes-Mallows* foi obtido árvores com nós isolados e modelos lineares iguais as equações de regressão 5.6 e 5.7 respectivamente. Para o índice *Jaccard* foi obtida a árvore modelo apresentada na figura 5.7. Esta árvore aprendeu os LMs descritos nas equações 5.8 e 5.9.

$$J = 0.0657\ S + 0.0584\ \Gamma + 0.0091\ DBI + 0.4464 \tag{5.8}$$

$$J = 0.0336\ S + 0.2484\ \Gamma + 0.0047\ DBI + 0.353 \tag{5.9}$$

A partir destas equações é possível notar que o índice *C-index* era insignificante, novamente não aparece nos LMs. O índice *DBI* foi importante para dividir as instâncias no nodo raiz, mas apresentou insignificante participação nos LMs. Semelhante a regressão linear, o índice *Gamma* tem contribuição positiva sobre o valor de *Jaccard*. A tabela 5.2 apresenta a comparação dos resultados de regressão linear e árvore modelo. É possível notar que os índices *Rand* e *Fowlkes-Mallows* têm resultados iguais para os dois algoritmos. A diferença é apenas para os resultados de *Jaccard*.

A análise dos modelos permite verificar que os índices internos de validação de agrupamento mais adequados para avaliar os conjuntos de dados gerados usando *DBSCAN* são *Silhouette* e *Gamma*.

Tabela 5.2: Resultado das métricas de avaliação para o algoritmo *DBSCAN*.

	$J(\%)$		$R(\%)$		$FM(\%)$	
	Reg. Linear	M5	Reg. Linear	M5	Reg. Linear	M5
Coeficiente de Correlação	65.48	67.59	72.59	72.59	64.29	64.29
Raiz do erro quadrático relativo	75.58	73.73	68.78	68.78	76.59	76.59

Capítulo 6

Validação

Este capítulo descreve o processo realizado a fim de validar os modelos obtidos nos experimentos apresentados no capítulo 5. Desta maneira, assim como nos estudos de caso a validação também foi divida em duas partes, uma para os modelos gerados utilizando o algoritmo de agrupamento *k-means* e outra para os modelos gerados utilizando o algoritmo de agrupamento *DBSCAN*.

Para a validar os modelos do estudo de caso 1, foram usados 5 *datasets* reais extraídos do site UCI[1] descrito na tabela 6.1. Para esta validação não foram gerados os gráficos dos conjuntos de dados para inspeção visual dos resultados de agrupamento, pois todos eles possuem mais que duas dimensões. Na Tabela 6.1 na primeira coluna tem-se o nome do *dataset*, como o mesmo está descrito no UCI. Nas colunas 2 e 3 são apresentados o total de instâncias e o total de atributos de determinado *dataset*. E por fim, a coluna 4 mostra o número de classes do *dataset* originalmente.

A validação foi realizada seguindo as mesmas etapas da metodologia descrita no capítulo 4, ou seja, foi feita a seleção dos *datasets*, de posse dos mesmos, foi executado o algoritmo de agrupamento *k-means* com k igual ao número de classes original (coluna 4 da Tabela 6.1). Com os dados já agrupados, os índices de validação foram calculados e transformados. A tabela 6.2 apresenta os valores dos índices internos e externos obtidos após este processo.

Conforme é possível observar na tabela 6.2 o único *dataset* que respeita o modelo gerado anteriormente é o *Habberman's*, pois bons valores para os índices externos impli-

[1] `https://archive.ics.uci.edu/ml/datasets.html`

Tabela 6.1: Descrição dos *datasets* utilizados na validação dos modelos usando o algoritmo k-means.

Dataset	**Instâncias**	**Atributos**	**Classes**
Balance Scale	625	4	3
Haberman's Survival	306	3	2
Hayes - Roth	160	5	3
Iris	150	4	3
Teaching Assistant Evaluation	150	5	3

caram em um valor alto para o índice *Gamma*. Para os outros *datasets* não foi possível encontrar uma relação entre os índices. Devido ao fato dos *datasets* representarem dados reais, não se tem a informação de suas características além de que utilizar a classe como gabarito para o agrupamento não é a melhor estratégia, pois os dados reais podem ser agrupados com base em suas características, sem respeitar a classe ao qual pertence.

Tabela 6.2: Índices de validação calculados para os *datasets* reais utilizados na validação dos modelos.

	DBI	**S**	**Γ**	**C**	**D**	**J**	**R**	**FM**
Balance Scale	0	0	0	0	1	0,34	0,34	0,58
Haberman's Survival	0,56	0,56	0,97	0,56	0	0,50	0,50	0,70
Hayes - Roth	1	1	0,99	1	0,32	0,33	0,34	0,57
Iris	0,89	0,96	1	0,88	0,65	0,34	0,35	0,58
Teaching Assistant Evaluation	0,40	0,37	0,98	0,40	0,38	0,34	0,35	0,58

Para validar os modelos do estudo de caso 2, foram usados 5 *datasets* sintéticos, utilizados na literatura [AS14, LLG, JPZ03], disponíveis no site[2]. A tabela 6.3 apresenta

[2] `http://cs.joensuu.fi/sipu/datasets/`

uma descrição dos *datasets*. Na primeira coluna tem-se o nome do *dataset*. Nas colunas 2 e 3 são apresentados o total de instâncias e o total de atributos de determinado *dataset*, sendo todos bi-dimensionais. E por fim, a coluna 4 mostra o número de classes do *dataset* originalmente.

Tabela 6.3: Descrição dos *datasets* utilizados na validação dos modelos usando o algoritmo *DBSCAN*.

Dataset	**Instâncias**	**Atributos**	**Classes**
Compound	399	2	6
Flame	240	2	2
Jain	373	2	2
Path	300	2	3
Spiral	312	2	3

Da mesma maneira que a validação anterior, esta validação seguiu as mesmas etapas da metodologia, utilizando os parâmetros adaptáveis para realizar o agrupamento, e ao final foram gerados gráficos dos conjunto de dados originais e do agrupamento resultante. As Fig. 6.1 e 6.2 apresentam os *datasets* utilizados e os agrupamentos resultantes.

Tabela 6.4: Índices de validação calculados para os *datasets* sintéticos utilizados na validação dos modelos.

	DBI	**S**	**Γ**	**C**	**J**	**R**	**FM**
Compound	0,24	0,40	1	1	0,80	0,94	0,90
Flame	0,94	0,81	0,53	0,54	0,54	0,54	0,73
Jain	1	1	0,69	0,28	0,95	0,95	0,97
Path	0,69	0,64	0,51	0,58	0,33	0,34	0,57
Spiral	0	0	0	0	1	1	1

Na tabela 6.4, são apresentados os valores dos índices para os *datasets* descritos anteriormente. Pode-se perceber, através de inspeção visual que os melhores resultados de

agrupamento foram obtidos nos *datasets* Compound, Jain e Spiral. Analisando os valores dos índices obtidos para os *dataset* Jain, validamos os modelos descritos na seção 5.2, já que eles mostram que os melhores índices para avaliar os agrupamentos são *Silhouette* e *Gamma* e isso se confirma. Já para o *dataset* Compound apenas o índice *Gamma* se confirma, pois foi obtido bom valor de *Gamma* e bom valor de *Jaccard*.

Os valores mais altos referente aos índices externos são dos *datasets* Compound, Jain e Spiral, conforme consta na tabela 6.4, tendo assim, relação direta com os valores dos índices internos, exceto para o Spiral que apesar do resultado do agrupamento ter ficado igual ao original, o valor dos índices internos é 0, pois foi o *dataset* com menor valor e também devido a sua forma, os centroides estão todos no mesmo lugar.

O contrário também é verdadeiro, segundo inspeção visual, os piores resultados de agrupamento foram obtidos nos *datasets* Flame e Path, que consequentemente apresentaram valores ruins nos índices externos e também refletiu nos índices internos *Silhouette* e *Gamma* que tiveram valores ruins se comparados aos valores referente aos *datasets* Compound, Jain. Conforme discutido anteriormente na seção 5.2, chegamos a conclusão que o *DBI* não é um bom índice para avaliar agrupamentos com característica de densidade, esta hipótese se confirma, pois como é possível observar, para os agrupamentos com resultados não tão bons, o *DBI* apresentou um valor alto, e para os agrupamentos com resultados bons o *DBI* apresentou um valor baixo.

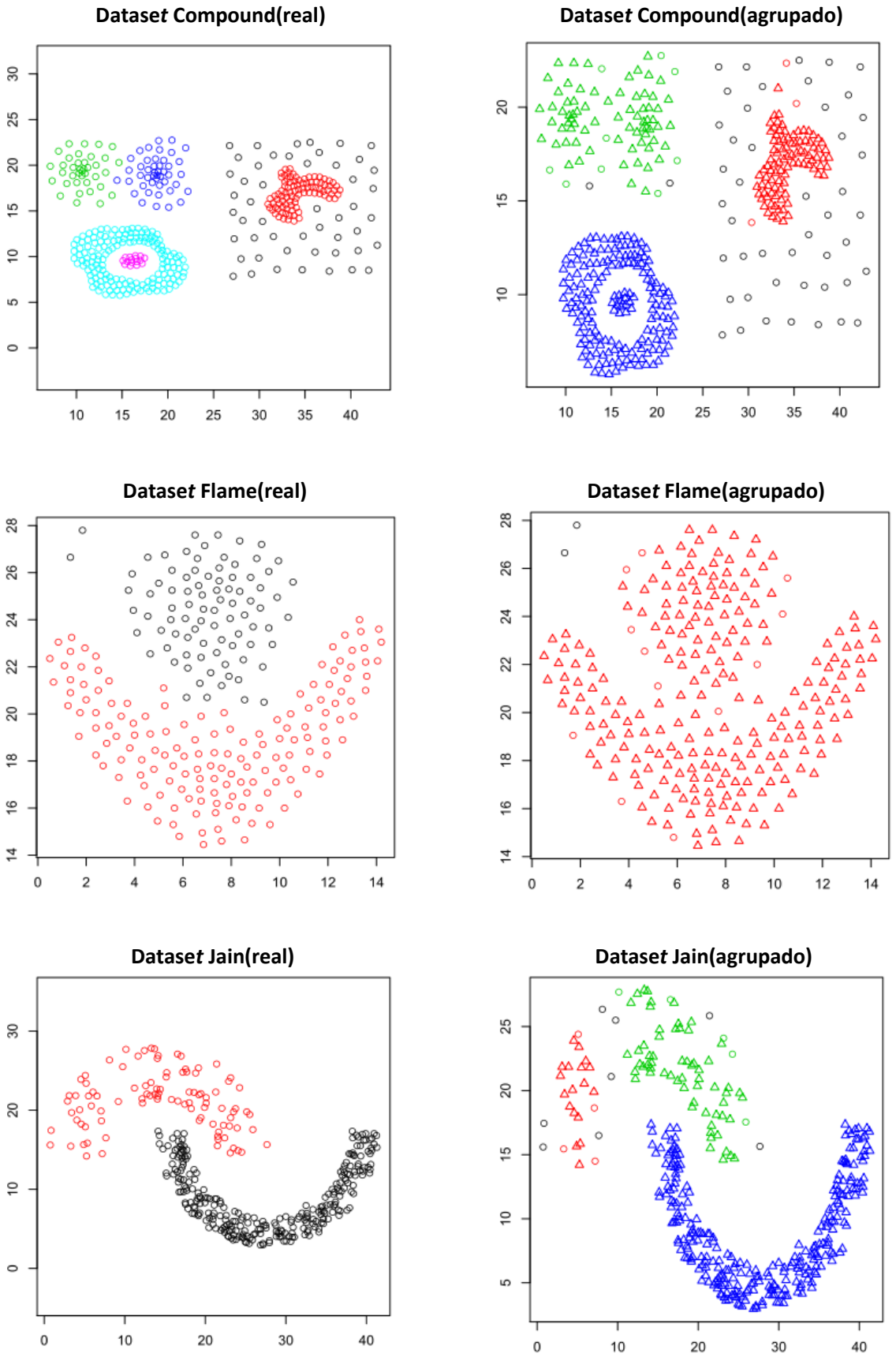

Figura 6.1: *Datasets* utilizados na validação dos modelos. A esquerda o conjunto de dados original e à direita o agrupamento gerado pelo algoritmo DBSCAN.

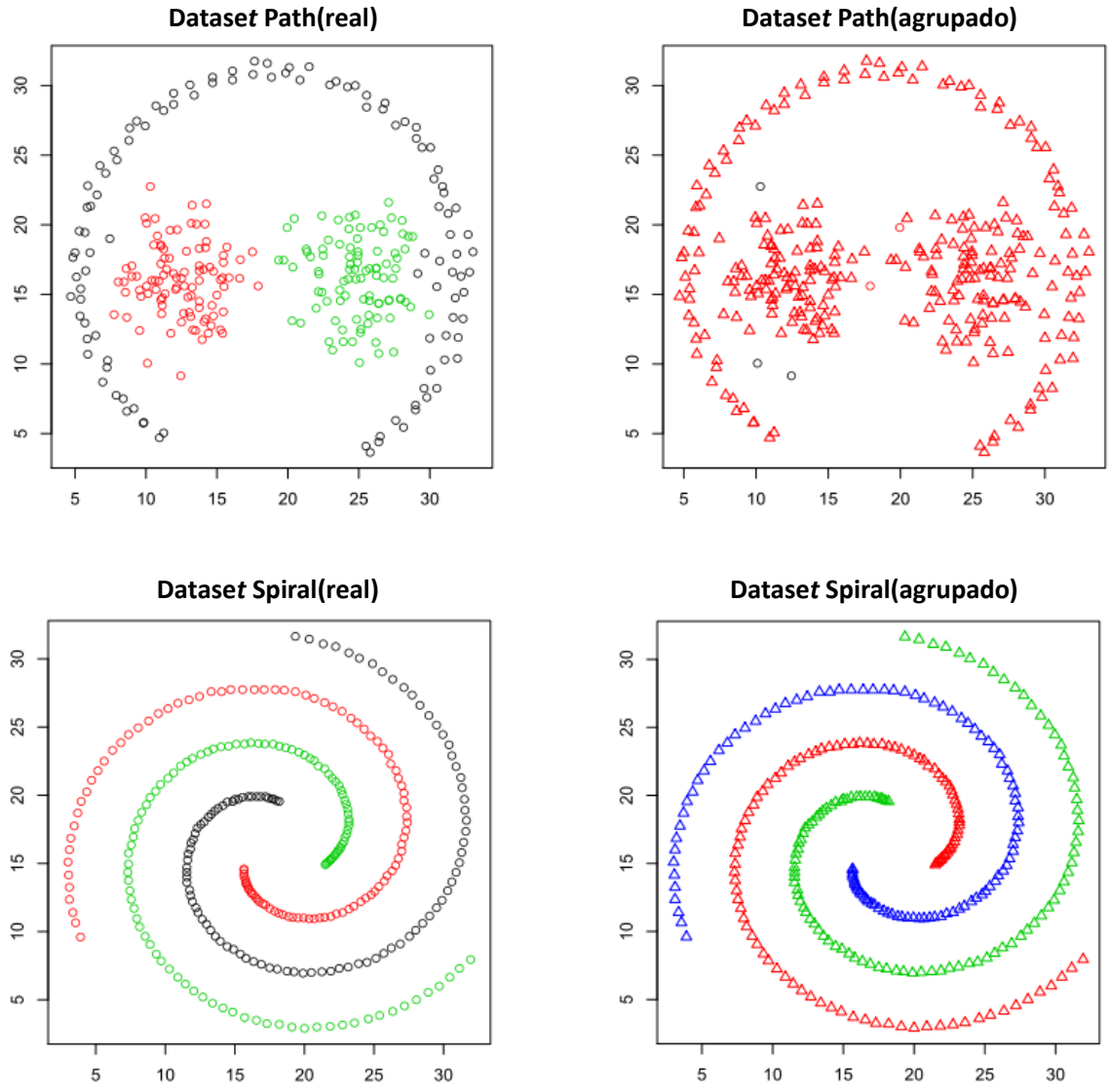

Figura 6.2: *Datasets* utilizados na validação dos modelos. A esquerda o conjunto de dados original e à direita o agrupamento gerado pelo algoritmo DBSCAN.

Capítulo 7

Conclusões e Trabalhos Futuros

Esta dissertação de mestrado propõe uma metodologia para a seleção dos índices internos de validação de agrupamento mais adequados. Foi investigado as relações entre os índices internos e externos a partir de um conjunto de modelos de regressão. Foram realizados experimentos com dois algoritmos de agrupamento distintos sobre os conjuntos de dados sintéticos gerados para este fim, usando diferentes configurações. A análise dos modelos de regressão permitiu a inferência do(s) índice(s) interno(s) mais adequado(s) para cada algoritmo de agrupamento.

Os resultados dos experimentos do estudo de caso 1 mostram que os índices internos *Silhouette* e *Gamma* são os mais adequados para avaliar os conjuntos de dados com com característica de compacidade usando *k-means*. Para o estudo de caso 2, os resultados mostram que os mesmos índices apresentaram melhor desempenho comparado aos outros, avaliando conjuntos de dados com múltipla densidade e utilizando o algoritmo *DBSCAN*.

Desta maneira, pode-se concluir que a metodologia proposta pode ser vista como uma guia para uma estratégia geral para a seleção do índice interno de validação mais adequado, no qual o método de agrupamento ou o algoritmo de regressão podem ser substituídos por outros mais eficazes ou eficientes em situações específicas.

Por fim, destaca-se como trabalhos futuros testar novos algoritmos com outras abordagens, como por exemplo algoritmos hierárquicos, gerar dados sintéticos com outras características, conforme abordado por Liu *et al.* [LLX+10] e incluir outros índices de validação de agrupamento internos e externos na metodologia.

Referências Bibliográficas

[AS09] Amparo Albalate and David Suendermann. A combination approach to cluster validation based on statistical quantiles. In *Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS'09. International Joint Conference on*, pages 549–555. IEEE, 2009.

[AS14] Loai AbdAllah and Ilan Shimshoni. Mean shift clustering algorithm for data with missing values. In *Data Warehousing and Knowledge Discovery*, pages 426–438. Springer, 2014.

[Ber06] P. Berkhin. A survey of clustering data mining techniques. In Jacob Kogan, Charles Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer Berlin Heidelberg, 2006.

[BH75] Frank B Baker and Lawrence J Hubert. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38, 1975.

[BL96] Michael JA Berry and Gordon Linoff. Data mining techniques for marketing, sales and customer support. john willey & sons. *Inc., 1997, 454 P*, 1996.

[BL97] Michael J Berry and Gordon Linoff. *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.

[CAC10] Santhana Chaimontree, Katie Atkinson, and Frans Coenen. Best clustering configuration metrics: Towards multiagent based clustering. In *Advanced Data Mining and Applications*, pages 48–59. Springer, 2010.

[Cas09] Marco Antonio Casanova. *Classificação automática de dados semi-estruturados*. PhD thesis, PUC-Rio, 2009. Figura do Kmeans.

[CCA+09] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

[CH74] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[CHY96] Ming-Syan Chen, Jiawei Han, and Philip S. Yu. Data mining: an overview from a database perspective. *Knowledge and data Engineering, IEEE Transactions on*, 8(6):866–883, 1996.

[DB79] David L Davies and Donald W Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.

[Dun74] Joseph C DunnĘ. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[Fac11] Katti et al. Faceli. *Inteligencia Artificial: Uma Abordagem de Aprendizagem de Maquina*. LTC, 2011.

[FM83] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.

[Goo63] NR Goodman. Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *Annals of mathematical statistics*, pages 152–177, 1963.

[GRS99] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: A robust clustering algorithm for categorical attributes. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 512–521. IEEE, 1999.

[Har01] Frank E Harrell. *Regression modeling strategies*. Springer Science & Business Media, 2001.

[HBV01] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

[Hin06] William W Hines. *Probabilidade e estatística na engenharia*. Livros Técnicos e Científicos, 2006.

[HKK05] Julia Handl, Joshua Knowles, and Douglas B Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005. Compactness,conectedness, spatial separation.

[HKP06] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining, southeast asia edition: Concepts and techniques*. Morgan kaufmann, 2006.

[HL76] Lawrence J Hubert and Joel R Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological bulletin*, 83(6):1072, 1976.

[HVB00] Maria Halkidi, Michalis Vazirgiannis, and Yannis Batistakis. Quality scheme assessment in the clustering process. In *Principles of Data Mining and Knowledge Discovery*, pages 265–276. Springer, 2000.

[HW79] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108, 1979.

[JMF99] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999. Classificação tirada do livro inteligencia artificial.

[JPZ03] Daxin Jiang, Jian Pei, and Aidong Zhang. Dhc: a density-based hierarchical clustering method for time series gene expression data. In *Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on*, pages 393–400. IEEE, 2003.

[KR87] Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

[KR09] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[KST09] Vipin KUMAR, Michael STEINBACH, and PN Tan. Introdução ao data mining-mineração de dados. *Ciência Moderna*, 2009. Figura do DBSCAN.

[LLG] Yonggang Lu, Ming Liu, and Rongmin Gao. A spectral clustering algorithm based on hierarchical method. *Agents and Data Mining Interaction*, page 111.

[LLX+10] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 911–916, Washington, DC, USA, 2010. IEEE Computer Society.

[MB02] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(12):1650–1654, 2002.

[MC85] Glenn W Milligan and Martha C Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[Mil81] Glenn W Milligan. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199, 1981.

[NH02] Raymond T. Ng and Jiawei Han. Clarans: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 14(5):1003–1016, 2002.

[Q+92] John R Quinlan et al. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. Singapore, 1992. Algoritmo M5.

[Ran71] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[Rou87] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[SEKX98] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, 2(2):169–194, 1998.

[TSK+06] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston, 2006.

[VCH10] Lucas Vendramin, Ricardo JGB Campello, and Eduardo R Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235, 2010.

[Wei05] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[WW96] Yong Wang and Ian H Witten. Induction of model trees for predicting continuous classes. 1996.

[XB91] Xuanli Lisa Xie and Gerardo Beni. A validity measure for fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 13(8):841–847, 1991.

[XW+05] Rui Xu, Donald Wunsch, et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[XW09] R Xu and DC Wunsch. Clustering. piscataway, 2009.

[ZWL12] Hongfang Zhou, Peng Wang, and Hongyan Li. Research on adaptive parameters determination in dbscan algorithm. *Journal of Information & Computational Science*, 9(7):1967–1973, 2012.