UMA AVALIAÇÃO DA UTILIZAÇÃO DE MATRIZES DE AFINIDADES NA VALIDAÇÃO DE AGRUPAMENTOS DE DADOS

RAFAEL XAVIER VALENTE

UMA AVALIAÇÃO DA UTILIZAÇÃO DE MATRIZES DE AFINIDADES NA VALIDAÇÃO DE AGRUPAMENTOS DE DADOS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Engenharia Elétrica.

Orientador: Prof. Antônio de Pádua Braga

Belo Horizonte

Outubro de 2013

Valente, Rafael Xavier

D1234p Uma Avaliação da Utilização de Matrizes de Afinidades na Validação de Agrupamentos de Dados / Rafael Xavier Valente. — Belo Horizonte, 2013
xviii, 71 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Prof. Antônio de Pádua Braga

1. Computação — Teses. 2. Redes — Teses.
I. Orientador.II. Título.

CDU 519.6*82.10

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

Dedico este trabalho aos meus pais e irmãs.

Agradecimentos

Ao orientador Prof. Antônio de Pádua Braga pelos valiosos ensinamentos e pela paciência ao lidar com todos os meus questionamentos. Por me ensinar que ciência se faz de forma gradual e que os pequenos passos são parte fundamental de grandes conquistas. Sou realmente muito grato por estes vários anos de convívio como aluno de mestrado e iniciação científica. Seu conhecimento técnico e humano foram definitivos para a minha formação profissional.

Aos membros da banca de defesa Prof. Felipe Campelo e Prof. Lenin Morais, pelas valiosas contribuições com este trabalho.

Aos demais professores e funcionários do PPGEE.

Aos amigos do LITC pelo entusiasmo na discussão de idéias e trocas de experiências.

Aos amigos do PPGEE, da Elétrica 2006/1 e meus eternos amigos de Viçosa, um muito obrigado pelo constante apoio. Ao amigo Rafael Cruz pela ajuda com a revisão e formatação do texto.

Aos colegas da Radix Engenharia e Software pelo incentivo e complacência, em especial João Kudo e Gessé Dafé, pelos ensinamentos e exemplo de profissionalismo.

Aos meus pais Francisco e Suely, pelo apoio incondicional. Não tenho dúvidas de que sem o incentivo e carinho deles, dificilmente teria chegado até aqui. Agradeço também às minhas irmãs Flávia e Raquel, por sempre estarem presentes e me ajudando de todas as formas possíveis.

À Leciane Giordano, pela dedicação, carinho e compreensão durante esta longa jornada.

À todos aqueles que de alguma forma participaram desse trabalho.

Resumo

Diferentemente de um problema de aprendizado de máquina supervisionado, onde busca-se encontrar uma função aproximadora a partir de um conjunto de dados rotulados, os problemas não supervisionados não possuem rótulos para guiar o processo de aprendizagem. Sendo assim, um critério deve ser adotado para o estabelecimento dos agrupamentos. O problema desta abordagem é que usualmente as funções objetivo utilizadas são degeneradas em relação ao número de agrupamentos, ou seja, a otimização da função alvo não provê o número ótimo de agrupamentos para determinado conjunto de dados. Neste caso, é realizado o particionamento dos dados para alguns valores de número de grupos e de acordo com alguma métrica as partições são avaliadas comparativamente para selecionar a quantidade ótima de grupos.

Neste trabalho, procura-se implementar uma nova métrica para a identificação do número de grupos de bases de dados que possuam agrupamentos compactos. Para tanto, utiliza-se da matriz de partição *fuzzy* obtida através do método Fuzzy C-Means (FCM) e calcula-se uma matriz de proximidade entre os elementos. A partir da matriz de proximidade são extraídas medidas estatísticas dos grupos para compor um índice comparativo, utilizado para estimar a partição que melhor se adequa à métrica proposta. Além disto, a matriz de proximidade possibilita ao usuário final visualizar os agrupamentos em duas dimensões para a validação dos resultados obtidos.

A fim de demonstrar a validade do método proposto, são realizados experimentos com bases de dados sintéticas e de referência. Os resultados obtidos para os casos controlados, onde a função geradora dos dados é conhecida, corroboram a hipótese da métrica desenvolvida. Já para as bases de referência, os resultados obtidos são comparados com outras métricas da literatura para a sua validação. Os resultados experimentais obtidos neste caso mostram que as abordagens apresentadas são consistentes com outras métricas bem conhecidas. Nestes casos, a matriz de proximidade apresentada é primordial para a validação dos resultados e visualização da conformidade da partição obtida com a estrutura intrínseca dos dados.

Palavras-chave: Aprendizado Não Supervisionado, Métodos de Agrupamentos, Índices de Validação de Agrupamentos, Matriz de Proximidade.

Abstract

Differently from a supervised machine learning problem, where one seeks to find an approximate function from a labeled dataset, the unsupervised problems does not contain any information to guide the learning process. In this case, a criterion must be adopted for the establishment of the partitions. The problem with this approach is that usually the objective functions commonly used are degenerated according to the number of groups, thus the simple optimization of the adopted criterion is not able to provide the optimum number of partitions for a given dataset. Therefore, partitions for differente number of groups are performed and according to another metric these partitions are comparatively evaluated to select the optimum number of groups.

In this work a new metric is proposed to identify the number of groups from datasets which can be clustered in compact clusters. In order to achieve the objective, the fuzzy partition matrix obtained from an algorithm like Fuzzy C-Means (FCM) is used to calculate a proximity matrix between the objects. Some factors are then calculated from the proximity matrix to compose the final index that will be used to compare the partitions and select the one which most agree which the proposed metric. Yet, the proximity matrix calculated makes it possible for the final user to visualize the clusters in two dimensions to validate the obtained results.

To demonstrate the validity of the proposed metric, experiments with synthetics and real datasets are provided. The results obtained for the controlled cases, where the datasets generator functions are known, show the validity of the development metric. For the real datasets, the obtained results are compared with other metrics to validate it. In this case, the results obtained show the new approach are consistent with other well-known metrics. In these cases, the proximity matrix presented are primordial to visualize the partitions and consequently validate it against the intrinsic structures of the datasets.

Keywords: Fuzzy Clustering, Validity Index, Proximity Matrix.

Lista de Figuras

Sumário

Capítulo 1

Introdução

Para realizar uma tarefa de aprendizado de máquina, deve-se tomar várias decisões acerca dos algoritmos, funções objetivo, métodos de validação, etc, que irão afetar os resultados finais do modelo. Particularmente, a função objetivo adotada tem um impacto majoritário nas características do modelo visto que de fato esta encarna o método por si só. Tal afirmação é muito clara nos problemas de aprendizado que possuem objetivos bem definidos, como aqueles que pertencem ao escopo do aprendizado supervisionado. Para tais problemas existe um conjunto de aprendizado indutivo $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ e uma família de funções aproximadoras $f(\mathbf{x}, \mathbf{w})$, onde o objetivo de aprendizagem tipicamente define-se em encontrar o vetor de parâmetros $\mathbf{w}$ que resulta em $f(\mathbf{x}_i, \mathbf{w}) \approx y_i \ \forall(\mathbf{x}_i, y_i) \in D$. Este problema tem um objetivo bem definido que é usualmente representado pela função de erro quadrático $\sum_i (y_i - f(\mathbf{x}_i, \mathbf{w}))^2$. Decisões sobre a utilização de métodos de regularização ou validação cruzada para selecionar os parâmetros finais do modelo irão com toda certeza afetar o modelo final obtido.

Problemas não supervisionados, por sua vez, são muito menos objetivos visto que não existem rótulos ou supervisores para guiar o processo de aprendizagem, o que faz com que o modelo final obtido seja muito mais dependente do que o usuário almeja quando escolhe um determinado método ou função objetivo. Algoritmos não supervisionados dependem de um conjunto de dados não rotulados $D_U = \{\mathbf{x}_i\}_{i=1}^{N_U}$ e um critério de agrupamento para obter uma determinada partição D_U. O critério mais conhecido e aceito na literatura é aquele representado pela função objetivo do algoritmo K-Means [MacQueen et al., 1967] J_{kmeans}, que resulta no problema de minimização quadrática das distâncias Euclidianas dos padrões $\mathbf{x}_i$ em relação aos centróides dos grupos. Este é o método usualmente utilizado quando o objetivo for obter agrupamentos compactos baseados em centróides no espaço Euclidiano. Para um determinado número de grupos k, a função objetivo J_{kmeans} resulta em uma partição estatisticamente

consistente com uma tendência central, a qual é particularmente útil para estimar as densidades de probabilidade $P(\mathbf{X})$ que geraram D_U. Entretanto, J_{kmeans} não é capaz de produzir o número ótimo de agrupamentos de acordo com o critério de compacticidade determinado pela função objetivo do algoritmo. De fato, J_{kmeans} é decrescente quando k aumenta, pois os centróides tendem aos próprios padrões $\mathbf{x}_i$ quando $k \rightarrow N_U$. Este conflito entre o valor da função J_{kmeans} e a compacticidade dos agrupamentos é inerente aos problemas de agrupamento, ocorrendo assim também em outros métodos baseados em centróides, como o K-Medoids [Rousseeuw & Kaufman, 1987] e o Fuzzy C-Means (FCM) [Bezdek, 1981].

A dificuldade em obter o número ótimo de agrupamentos ocorre porque a função objetivo que caracteriza o problema não pode ser utilizada para selecionar o número de agrupamentos. Para superar esta questão, vários métodos para a validação dos agrupamentos foram propostos na literatura [Pal & Bezdek, 1995; Wang & Zhang, 2007; Kaufman & Rousseeuw, 2009]. Ao invés de utilizar uma função objetivo diretamente, a maioria destes métodos é baseada em estatísticas intra-grupos e entre-grupos para representar os conceitos de compacticidade e separabilidade que fornecem a qualidade de uma determinada partição.

O problema destes métodos para a validação de agrupamentos é que eles sozinhos não são capazes de validar a premissa de que o conjunto de dados possua agrupamentos compactos. Nesta dissertação é proposto então um novo método para a validação de agrupamentos que possui uma fase de visualização posterior capaz de validar esta premissa. O método é baseado em estatísticas extraídas das matrizes de proximidade *fuzzy* [Loia et al., 2003] dos dados. Mesmo que a idéia tenha sido explorada em um conceito de agrupamentos *fuzzy*, o conceito pode ser estendido para outros métodos de agrupamento baseados em centróides. A noção de que as afinidades internas de um agrupamento devem ser mais fortes do que as afinidades entre os elementos de diferentes agrupamentos é representada pelas relações entre as submatrizes diagonais e as submatrizes não diagonais da matriz de proximidade, que pode ser representada em forma bloco diagonal [Queiroz et al., 2009] e possibilita a visualização das relações entre os padrões. A compacticidade de uma partição de agrupamentos é medida de acordo com as discrepâncias das magnitudes das submatrizes diagonais e não diagonais. Como será mostrado nas próximas seções, a nova métrica possui um máximo indicando o número ótimo de agrupamentos compactos.

Para melhor entendimento do leitor, o restante do texto encontra-se estruturado da seguinte forma: no capítulo 2 é realizada uma revisão da temática de análise de agrupamentos. É apresentado um panorama geral do processo e discutido suas principais características, como a representação dos dados, as métricas de distância, os tipos

de algoritmos e os métodos de validação. Em seguida, no capítulo 3, é apresentado o problema da escolha do número de grupos. Será mostrado que as funções objetivo dos principais algoritmos não possuem a capacidade de selecionar o número de grupos, além de discutir as métricas de validação que serão utilizadas como comparação nos experimentos. No capítulo 4 o método proposto é apresentado e são discutidos os fundamentos teóricos que sustentam o seu desenvolvimento. No capítulo 5 são apresentados os resultados experimentais da nova métrica em bases de dados sintéticas e reais. Além disto, será mostrado como a etapa de visualização proposta é importante para validar os resultados obtidos. Por fim, no capítulo 6 são descritas as principais conclusões deste trabalho, assim como as propostas de continuidade.

1.1 Contribuições

O método desenvolvido neste trabalho e os resultados obtidos estão publicados em [Valente et al., 2013].

Capítulo 2

Análise de Agrupamentos

O objetivo da análise de agrupamentos é revelar a estrutura intrínseca de um conjunto de dados através da formação de grupos estatisticamente consistentes. Busca-se portanto separar um conjunto de objetos em grupos onde os elementos de um mesmo grupo sejam similares entre si e os elementos de grupos distintos sejam diferentes.

Na literatura, é possível encontrar diversos trabalhos que tentam propor uma definição formal ao termo agrupamento, como em [Xu & Wunsch, 2008], [Gordon, 1999], [Jain, 2010] e [Kaufman & Rousseeuw, 1990]. Entretanto, o agrupamento é uma entidade subjetiva, visto que seu significado e interpretação requerem um conhecimento apurado do domínio a ser trabalhado, estando assim intimamente relacionado ao objetivo que se deseja alcançar com a análise. Embora não exista um consenso quanto à definição, é possível observar que busca-se sempre descrever os agrupamentos em termos da homogeneidade dos elementos dos grupos e da separação entre eles.

Desta forma, é razoável definir que um agrupamento ideal é um conjunto de pontos compacto e isolado, conforme colocado por [Jain, 2010], e que a análise de agrupamentos pode ser descrita como um conjunto de técnicas estatísticas para a descoberta se observações de uma população podem ser agrupadas através de comparações quantitativas de múltiplas características. A análise de agrupamentos pode ser então definida como: dada uma representação de N objetos, encontrar k grupos baseado em uma métrica de proximidade tal que as similaridades dos objetos de um mesmo grupo sejam altas enquanto as similaridades de objetos em grupos distintos sejam baixas.

Estas e muitas outras questões são discutidas em [Jain, 2010], bem como o seguinte conjunto de dilemas que são recorrentes em problemas de agrupamento:

- O que é um grupo?
- Quais características devem ser utilizadas?

- Os dados devem ser normalizados?
- Como deve ser definida a proximidade entre os objetos?
- Qual o número de agrupamentos presente nos dados?
- Qual método de agrupamento deve ser utilizado?
- Os agrupamentos encontrados são válidos?

Este capítulo tem como objetivo discutir as principais questões envolvidas no processo da análise de agrupamentos, assim como revisar as principais abordagens para lidar com os dilemas existentes. Desta forma, o restante do capítulo é dividido da seguinte forma: Inicialmente será dada uma visão geral do problema de agrupamentos, assim como a representação matemática do mesmo e os principais termos a serem utilizados no restante da dissertação. Em seguida é fornecida uma visão geral do processo de análise de agrupamentos, explicando a importância de cada fase e como estas se relacionam. Logo após é fornecida uma revisão sobre a representação dos dados, seus tipos e particularidades. A seguir é tratada a questão das métricas de proximidade, conceito importante que está diretamente relacionado à qualidade das partições obtidas. Com os conceitos já explicados, são descritos os vários tipos de algoritmos de agrupamento existentes, dando especial atenção ao tipo de problema que cada um visa resolver. Por fim, é realizada uma revisão dos métodos de validações de partições, contemplando os tipos existentes e suas características principais.

2.1 Representação dos Dados

A representação dos dados é um dos fatores que mais influênciam os resultados obtidos por um algoritmo de análise de agrupamentos [Jain, 2010]. Se a representação (escolha dos atributos) está de acordo com o objetivo da análise, as estruturas dos agrupamentos tendem a ser compactas e bem separadas entre si, tornando o problema mais simples de ser resolvido, permitindo que mesmo algoritmos mais simples sejam potencialmente capazes de encontrá-los. Porém, infelizmente não existe um procedimento padrão para a escolha da representação, e portanto, o conhecimento a priori do domínio a ser trabalhado tem papel decisivo na condução desse processo. Neste contexto, busca-se nessa seção apresentar os tipos de dados que ocorrem em problemas reais, assim como os tratamentos existentes para cada caso.

Suponha que existam N objetos a serem agrupados, que podem ser por exemplo pessoas, flores, palavras, países ou qualquer outro objeto. Os algoritmos de agru-

pamento tipicamente possuem dois formatos de entrada: na primeira os objetos são representados por vetores multidimensionais contendo d dimensões, ou características, ou atributos, como altura, idade, sexo, cor, etc. Estas medidas são normalmente arranjadas em uma matriz $N \times p$, onde cada linha corresponde a um objeto e as colunas correspondem aos atributos; na segunda opção o método recebe uma coleção de proximidades que devem existir para todos os pares de objetos. Estas proximidades formam uma matriz de tamanho $N \times N$ e podem ser dadas de duas formas: dissimilaridades ou distância (que mede quanto um objeto é diferente do outro) e similaridades (que mede o quanto um objeto é parecido com o outro) [Jain, 2010]. Em suma, tem-se que os dados podem ser organizados na forma de uma matriz $N \times d$, onde as linhas correspondem aos objetos (ou observações, padrões, registros, etc) e as colunas correspondem às características (atributos, dimensões) [Kaufman & Rousseeuw, 1990].

Sobre a escolha das p dimensões a serem utilizadas, não existe fundamentação teórica que sugira como deve ser feita a escolha dos padrões e características a serem utilizadas para situações específicas. De fato, o processo de geração dos dados não é diretamente controlado e a função geradora $P(X)$ não é conhecida [Jain et al., 1999]. Assim, o papel do usuário ao escolher a representação dos padrões é de coletar os fatos e conjecturas sobre os dados, de preferência realizando um processo de extração e seleção de características. Ainda assim, uma investigação criteriosa das características disponíveis pode mostrar que a partir da transformação dos dados é possível melhorar significativamente os resultados das análises de agrupamentos. Uma boa representação dos padrões usualmente resulta em agrupamentos simples e de fácil compreensão, enquanto uma representação ruim pode resultar em agrupamentos complexos cuja estrutura é impossíveis de distinguir [Jain et al., 1999].

É também um procedimento importante selecionar apenas as características mais descritivas e descriminativas do conjunto de entrada para serem utilizadas nas análises. Técnicas de seleção de características identificam um subconjunto das características existentes, enquanto extração de características calculam novas características a partir das originais. Em ambos os casos, o objetivo é melhorar os agrupamentos obtidos ou melhorar a eficiência computacional. O tópico de seleção de características é bem explorado no ramo de classificação de padrões [Duda et al., 2000], [Guyon & Elisseeff, 2003] e [Guyon, 2006]. Entretanto, no ramo de análise de agrupamentos, onde não existem rótulos, o processo de seleção de características deve ser realizado para cada finalidade, e envolve normalmente um processo de tentativa e erro onde vários subconjuntos de características são selecionados, os padrões resultantes são agrupados, e o resultado é avaliado a partir de uma métrica de validação [Jain et al., 1999]. Em contraste, alguns métodos de extração de características como o PCA [Jolliffe, 2005]

não dependem dos rótulos e podem ser utilizados diretamente.

Um padrão pode representar tanto um objeto físico (carro, cadeira, bola) ou uma noção abstrata de uma entidade (estilo de escrita, etc). Como observado previamente, os padrões são convenientemente representados por vetores multidimensionais, onde cada dimensão é uma única característica [Duda et al., 2000]. Conforme descrito em [Jain et al., 1999] e [Gan et al., 2007], estas características podem ser tanto quantitativas quanto qualitativas. Dentre os vários tipos de dados existentes, os mais utilizados são:

- Características Contínuas
- Características Binárias
- Características Nominais
- Características Ordinais

Nas subseções a seguir cada tipo de dados será explicado com mais profundidade, além de fornecer os métodos e transformações úteis que podem ajudar no processo de agrupamento.

2.1.1 Características Escalares

As características escalares são representadas por valores contínuos escalares, e servem para representar atributos como altura, peso, temperatura, idade, custo, etc. Neste caso é importante que a dimensão em análise siga uma escala linear, ou seja, os intervalos devem manter a mesma importância por toda a escala. Como exemplo, pode-se citar a diferença entre as alturas de duas pessoas com $1,50m$ e $1,60m$ que é igual a diferença das alturas de duas pessoas com $1,70m$ e $1,80m$.

As características escalares são usualmente obtidas por processos de medição, o que faz com que as mesmas tenham unidade de medida. Portanto, como no exemplo anterior em que as alturas foram dadas em metros, as mesmas poderiam ter sido dadas em centímetros ou milímetros. Este processo de escolha das unidades de medidas pode distorcer diretamente a análise de agrupamentos, visto que ao aumentar os valores absolutos de um determinado atributo, o mesmo passa a ter maior importância no processo de agrupamento [Kaufman & Rousseeuw, 1990].

Uma maneira de acabar com a dependência da escolha das unidades de medidas, e consequentemente não impor diferentes níveis de importância aos atributos, é normalizar os dados. A operação de normalização possui a capacidade de converter as medidas originais de uma característica para variáveis adimensionais. Para isto, deve-se

calcular as médias amostrais de cada uma das características que se deseja normalizar, conforme a Equação 2.1.

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf}) \tag{2.1}$$

Em seguida, deve-se calcular uma medida de dispersão dos dados. Esta medida pode ser dada pela diferença entre os valores máximos e mínimos de cada dimensão, mas a medida usualmente utilizada é o desvio padrão amostral, dado pela Equação 2.2.

$$std_f = \sqrt{\frac{1}{n-1}\{(x_{1f} - m_f)^2 + (x_{2f} - m_f)^2 + \cdots + (x_{nf} - m_f)^2\}} \tag{2.2}$$

Entretanto, devido à sua natureza quadrática, o desvio padrão é muito afetado pela presença de *outliers* nos dados [Kaufman & Rousseeuw, 1990]. Assim, uma medida mais robusta pode ser utilizada para minimizar este efeito. Um exemplo de métrica a ser utilizada é o *Desvio Absoluto Médio*, definido pela Equação 2.3.

$$s_f = \frac{1}{n}\{\|x_{1f} - m_f\| + \|x_{2f} - m_f\| + \cdots + \|x_{nf} - m_f\|\} \tag{2.3}$$

Neste caso é possível observar que a contribuição de cada padrão na composição da medida de dispersão é proporcional apenas ao valor absoluto da diferença $\|x_i - m_i\|$, diminuindo a influência dos *outliers*.

Assumindo que a dispersão calculada é não nula (caso contrário todos os registros possuem o mesmo valor para a característica, o que não agrega nenhum tipo de informação, devendo portanto ser removida da análise), os novos valores normalizados são definidos conforme

$$z_{if} = \frac{x_{if} - m_f}{s_f} \tag{2.4}$$

e são conhecidos como *z-scores* [Kaufman & Rousseeuw, 1990]. Estas medidas são adimencionais porque tanto o numerador quanto o denominador são expressos na mesma unidade. Por construção, o valor z possui média nula e dispersão unitária. Quando a normalização é utilizada, o vetor correspondente aos dados originais deve ser deixado de lado e substituído pelos novos valores calculados em todas as análises a serem realizadas.

A discussão realizada anteriormente pode passar a impressão de que a normalização dos dados é benéfica em todas as situações. Todavia, esta se apresenta apenas como uma opção que pode ser ou não útil em determina aplicação. Não raro as variáveis podem possuir valor absoluto significativo, apresentando dominância intrínseca

ao problema, não devendo portanto serem normalizadas. Além disso, muitas vezes a normalização pode destruir a estrutura original dos dados.

Sendo assim, é importante ressaltar que a normalização dos dados não resolve o problema da representação de dados numéricos. De fato, a escolha das unidades de medidas e escalas abordam a temática de ponderação das variáveis. Sempre que uma variável é expressa em unidades menores, o alcance resultante do atributo é maior, o que terá um efeito maior na estrutura resultante do agrupamento.

Entretanto, ainda que a normalização dos dados não resolva o problema da representação de dados numéricos, a mesma se apresenta como uma boa tentativa inicial para quando não existirem outras informações dos dados em questão. Assim, deve-se manter sempre em mente o que foi discutido anteriormente, referente a fato que intrinsecamente algumas variáveis são mais importantes que outras em aplicações específicas e a atribuição dos pesos deve ser baseado em conhecimento subjetivo sobre o negócio. Por fim, o dilema da normalização é inevitável no paradigma atual de análise de agrupamentos, deixando a escolha da decisão final para o usuário.

2.1.2 Características Binárias

Os atributos binários são aqueles que possuem apenas dois estados, sendo normalmente utilizados para representar a presença ou ausência de algum fator. Como exemplo tem-se a identificação de gênero Masculino/Feminimo, Fumante/Não-Fumante, Sim/Não para uma questão específica, etc. A codificação destas características são normalmente 0 e 1, apesar de que quaisquer números ou símbolos como S/N possam ser utilizados.

Apesar da representação simples, é crucial salientar que existem dois tipos de variáveis binárias que dependem fortemente do contexto em que são aplicadas, conforme explicado em [Kaufman & Rousseeuw, 1990] e [Gan et al., 2007]. Tomemos como exemplo uma variável do tipo sexo, tal que os estados possíveis são Masculino e Feminino. Ambos possuem o mesmo peso, não havendo diferença sobre qual estado será codificado como 0 ou 1. Esta variável é portanto do tipo simétrica, pois o problema independe da codificação.

Mas o cenário muda drásticamente quando a variável binária é do tipo assimétrica, situação na qual os estados não possuem a mesma importância. Exemplos de variáveis assimétricas são dados por atributos que identificam a presença ou ausência de uma determinada situação rara. Como exemplo pode-se citar a presença de uma doença incomum. O fato de dois indivíduos estarem contaminados por esta doença é muito mais informativo do que o fato de dois outros indivíduos quaisquer não estarem contaminados.

Por fim, devido à simplicidade da representação das variáveis binárias, não há a necessidade de realizar transformações nos dados, mantendo o foco centrado na análise de simetria da característica.

2.1.3 Características Nominais

Quando um determinado atributo é qualitativo e possui mais de dois estados, define-se o conceito de atributo nominal [Kaufman & Rousseeuw, 1990]. Neste tipo de atributos, os estados são codificados com a própria descrição do estado ou por números que independem dos valores.

Para ilustrar o conceito de variável nominal, toma-se como exemplo um atributo do tipo nacionalidade, onde os estados são representados pelos valores: Brasileiro=1; Argentino=2; Francês=3. Neste caso um número maior ou menor não possui qualquer significado sobre a importância de determinado estado.

2.1.4 Características Ordinais

Os atributos ordinais são muito parecidos com os atributos nominais, mas com a diferença de que a ordem dos estados é significativa [Kaufman & Rousseeuw, 1990], o que retira a arbitrariedade da codificação $1, \cdots, M$. A distância entre dois estados passa a ser maior quanto maior for sua diferença entre os códigos, fazendo com que a discrepância entre os estados 1 e M seja a maior.

As variáveis ordinais são muito úteis para registro de opiniões e julgamentos que não podem ser medidos objetivamente. Como exemplo você pode pedir alguém para avaliar uma música em uma escala de 1 a 5, caso em que a preferência das pessoas é modelado como uma variável ordinal de 5 estados.

Outra situação em que os tipos ordinais são necessários é na discretização de quantidades escalares. Em situações deste tipo todo o intervalo de valores possíveis da variável é particionado em um número finito de estados. Este tipo de técnica é utilizada por exemplo nas categorias do UFC e nas temidas categorias de imposto de renda no Brasil.

Por fim, ocorrem casos onde as informações medidas são extremamente ruidosas e não confiáveis. Pode-se então criar um índice de ordenação para a variável para que os erros de medição tenham uma influência menor nos resultados finais do agrupamento [Kaufman & Rousseeuw, 1990].

2.1.5 Outros Tipos

Em bases de dados reais, podem ocorrer ainda outros tipos de atributos. Em [Gan et al., 2007] são descritos os tipos de dados transacionais e simbólicos, além dos dados que caem nas categorias de séries temporais. Já em [Jain et al., 1999] são descritas as variáveis estruturadas representadas por árvores. Por fim tem-se as variáveis de razão, conforme explicado em [Kaufman & Rousseeuw, 1990].

2.2 Métricas de Proximidade

Conforme visto anteriormente, os agrupamentos são formados por grupos de objetos similares entre si, enquanto objetos em grupos distintos não o são. Assim, surge naturalmente a questão de como definir a proximidade dos objetos, ou seja, como avaliar quantitativamente a dissimilaridade (distância) ou a similaridade entre um par de objetos, um objeto e um grupo ou ainda entre um par de grupos.

Neste contexto, o termo *proximidade* pode ser utilizado como uma generalização dos termos similaridade e dissimilaridade. A utilização de uma métrica ou outra é fortemente dependente do problema e dos tipos de atributos envolvidos. O objetivo desta seção é apresentar a definição matemática formal do que constitui uma métrica, além de revisar as métricas da literatura usualmente utilizadas em problemas reais.

2.2.1 Definições

A seguir são fornecidas as definições formais de métricas de dissimilaridade $D(\mathbf{x}_i, \mathbf{x}_j)$ e similaridade $S(\mathbf{x}_i, \mathbf{x}_j)$.

2.2.1.1 Dissimilaridade

Segundo [Xu & Wunsch, 2008] e [Kaufman & Rousseeuw, 1990], para uma função $D(\mathbf{x}_i, \mathbf{x}_j)$ ser considerada uma métrica de dissimilaridade, as seguintes propriedades devem ser satisfeitas :

1. Simetria:
$$D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_j, \mathbf{x}_i) \tag{2.5}$$

2. Positividade:
$$D(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \quad \forall\, \mathbf{x}_i, \mathbf{x}_j \tag{2.6}$$

3. Reflexividade:

$$D(\mathbf{x}_i, \mathbf{x}_j) = 0 \quad \Longleftrightarrow \quad \mathbf{x}_i = \mathbf{x}_j \tag{2.7}$$

4. Desigualdade Triangular:

$$D(\mathbf{x}_i, \mathbf{x}_j) \leq D(\mathbf{x}_i, \mathbf{x}_k) + D(\mathbf{x}_k, \mathbf{x}_j) \quad \forall\, \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \tag{2.8}$$

2.2.1.2 Similaridade

Para o caso de funções de similaridade $S(\mathbf{x}_i, \mathbf{x}_j)$, as propriedades a serem satisfeitas segundo so trabalhos de [Xu & Wunsch, 2008] e [Kaufman & Rousseeuw, 1990] são:

1. Simetria

$$S(\mathbf{x}_i, \mathbf{x}_j) = S(\mathbf{x}_j, \mathbf{x}_i) \tag{2.9}$$

2. Positividade:

$$0 \leq S(\mathbf{x}_i, \mathbf{x}_j) \leq 1 \quad \forall\, \mathbf{x}_i, \mathbf{x}_j \tag{2.10}$$

3. Reflexividade:

$$S(\mathbf{x}_i, \mathbf{x}_j) = 1 \quad \Longleftrightarrow \quad \mathbf{x}_i = \mathbf{x}_j \tag{2.11}$$

2.2.2 Métricas

A métrica mais conhecida e utilizada na literatura é a distância Euclidiana, representada pela Equação 2.12 e que representa a distância geométrica entre dois vetores no espaço de características.

$$D_{Euclidiana}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{2.12}$$

O problema é que, para casos em que nem todos os atributos são escalares, as distâncias baseadas em medidas puramente geométricas não representam corretamente o problema.

Desta maneira, diversas métricas para os diferentes tipos de características foram propostas na literatura. Nos trabalhos [Gower & Legendre, 1986], [Kaufman & Rousseeuw, 1990] e [Pękalska & Duin, 2005] é possível encontrar uma grande variedade de métricas para os mais variados tipos de atributos, incluindo outras métricas para características escalares e métricas para variáveis binárias simétricas e assimétricas, além de métodos de transformação para as variáveis nominais e ordinais. Para padrões

que possuam vários tipos de características, a métrica proposta em [Ichino & Yaguchi, 1994] pode ser utilizada para o computo das distâncias entre os pares de objetos.

Por fim, tem-se também as distâncias calculadas a partir de funções de *kernel*, descritas detalhadamente em [Shawe-Taylor & Cristianini, 2004] e [Filippone et al., 2008]. Os métodos que utilizam estes tipos de funções são conhecidos como *Spectral Clustering*.

2.3 Algoritmos

Devido à grande gama de problemas existentes no contexto de análise de agrupamentos, uma infinidade de algoritmos para os mais diferentes propósitos foram propostos na literatura. Porém, de uma forma geral, os algoritmos de agrupamento podem ser divididos entre os métodos particionais e hierárquicos. O restante da discussão desta seção é baseada na revisão do tema realizada em [Xu et al., 2005].

Os algoritmos particionais têm como objetivo dividir o espaço de características em áreas de influência baseadas na distância aos centros das partições. Assim, cada agrupamento pode ser representado pelo centróide de seus elementos. O algoritmo clássico que representa este tipo de modelos é o K-Means, descrito inicialmente em [MacQueen et al., 1967]. Além disto, os métodos particionais podem ser classificados como *Hard* ou *Soft*. Nos modelos *Hard* os elementos a serem agrupados pertencem a apenas um grupo, enquanto nos algoritmos *Soft* cada elemento pertence a todos os grupos com um determinado nível de pertinência. Estes métodos são baseados nas ideias de lógica *fuzzy*, proposta em [Zadeh, 1965]. O algoritmo que representa esta classe de modelos é o Fuzzy C-Means (FCM) [Bezdek, 1981].

Já os algoritmos hierárquicos têm como objetivo organizar os dados em uma estrutura hierárquica de acordo com as similaridades entre os padrões. Desta maneira, o resultado do processo de agrupamento hierárquico é dado por uma árvore binária ou dendrograma representando o particionamento. Diferentemente dos métodos particionais, as técnicas hierárquicas não necessitam do número de grupos fornecido à *priori*. Depois de formada a árvore dos resultados, cabe ao usuário escolher em qual nível da mesma deve ser realizado o corte, procedimento que está intimamente ligado ao objetivo final almejado no processo de agrupamento.

Além dos métodos tradicionais já apresentados, existem também na literatura outras abordagens para o problema de agrupamentos. Como exemplo pode-se citar os algoritmos baseados em grafos [Schaeffer, 2007], baseados em modelos estatísticos [Fraley & Raftery, 2002], representado pelo algoritmo *Expectation Maximization*

[McLachlan & Krishnan, 2007], baseados em densidade [Ester et al., 1996], baseados em redes neurais [Kohonen, 1990], funções de *kernel* [Ng et al., 2002] e ainda algoritmos baseados na teoria do aprendizado estatístico descrito em [Vapnik, 1998], como o método proposto por [Ben-Hur et al., 2002].

2.4 Métricas de Validação

Um dos maiores desafios em problemas de análise de agrupamentos é avaliar os resultados obtidos sem nenhum tipo de informação auxiliar. Uma prática usual é utilizar métricas de validação para mensurar a qualidade das partições obtidas. Estas métricas podem ser externas ou internas.

As métricas externas são dependentes de outras informações que não os dados a serem agrupados. Como exemplo pode-se citar a imposição de uma estrutura pré-definida, como agrupamentos circulares ou em espiral, e rótulos de classe para bases de dados que os contêm. É importante observar que os rótulos de classes não possuem obrigatoriamente uma coerência estatística, podendo desvirtuar completamente o processo de agrupamento.

Já as métricas internas dependem apenas das informações intrínsecas aos dados, como os próprios padrões ou os centróides resultantes da aplicação de um algoritmo particional. Nesta dissertação somente métricas internas serão consideradas.

Um comparativo entre a utilização de métricas internas e externas pode ser encontrado em [Rendón et al., 2011].

2.5 Conclusões do Capítulo

Neste capítulo foi realizada uma revisão dos aspectos teóricos e práticos que permeiam a temática da análise de agrupamentos. Foram discutidos os principais dilemas [Jain, 2010] e as maneiras de abordá-los. Para tal, foram fornecidas explicações sobre os principais tipos de atributos que ocorrem em problemas de agrupamento e suas respectivas tratativas. Além disto, foram apresentadas as definições formais das métricas de proximidade, assim como os tipos existentes e os problemas relacionados. Por fim, foi realizada a revisão dos tipos de algoritmos existentes e definidos os tipos de validação de agrupamentos, tema que será extensivamente explorado no capítulo 3.

Capítulo 3

O Problema da Seleção do Número de Grupos

Diferentes ideias e métodos foram propostos na literatura para tentar determinar um critério global que forneça uma maneira objetiva de descobrir o número de agrupamentos. As revisões realizadas por [Rand, 1971], [Milligan & Cooper, 1985], [Halkidi et al., 2001] e [Bouguessa et al., 2006] exploram o tema de uma maneira geral, fornecendo informações acerca da importância do processo de validação e os indicadores que podem ser utilizadas para este fim.

Dentre os diferentes critérios existentes, a escolha da teoria de informação proposta em [Shannon & Weaver, 1949] é uma alternativa natural para validar processos de agrupamento, visto que a mesma busca quantificar a quantidade e a qualidade da informação existente. Neste contexto, pode-se citar os trabalhos [Celeux & Soromenho, 1996], [Gokcay & Principe, 2002], [Jenssen et al., 2003] e [Ayad & Kamel, 2008] como exemplos de tentativas em estimar o número de grupos através de medidas de entropia da informação.

Outra alternativa muito utilizada para deteminar o número de grupos é a construção de métodos baseados em estatística. Neste escopo são exploradas várias abordagens: nos trabalhos [Fred, 2001], [Ayad & Kamel, 2010], [Ghosh & Acharya, 2011] e [Vega-Pons & Ruiz-Shulcloper, 2011], a ideia da construção de *ensembles* é utilizada para criar um sistema de votos, onde diferentes combinações de métodos de agrupamentos e validação fornecem palpites para o valor do número de grupos e o valor mais votado é escolhido como resposta final do modelo; já os trabalhos [Roth et al., 2002], [Lange et al., 2004] e [Pascual et al., 2010] exploram a idéia da reamostragem dos dados para selecionar as partições consideradas mais estáveis de acordo com o número de grupos; por fim, a ideia de explorar testes estatísticos é abordada em [Hamerly, 2007].

Além disto, existem os métodos baseados em grafos [Pal & Biswas, 1997], [Günter & Bunke, 2003] e [Barzily et al., 2009], assunto que possui uma teoria extensa e já estabelecida na literatura. Temos também métodos para lidar exclusivamente com problemas de alta dimensionalidade, conforme descrito em [Kim et al., 2004a]. Ainda no contexto de grandes volumes de dados, existem os métodos baseados em visualização 2D das partições, conforme implementado em [Huang & Lin, 2000], [Huang et al., 2001], [Ng & Huang, 2002], [Chen & Liu, 2003] e [Chen & Liu, 2004]. Estes métodos são baseados no algoritmo *FastMap* [Faloutsos & Lin, 1995] e se apresentam como uma alternativa interessante para a estimativa do número de grupos.

Por fim, existem as métricas baseadas nos conceitos de separabilidade e compacticidade dos agrupamentos, que possuem uma literatura muito extensa. Para o caso de partições rígidas, onde cada elemento pertence a apenas um grupo, pode-se citar como referência os trabalhos [Rousseeuw, 1987], [Tibshirani et al., 2001], [Sun et al., 2004] e [Tibshirani & Walther, 2005]. Para os casos de partições *fuzzy*, onde os padrões possuem níveis de pertinência em relação aos agrupamentos, tem-se como principais métodos aqueles descritos em [Bezdek & Pal, 1995], [Bezdek & Pal, 1998], [Kwon, 1998], [Boudraa, 1999], [Kim et al., 2003], [Kim et al., 2004b], [Pakhira et al., 2004], [Tang et al., 2005], [Wang & Zhang, 2007] e [Izakian & Pedrycz, 2013].

Devido ao extenso número de técnicas e algoritmos propostos no contexto de análise e validação de agrupamentos, faz-se necessário limitar a extensão da revisão no presente trabalho. Assim, foi escolhido o algoritmo FCM [Bezdek, 1981] para gerar as partições a serem analisadas na dissertação. O restante deste capítulo tem como objetivo descrever o algoritmo FCM e demonstrar que o mesmo não é capaz de estimar o número de agrupamentos sozinhos. Além disto, são apresentadas as métricas de validação para agrupamentos *fuzzy* que serão utilizadas nos experimentos.

3.1 Fuzzy C-Means (FCM)

O algoritmo de agrupamento FCM [Bezdek, 1981], caracterizado pela função objetivo da Equação 3.1, estende o conceito de partições rígidas do método K-Means [MacQueen et al., 1967], introduzindo o conceito de funções de pertinência e partições *fuzzy*.

$$J_{fcm}(U,V) = \sum_{i=1}^{N_U}\sum_{j=1}^{c} u_{ij}^{m}\|\mathbf{x}_i - \mathbf{v}_j\| \tag{3.1}$$

onde c é o número de clusters, N_U é o número de padrões do conjunto de dados utilizado na análise, u_{ij} é a pertinência do padrão $\mathbf{x}_i$ em relação ao cluster i, $\mathbf{v}_j$ é o centróide do

cluster j e $||\mathbf{x}_i - \mathbf{v}_j||$ é a distância Euclidiana entre o padrão $\mathbf{x}_i$ e centróide $\mathbf{v}_j$.

Mesmo que o FCM seja robusto comparado com outros algoritmos de agrupamento, este ainda sofre de dois grandes problemas:

- por se tratar de um problema de otimização combinatória, é possível mostrar que o mesmo se caracteriza como um problema NP-Completo. Assim, as heurísticas usualmente utilizadas fornecem soluções sub ótimas em relação à função objetivo descrita na Equação 3.1, não existindo garantia de que o mínimo global da função seja sempre encontrado quando o algoritmo converge.

- assim como o K-Means, o número de clusters c deve ser fornecido à *priori*, o que implica que o usuário deve prover informações externas ao modelo.

Embora uma solução polinomial para um problema NP-Completo seja uma questão em aberto na literatura, várias abordagens foram propostas para inferir o valor de c. Nestes casos uma métrica de validação de agrupamentos é calculada para um intervalo de valores de c e, baseado no critério do índice utilizado, um valor de c é escolhido. Esta é uma tarefa complicada, visto que a não ser que a função de densidade de probabilidade $P(\mathbf{X})$ que gerou os dados seja conhecida, não é possível realizar afirmações em relação à estrutura real do conjunto de dados, mesmo que os rótulos sejam conhecidos.

3.2 O Problema da Função Objetivo

Uma vez que a função objetivo do FCM representada pela Equação 3.1 é baseada apenas nas distâncias dos elementos aos centros dos grupos, pode-se mostrar que a mesma não consegue identificar a quantidade de agrupamentos em uma base de dados. Isto ocorre pois sempre que o valor de c aumenta, são criados mais agrupamentos, até o caso limite onde o número de grupos é o mesmo número de padrões, situação na qual todos os elementos são os próprios centros dos grupos aos quais pertencem e o somatório das distâncias da função objetivo J_{FCM} se anula.

Para ilustrar este fato, considere os dados da Figura 3.1, onde existem nove agrupamentos compactos e bem separados.

Aplicando o FCM aos dados da Figura 3.1 para os valores de número de grupos $c = 1, \cdots, 15$, obtém-se os resultados mostrados nas Figuras 3.2 e 3.3. Para cada valor de c a base foi reamostrada 50 vezes e construído um *boxplot* para representar os valores obtidos. É possível observar que a função objetivo J_{FCM} sempre decresce ao aumentar o número de grupos c.

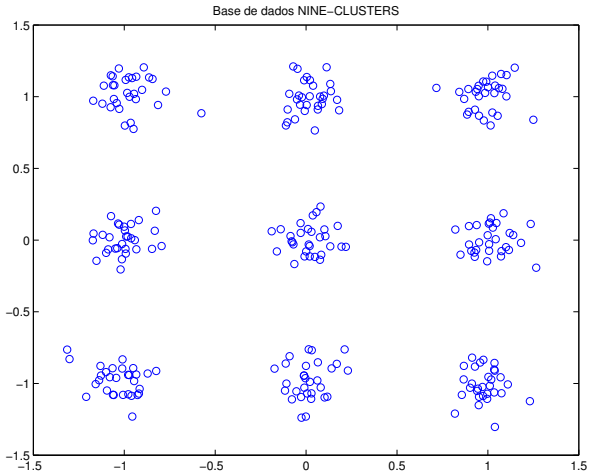


Figura 3.1: Base de dados sintética para demonstração do conflito de objetivos do FCM.

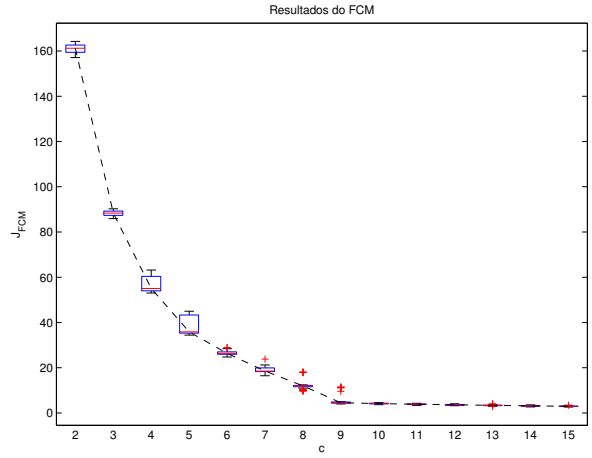


Figura 3.2: Boxplot dos resultados do FCM para a base de dados da Figura 3.1. Foram realizadas 50 reamostragens no total.

A partir das Figuras 3.2 e 3.3 é possível observar também a semelhança da curva de J_{FCM} com uma curva "pareto-ótima", de problemas de otimização multiobjetivo [Marler & Arora, 2004]. Esta semelhança já foi detectada antes em [Tibshirani & Walther, 2005], onde é traçado um paralelo entre o conflito do número de grupos com a função objetivo com o dilema viés-variância [Geman et al., 1992] presente nos problemas de aprendizado de máquina supervisionados. Neste caso, as métricas para validação de agrupamentos atuam como decisores de um problema de otimização multiobjetivo.

Além disto, soluções de regularização para funções objetivo de métodos de agrupamento já foram propostas nos trabalhos [Kothari & Pitts, 1999], [Li et al., 2008] e

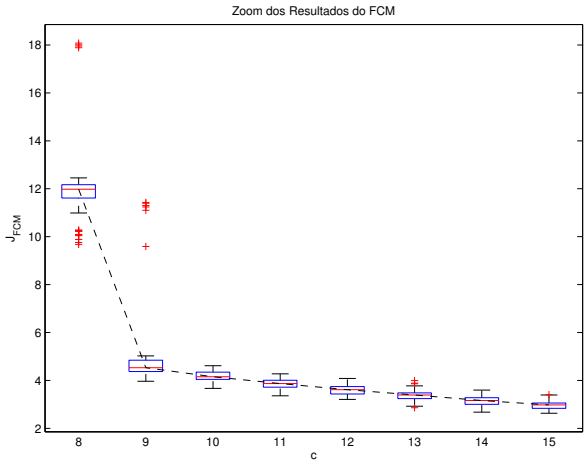


Figura 3.3: Zoom dos resultados da Figura 3.2 para $c \geq 8$.

[Lindsten et al., 2011]. Todas estas formulações buscam minimizar o efeito do aumento do número de agrupamentos através de um fator de penalização na função objetivo. O problema destas abordagens é que estes fatores dependem de um novo parâmetro de ponderação e nenhum dos métodos propõe uma maneira objetiva de encontrá-lo, ou seja, na prática ocorre apenas uma troca entre encontrar o parâmetro c que representa o número de grupos e um parâmetro λ que representa o fator de ponderação da penalização.

3.3 Métricas de Validação *Fuzzy*

As análises dos métodos de validação escolhidos para serem utilizados neste trabalho são divididas entre os métodos que utilizam apenas a matriz de partição e os métodos que utilizam tanto a matriz de partição como os próprios dados para o cálculo do índice.

3.3.1 Métricas Que Utilizam Somente a Matriz de Partição U

3.3.1.1 Partition Coefficient (PC)

O índice *Partition Coefficient* (PC), descrito em [Bezdek, 1973], é uma métrica de validação que mede a quantidade de interseção *fuzzy* entre os elementos da matriz de partição **U**. É descrito pela Equação 3.2.

$$PC = \frac{1}{N_U} \sum_{i=1}^{N_U} \sum_{j=1}^{c} u_{ij}^2 \tag{3.2}$$

No melhor cenário possível para o FCM, a matriz de partição **U** se assemelha a uma partição rígida, caso no qual cada elemento só pertence a um agrupamento. Já no pior caso, todos os elementos estarão igualmente distribuidos entre todos os c agrupamentos, resultando em uma quantidade de $1/c$ para todos os valores da matriz de partição. Da Equação 3.2 tem-se então que o valor mínimo para o índice PC é $1/c$. O número ótimo de agrupamentos c é dado pelo valor máximo do índice.

3.3.1.2 Partition Entropy (PE)

Bezdek também propôs o índice *Partition Entropy* (PE) [Bezdek, 1975] como definido pela Equação 3.3.

$$PE = -\frac{1}{N_U}\sum_{i=1}^{N_U}\sum_{j=1}^{c} u_{ij} log_a(u_{ij}) \tag{3.3}$$

onde a é a base do logaritmo.

No melhor caso, quando **U** se assemelha a uma partição rígida, todos os fatores $log_a(u_{ij})$ são 0, fazendo com que o valor de PE também se anule. Já no pior caso, o índice PE somará $log_a(c)$, o que com faz com que o intervalo da métrica seja $0 \leq PE \leq log_a(c)$. O número ótimo de agrupamentos é dado pelo valor mínimo de PE, mas neste trabalho o sinal da métrica é trocado para transformar a Equação 3.3 em um problema de maximização.

3.3.1.3 Modified Partition Coefficient (MPC)

Considerando o conflito que ocorre para os métodos de agrupamento baseados em centróides, pode-se mostrar que o mesmo ocorre para os índices PC e PE dados pelas Equações 3.2 e 3.3 quando o valor de c aumenta. Assim, o índice *Modified Partition Coefficient* (MPC) [Dave, 1996] foi proposto como um esforço para tentar corrigir a tendência monotônica que índices como PC e PE possuem. O índice MPC é dado pela Equação 3.4.

$$MPC = 1 - \frac{c}{c-1}(1 - PC) \tag{3.4}$$

Assim como para o índice PC, o número ótimo de agrupamentos é dado pelo valor máximo de MPC. Neste caso o intervalo da métrica é $0 \leq MPC \leq 1$.

3.3.2 Índices que Utilizam a Matriz de Partição e os Dados

3.3.2.1 Fukuyama Sugeno (FS)

O índice de validação *Fukuyama and Sugeno* (FS) [Fukuyama & Sugeno, 1989] é dado pela Equação 3.5

$$FS = \sum_{i=1}^{c}\sum_{j=1}^{N_U} u_{ji}^{m}\|\mathbf{x}_j - \mathbf{v}_i\| - \sum_{i=1}^{c}\sum_{j=1}^{N_U} u_{ji}^{m}\|\mathbf{v}_i - \overline{\mathbf{v}}\| \tag{3.5}$$

tal que $\overline{\mathbf{v}} = \sum_{i=1}^{c} \frac{\mathbf{v}_i}{c}$.

É possível observar que o primeiro termo da Equação 3.5 é exatamente a expressão para J_{FCM}, conforme a Equação 3.1. Esta expressão está relacionada com a compacticidade geométrica dos agrupamentos do conjunto de dados. Já o segundo termo é uma medida da dispersão dos centróides, por exemplo, está relacionado à proximidade entre os centros dos agrupamentos. O valor ótimo de c, c^*, é dado pela partição que provê o menor valor para o índice FS, confirmando com a função objetivo do FCM. Assim como PE, o sinal da equação do índice é invertido para a métrica ser analisada como um problema maximização.

3.3.2.2 Xie Beni (XB)

O índice *Xie Beni* (XB) [Xie & Beni, 1991] também é baseado na definição de compacticidade e separabilidade dos agrupamentos, mas a questão da separabilidade é abordada de uma maneira diferente, como mostrado na Equação 3.6.

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N_U} u_{ji}^{m}\|x_j - v_i\|}{N_U min_{i,j}\|v_i - v_j\|} \tag{3.6}$$

Enquanto o numerador indica a compacticidade da partição através da equação de J_{FCM}, o denominador indica a força de separação entre os agrupamentos. O valor ótimo do número de grupos é dado pela partição que possuir o valor mínimo para XB. O sinal desta métrica também é invertido para a análise ser realizada como um problema de maximização.

3.4 Conclusões do Capítulo

Neste capítulo foram apresentadas as principais abordagens da literatura para selecionar o número de grupos em problemas de agrupamento. Foi descrito como a função objetivo do algoritmo FCM J_{FCM} se comporta de maneira decrescente à medida que

o número de agrupamentos c aumenta. Além disto, foi discutido como a função J_{FCM} apresenta-se como uma curva "pareto-ótima" de um problema de otimização multiobjetivo e traçado um paralelo do conflito com o dilema bias-variancia dos problemas de aprendizado supervisionado. Foram apresentadas também as métricas de validação para agrupamentos *fuzzy* que serão utilizadas na seção de experimentos.

Capítulo 4

Método Proposto

Conforme analisado no capítulo **??**, as funções objetivo dos algoritmos de agrupamento não são capazes de identificar o número de grupos de uma base de dados. Para tal fim existem várias métricas na literatura que, através de testes comparativos para vários valores de c, propõem uma maneira objetiva para guiar esta escolha.

O problema é que, para cada métrica proposta, uma série de premissas deve ser assumida. Além disto, como muitas vezes a estrutura geométrica dos dados é desconhecida, a simples utilização destas métricas não provê a informação necessária para a tomada de decisão. Uma alternativa para transpor este problema é expor visualmente os agrupamentos obtidos para uma dada partição, e transferir a avaliação final do processo de análise de agrupamentos para o usuário.

Neste capítulo é proposto um método para validação de agrupamentos, composto de uma etapa qualitativa e de uma etapa quantitativa. Na etapa quantitativa, medidas estatísticas são extraídas de uma matriz de proximidade *fuzzy* obtida a partir da matriz de partição oriunda de um algoritmo particional fuzzy, como o FCM [Bezdek, 1981]. Na etapa qualitativa é proposta uma maneira de visualização para a avaliação da qualidade das partições obtidas.

4.1 Representação por Matrizes de Afinidade

Dado um conjunto de dados $D_U = \{\mathbf{x}_i\}_{i=1}^{N_U}$, onde N_U é o número total de padrões, os elementos a_{ij} da matriz de afinidade $\mathbf{A}$ correspondentes possuem uma medida de similaridade para os pares $(\mathbf{x}_i, \mathbf{x}_j)$. Devido à propriedade reflexiva da definição de similaridade, a matriz de afinidade é simétrica e tem-se que $a_{ij} = a_{ji}$. Em problemas de agrupamento, a similaridade é normalmente dada pela distância entre os vetores $\mathbf{x}_i$ e $\mathbf{x}_j$, a qual é uma métrica reflexiva, resultando assim em uma matriz simétrica. Outras

formas de afinidade podem ser utilizadas em problemas de agrupamentos, como o produto escalar, proximidades de partição e funções de *kernel* [Shawe-Taylor & Cristianini, 2004].

Considere que um método de agrupamento como o FCM [Bezdek, 1981] tenha sido aplicado ao conjunto de dados D_U e que $\mathbf{x}_i \in D_U$ foi ordenado de acordo com o agrupamento predominante, tal que os elementos do agrupamento 1 são indexados primeiro, em seguida os elementos do agrupamento 2 e assim por diante. Para um conjunto ordenado D_U, uma matriz de afinidade representativa, como uma matriz de *kernel* $\mathbf{K}$, pode ser visualizada na forma bloco-diagonal da Equação 4.1, onde as submatrizes $\mathbf{K}_{ii}$ representam a afinidade dos elementos de um mesmo grupo e as submatizes $\mathbf{K}_{ij}(\forall i \neq j)$ representam as afinidades dos elementos de diferentes grupos.

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} & \cdots & \mathbf{K}_{1c} \\ \mathbf{K}_{21} & \mathbf{K}_{22} & \cdots & \mathbf{K}_{2c} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{K}_{c1} & \mathbf{K}_{c2} & \cdots & \mathbf{K}_{cc} \end{bmatrix} \tag{4.1}$$

4.2 Matriz de Proximidade Fuzzy

Dada uma matriz de partição $\mathbf{U}_{N\times c}$, obtida de um método de agrupamento fuzzy como o FCM, a matriz de proximidade $\mathbf{P}_{N\times N}$ correspondente é definida pela Equação 4.2.

$$\mathbf{P}(k,l) = \sum_{i=1}^{c} min(\mathbf{U}(i,k), \mathbf{U}(i,l)), \quad \forall k,l \in \{1 \cdots N\} \tag{4.2}$$

Cada elemento p_{kl} de $\mathbf{P}$ contém uma medida de afinidade dos padrões $\mathbf{x}_k$ e $\mathbf{x}_l$. $\mathbf{P}$ também é simétrica, visto que $min(\mathbf{U}(i,k), \mathbf{U}(i,l)) = min(\mathbf{U}(i,l), \mathbf{U}(i,k))$. Desta forma, a matriz $\mathbf{P}$ pode ser representada da forma bloco-diagonal da Equação 4.1 se os padrões associados com as linhas de $\mathbf{U}$ forem ordenados de acordo com o agrupamento com o qual possuem maior pertinência, por exemplo, o agrupamento com maior valor na matriz $\mathbf{U}$.

A matriz de proximidade correspondente obtida a partir da aplicação da Equação 4.2 aos dados da Figura 4.1 é apresentada na Figura 4.2. Fica evidente então que as submatrizes da diagonal principal representando as afinidades internas de um agrupamento são bem definidas e possuem maior magnitude que as submatrizes fora da diagonal principal. Isto ocorre devido à compacticidade dos dados da Figura 4.1 e à escolha de c como sendo o mesmo número de funções geradoras do conjunto de dados. A atribuição ótima dos padrões aos agrupamentos é aquela que melhor destaca

a compacticidade na matriz **P** ou, em outras palavras, aquela que maximiza as magnitudes da diagonal principal e minimiza as magnitudes das submatrizes fora dela. Este conceito será explorado a seguir para descrever o método proposto.

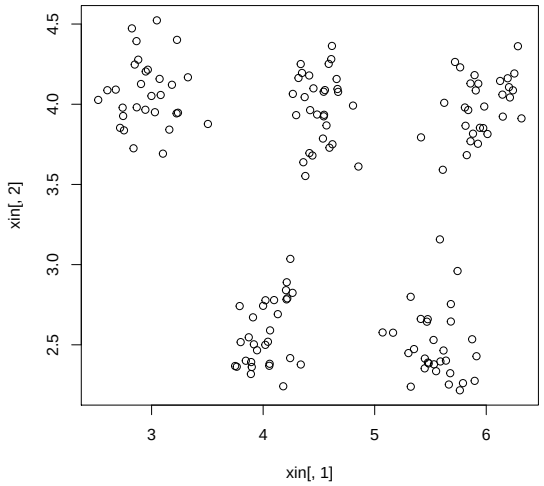


Figura 4.1: Dados amostrados de 5 diferentes distruições Gaussianas bidimensionais. Para cada distribuição foram amostrados 30 padrões, resultando em 150 padrões no total.

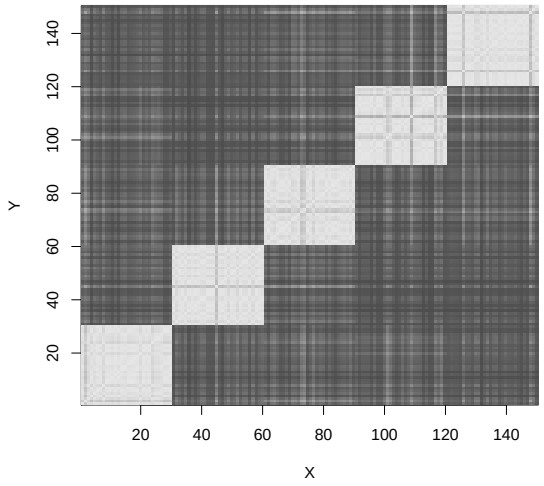


Figura 4.2: Matriz de proximidade dos dados da Figura 4.1.

4.3 Estimando o Número de Grupos a Partir da Matriz de Proximidade

A partir dos conceitos de compacticidade e separabilidade, é possível extrair métricas estatísticas que, combinadas, possuem a capacidade de validar uma determinada partição.

É de certa forma intuitivo pensar que uma escolha correta do número de agrupamentos c deve maximizar as afinidades internas de um agrupamento e minimizar as afinidades entre os grupos. Em outras palavras, de acordo com a representação da matriz de proximadade da Equação 4.2, um algoritmo de agrupamento pode ser descrito como um problema de otimização de duas funções extraídas a partir da matriz $\mathbf{P}$, dadas em forma geral pelas Equações 4.3 e 4.4

$$f_1(\mathbf{P}, c) = \phi_w(\mathbf{P}_{ii}) \tag{4.3}$$

$$f_2(\mathbf{P}, c) = \phi_i(\mathbf{P}_{ij}), i \neq j \tag{4.4}$$

onde $\mathbf{P}_{ii}$ representa as submatrizes da diagonal principal e $\mathbf{P}_{ij}$ representa as submatrizes fora da diagonal principal de $\mathbf{P}$, conforme a Equação 4.1 com $i, j = 1 \cdots c$.

Dadas as Equações 4.3 e 4.4 e assumindo que elas fornecem uma estimativa da homogeneidade dos valores de magnitude das submatrizes diagonais e não diagonais, o problema de otimização resultante para encontrar o valor c pode ser descrito como: *encontrar o número de agrupamentos c que maximiza* f_1 *e minimiza* f_2. O problema consiste então em expressar as funções f_1 and f_2 de uma maneira que elas representem propriamente o problema. A função objetivo deve portanto maximizar as magnitudes dos elementos $\mathbf{P}_{ii}$ e, ao mesmo tempo, minimizar as magnitudes dos elementos $\mathbf{P}_{ij}$. Pode-se dizer também que tanto as afinidades internas como as afinidades entre os grupos devem ter sua homogeinização maximizada, ou seja, espera-se que não exista uma discrepância de magnitudes muito grande dos elementos em um mesmo grupo, visto que tal fato pode indicar a falta de compacticidade dos agrupamentos obtidos. Desta forma, a média das magnitudes pode ser utilizada para obter as funções f_1 e f_2, calculada a partir de

$$g(\mathbf{P}, i_b, i_e, j_b, j_e) = \frac{1}{(i_e - i_b) + (j_e - j_b)} \sum_{i=i_b}^{i_e} \sum_{j=j_b}^{j_e} \mathbf{P}_{ij} \ , \tag{4.5}$$

onde $\mathbf{P}$ é a matriz de proximidade, a qual é ordenada de acordo com os agrupamentos

obtidos para certos valores de c, e i_b, i_e, j_b e j_e são os índices que marcam o início e fim das linhas e colunas correspondentes de uma determinada submatriz de $\mathbf{P}$.

As funções correspondentes f_1 e f_2 são aquelas representadas por

$$f_1(\mathbf{P}, c) = \frac{1}{c} \sum_{k=1}^{c} g(\mathbf{P}, i_b^{kk}, i_e^{kk}, j_b^{kk}, j_e^{kk}) \ e \tag{4.6}$$

$$f_2(\mathbf{P}, c) = \frac{1}{c^2 - c} \sum_{k=1}^{c} \sum_{l=1}^{c} g(\mathbf{P}, i_b^{kl}, i_e^{kl}, j_b^{kl}, j_e^{kl}), \ \ \forall k \neq l \tag{4.7}$$

tal que i_b^{kk}, i_e^{kk}, j_b^{kk} e j_e^{kk} são as coordenadas de início e fim que localizam as submatrizes de afinidade intra grupos em $\mathbf{P}$ e i_b^{kl}, i_e^{kl}, j_b^{kl} e j_e^{kl} são as coordenadas de início e fim que localizam as submatrizes de afinidade entre grupos na matriz de proximidade $\mathbf{P}$.

A otimização conjunta das funções $f_1(\mathbf{P}, c)$ e $f_2(\mathbf{P}, c)$ requereria uma abordagem multiobjetivo se estas possuíssem um comportamento conflitante. Todavia, o parâmetro c que maximiza as matrizes subdiagonais é o mesmo que minimiza as submatrizes não-diagonais. Assim, o problema de otimização matemática que representa o problema pode ser formulado como um problema de um único objetivo, sendo representado por uma função construída através da combinação linear de $f_1(\mathbf{P}, c)$ e $f_2(\mathbf{P}, c)$. Visto que quando o valor ótimo de c for escolhido a diferença entre as duas funções objetivo deve ser máxima, o problema de otimização resultante pode ser colocado como

$$\arg\max_{c} J(\mathbf{P}, c) \ , \tag{4.8}$$

onde $J(\mathbf{P}, c) = f_1(\mathbf{P}, c) - f_2(\mathbf{P}, c)$.

Para observar o comportamento de $J(\mathbf{P}, c)$, matrizes de proximidade $\mathbf{P}$ para $c = 2 \cdots 10$ foram obtidas e $J(\mathbf{P}, c)$ foi então calculada para cada valor de c. As funções $f_1(\mathbf{P}, c)$ e $f_2(\mathbf{P}, c)$ fornecem o valor médio das homogeneidades para cada submatriz. A partir dos resultados mostrados na Figura 4.3 é possível observar que o máximo da função ocorre em $c = 5$, que equivale ao número de agrupamentos do conjunto original dos dados.

No capítulo 5 o método será testado com outros conjuntos de dados sintéticos e bases de dados reais do repositório UCI [Bache & Lichman, 2013].

4.4 Conclusões do Capítulo

Neste capítulo mostrou-se que a compacticidade de partições geradas a partir de métodos de agrupamento baseados em centróides pode ser diretamente visualizada na forma de matrizes de proximidade quando estas são representadas na forma bloco-

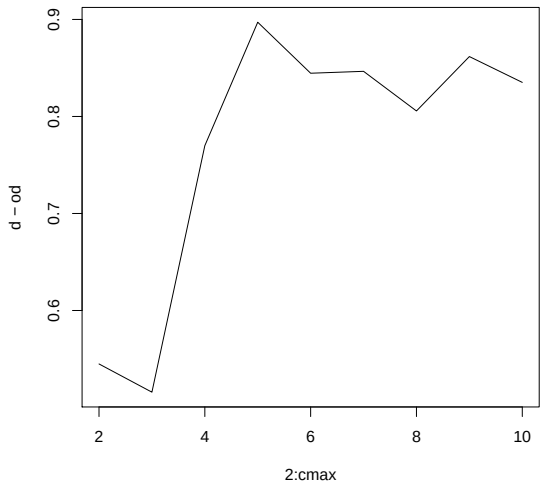


Figura 4.3: Função objetivo $J(\mathbf{P}, c) = f_1(\mathbf{P}, c) - f_2(\mathbf{P}, c)$ para valores de c variando de 2 a 10 para os dados da Figura 4.1. O máximo da função objetivo ocorre em $c = 5$, o número de agrupamentos do conjunto de dados.

diagonal. Foi discutido que medidas estatísticas extraídas das submatrizes em forma bloco-diagonal podem ser combinadas para gerar índices de validação para agrupamentos. A noção intuitiva de que as relações internas de cada grupo de um agrupamento devem ser mais fortes do que as relações dos elementos de diferentes grupos é demonstrada pela representação bloco-diagonal da partição. Para descrever o novo critério de valição de agrupamentos, foram exploradas as propriedades particulares das matrizes de proximidade, provenientes das matrizes de partição obtidas como resultado da aplicação do algoritmo FCM.

Capítulo 5

Experimentos

Neste capítulo, o método proposto é testado experimentalmente com bases de dados sintéticas e reais. O desempenho da métrica desenvolvida é comparado a métodos de validação de agrupamentos encontrados na literatura. Os resultados obtidos são analisados graficamente com o intuito de validar a visualização das matrizes de proximidade ordernadas como um método qualitativo para a análise de partições.

5.1 Metodologia

Os parágrafos a seguir descrevem a metodologia geral adotada na condução dos experimentos. Metodologia similar foi utilizada em [Wu & Yang, 2005] e [Xu & Wunsch, 2008].

Devido à natureza não supervisionada dos método de agrupamentos, situação na qual a função geradora do conjunto de dados $P(\mathbf{x})$ não é conhecida, não é possível traçar conclusões sobre o número de grupos existentes em um conjunto de dados. Sendo assim, os experimentos foram divididos em bases sintéticas, onde a função $P(\mathbf{x})$ é conhecida, e bases de dados reais do repositório UCI [Bache & Lichman, 2013], onde não existem informações à priori sobre a estrutura intrínseca dos dados.

Para demonstar a validade do método proposto, são realizadas comparações com outras cinco métricas de validação de agrupamentos: PC [Bezdek, 1973], CE [Bezdek, 1975], MPC [Dave, 1996], FS [Fukuyama & Sugeno, 1989], XB [Xie & Beni, 1991].

Para o algoritmo de agrupamentos, foi utilizado o FCM em todos os experimentos. O valor do parâmetro de fuzificação utilizado foi $m = 2$. O número máximo de 200 iterações foram realizadas em cada repetição e os centros iniciais são escolhidos aleatoriamente dentro do domínio do problema. Cada base sintética foi amostrada 50

vezes e as bases reais foram reamostradas aleatoriamente com 80% do seu tamanho em cada rodada.

5.2 Experimentos Com Bases de Dados Sintéticos

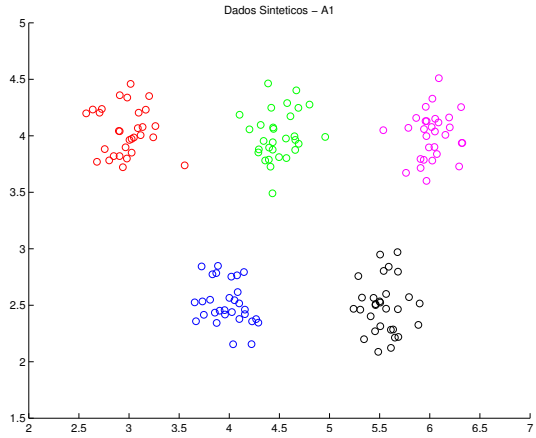
Conforme discutido anteriormente, os experimentos controlados são muito importantes em problemas envolvendo aprendizagem não supervisionada, visto que nestes casos a função geradora dos dados $P(\mathbf{x})$ é conhecida e é possível testar a validade do método proposto para situações específicas. Nesta seção busca-se contrastar a métrica proposta com outros métodos da literatura em relação ao número de grupos e robustez em relação à classes desbalanceadas e nível de superposição de agrupamentos.

Para testar o número de grupos, é utilizada a base de dados denominada *A1*, representada na Figura 5.1a. A base *A1* é composta por cinco gaussianas bidimensionais bem espaçadas, cada uma com 30 elementos. A fim de verificar a robustez das métricas em relação a agrupamentos desbalanceados, é utilizada a base *B1*, mostrada na Figura 5.1b. Esta base é composta por 3 agrupamentos, sendo que dois deles são compostos por 40 e um deles por 200 elementos. Por fim, visando verificar a robustez das métricas em relação ao nível de superposição das classes, são utilizadas as bases *C1* e *C2*, geradas por 4 gaussianas e se diferenciando pelo nível de espalhamento das mesmas, conforme representadas nas Figuras 5.1c e 5.1d respectivamente.

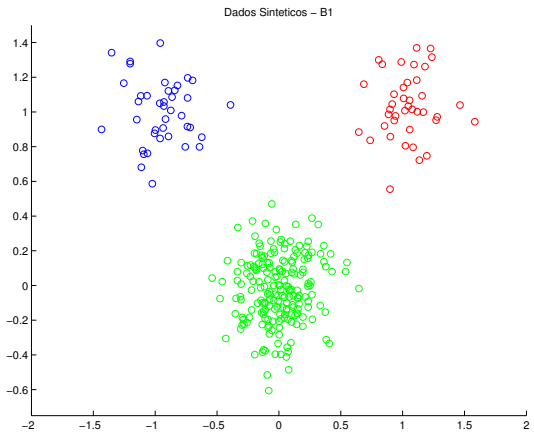
5.2.1 Resultados e Discussão

As Figuras 5.2a, 5.2b, 5.2c, 5.2d, 5.2e e 5.2f ilustram os resultados obtidos para as métricas nas 50 repetições para a base *A1*. É possível verificar através das imagens que todas as métricas conseguem encontrar o resultado $c = 5$ em todas as repetições, que é o número de funções geradoras. Os resultados obtidos são sumarizados na Figura 5.3, que representa o histrograma dos valores de c para os experimentos. É possível validar estes resultados através da Figura 5.4d, onde observa-se a coerência dos agrupamentos nas submatrizes da diagonal principal e a baixa interferência nas submatrizes fora da diagonal principal.

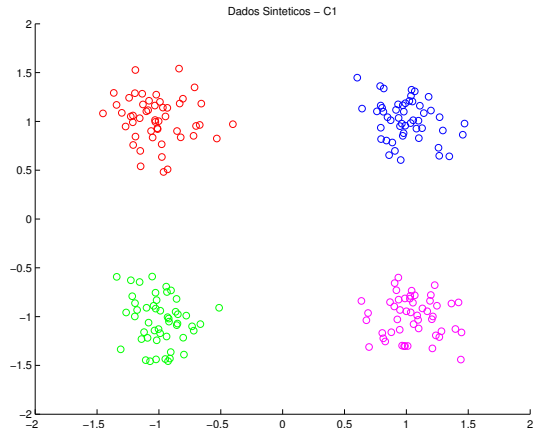
Para a base de dados *B1*, é possível observar os resultados das métricas nas Figuras 5.5a, 5.5b, 5.5c, 5.5d, 5.5e, 5.5f. Somente a métrica PE sofreu influência do desbalanceamento entre os grupos. O histrograma dos resultados obtidos pode ser visualizado na Figura 5.6. Assim como no caso para a base *A1*, através da visualização das matrizes de proximidade da Figura 5.7 é possível confirmar os resultados, visto que



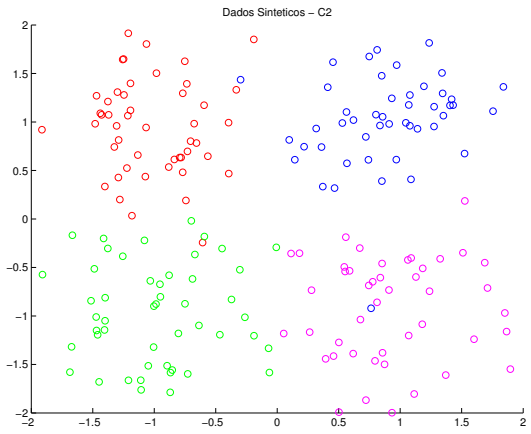
(a) Base de dados sintética *A1*.



(b) Base de dados sintética *B1*.



(c) Base de dados sintética *C1*.



(d) Base de dados sintética *C2*.

Figura 5.1: Bases de dados sintéticas *A1*, *B1*, *C1* e *C2*.

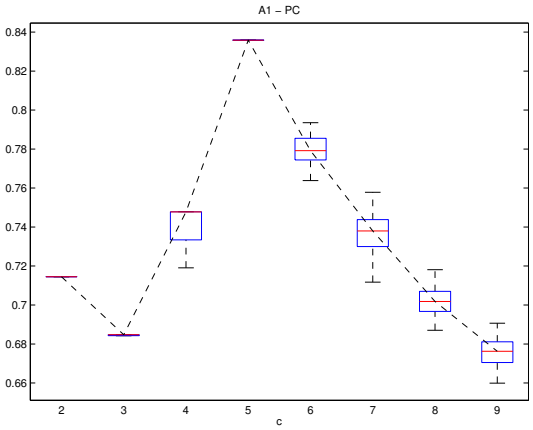
a matriz de proximidade para o caso $c = 3$ da Figura 5.7b é mais bem definida que as demais.

A fim de testar a robustez das métricas em relação ao nível de superposição dos agrupamentos, foram utilizadas as bases de dados *C1* e *C2*, tal que a função geradora de *C2* possui uma maior abertura das gaussianas.

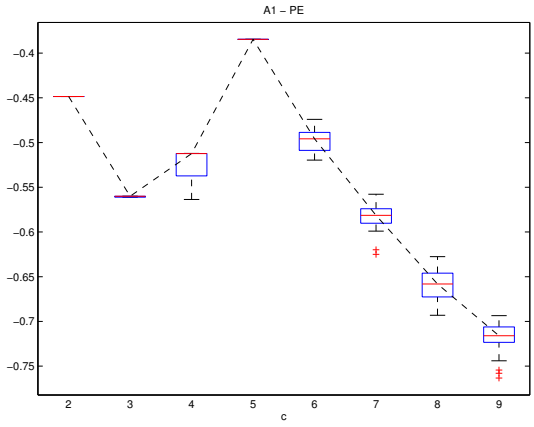
Para *C1*, todas as métricas são capazes de detectar os 4 agrupamentos existentes, conforme pode ser observado nas Figuras 5.8 e 5.9. As matrizes de proximidade da Figura 5.10 também corroboram com este resultado.

Já para a base *C2*, é possível observar os resultados das métricas através da Figura 5.11. O método proposto, identificado por BR, ficou bem dividido entre $c = 2$ e $c = 4$. As métricas PC e PE, que dependem apenas da matriz de partição $\mathbf{U}$, sofreram grande influência da superposição, não sendo capazes de manter o resultado apresentado para *C1*. Já as métricas MPC, FS e XB, que dependem também de outras

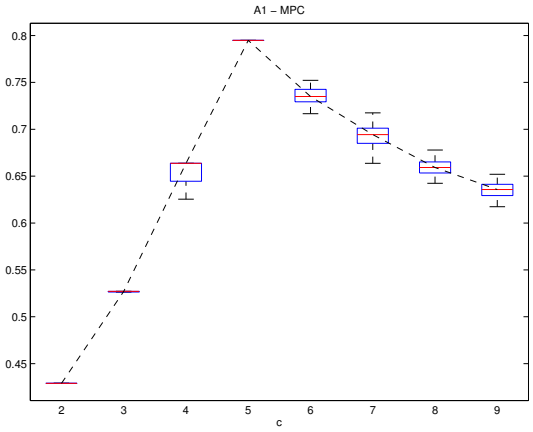
informações, como o número de grupos ou as distâncias entre eles, se mostraram robustas em relação à superposição dos agrupamentos. Os resultados podem ser visualizados no histograma da Figura 5.12.



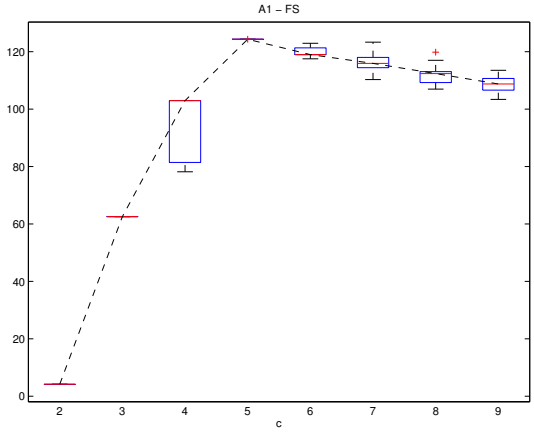
(a) Resultado para a métrica PC na base de dados *A1*.



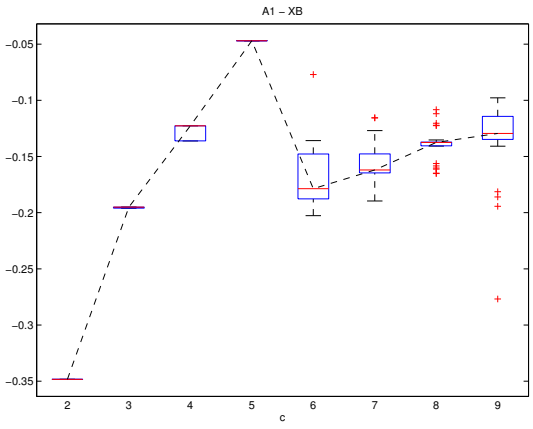
(b) Resultado para a métrica PE na base de dados *A1*.



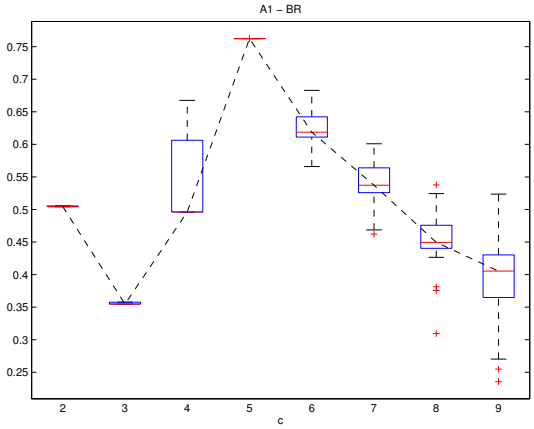
(c) Resultado para a métrica MPC na base de dados *A1*.



(d) Resultado para a métrica FS na base de dados *A1*.



(e) Resultado para a métrica XB na base de dados *A1*.



(f) Resultado para a métrica BR na base de dados *A1*.

Figura 5.2: Resultados das métricas para a base de dados *A1*.

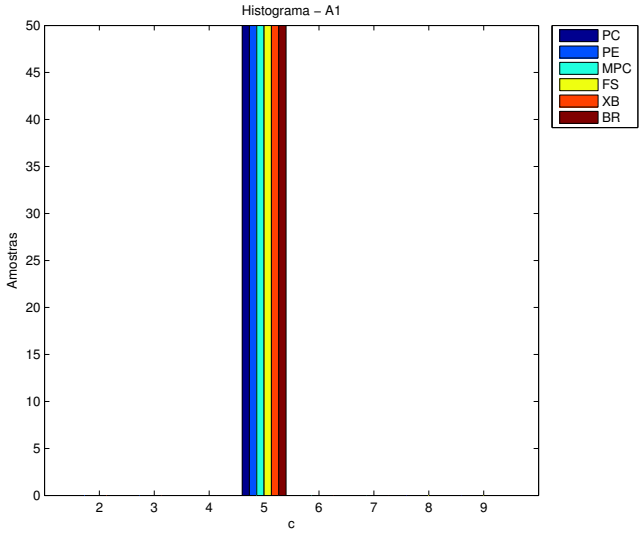
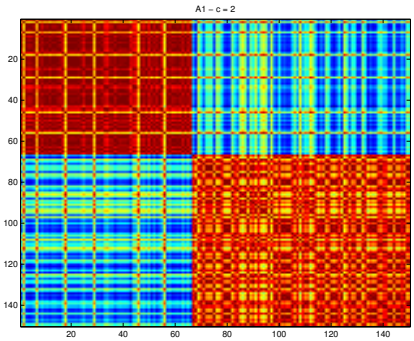
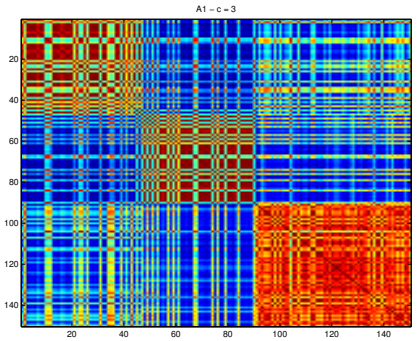


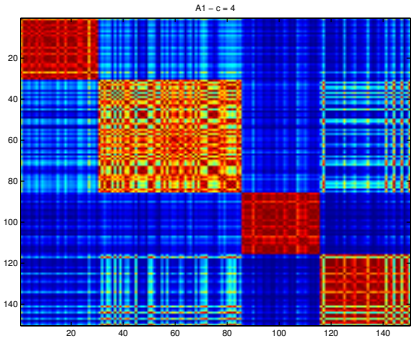
Figura 5.3: Histograma dos resultados para a base de dados *A1*. Todas as métricas são capazes de encontrar o número de grupos $c = 5$.



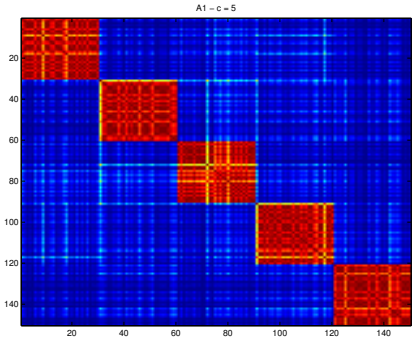
(a) Matriz de Proximidade para a base de dados $A1$ - $c = 2$.



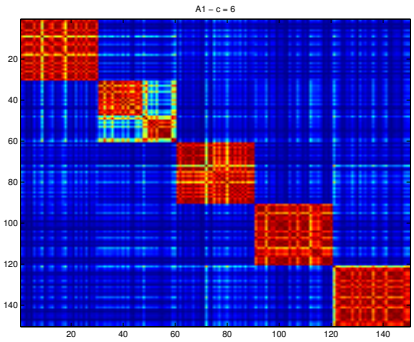
(b) Matriz de Proximidade para a base de dados $A1$ - $c = 3$.



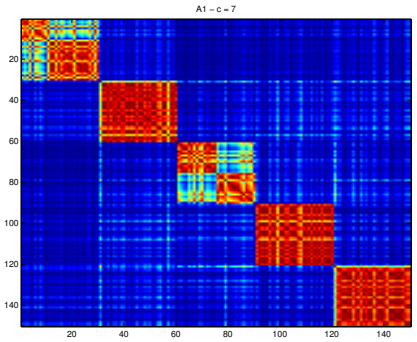
(c) Matriz de Proximidade para a base de dados $A1$ - $c = 4$.



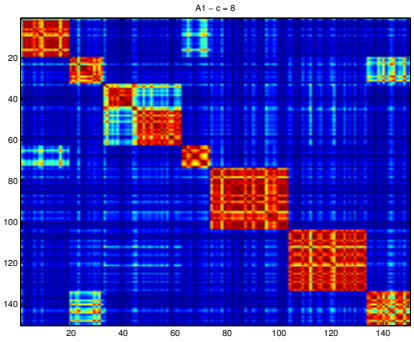
(d) Matriz de Proximidade para a base de dados $A1$ - $c = 5$.



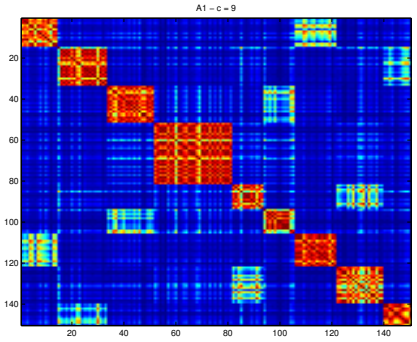
(e) Matriz de Proximidade para a base de dados $A1$ - $c = 6$.



(f) Matriz de Proximidade para a base de dados $A1$ - $c = 7$.

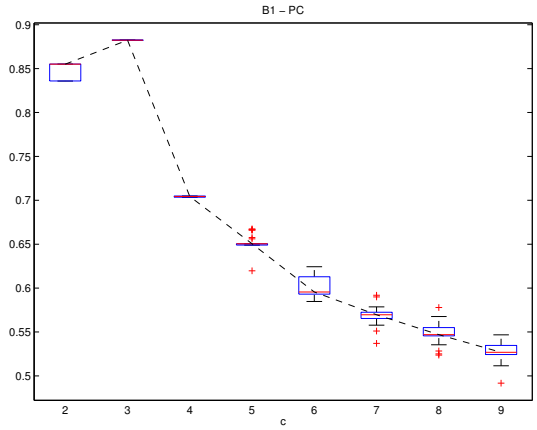


(g) Matriz de Proximidade para a base de dados $A1$ - $c = 8$.

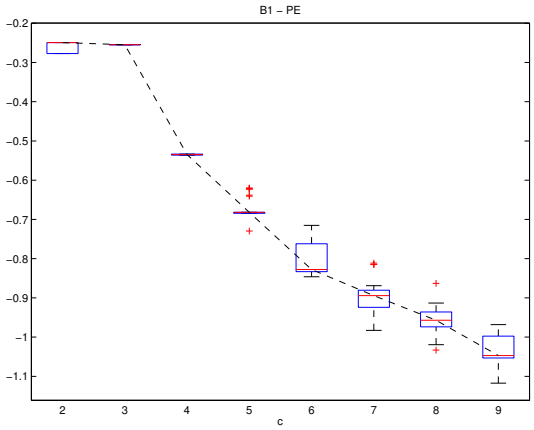


(h) Matriz de Proximidade para a base de dados $A1$ - $c = 9$.

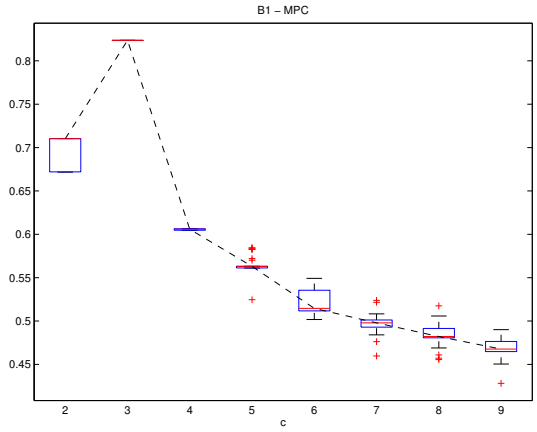
Figura 5.4: Matrizes de Proximidade para a base de dados $A1$.



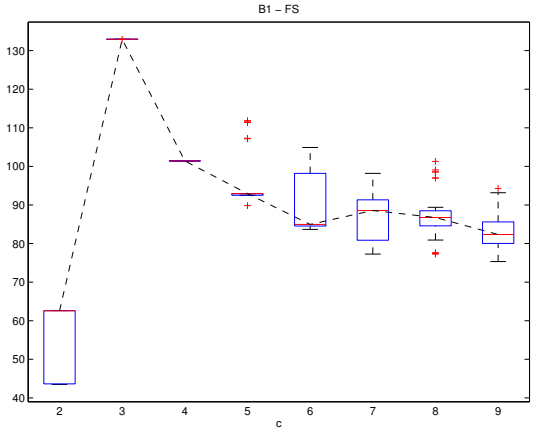
(a) Resultado para a métrica PC na base de dados *B1*.



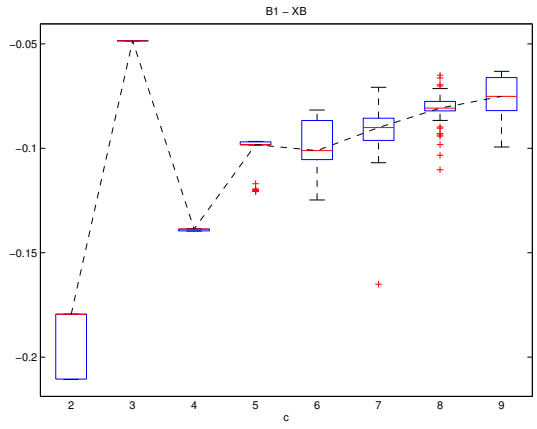
(b) Resultado para a métrica PE na base de dados *B1*.



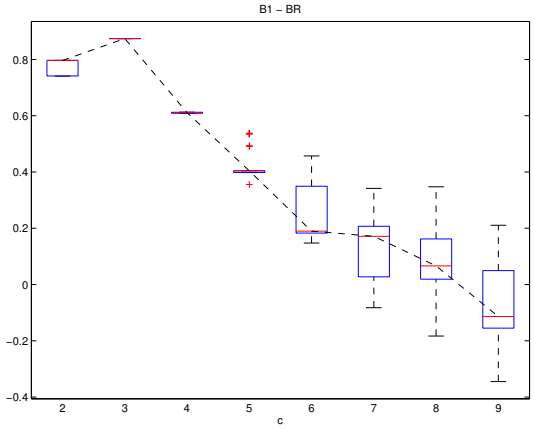
(c) Resultado para a métrica MPC na base de dados *B1*.



(d) Resultado para a métrica FS na base de dados *B1*.



(e) Resultado para a métrica XB na base de dados *B1*.



(f) Resultado para a métrica BR na base de dados *B1*.

Figura 5.5: Resultados das métricas para a base de dados *B1*.

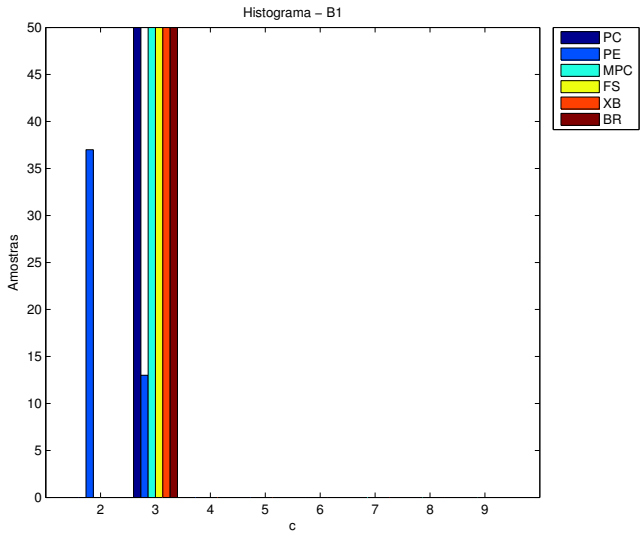
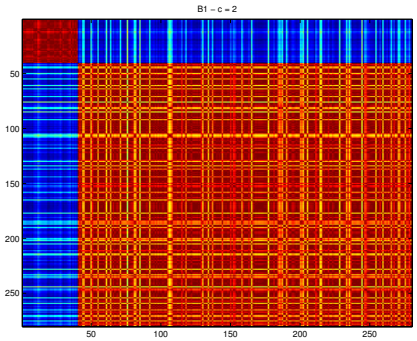
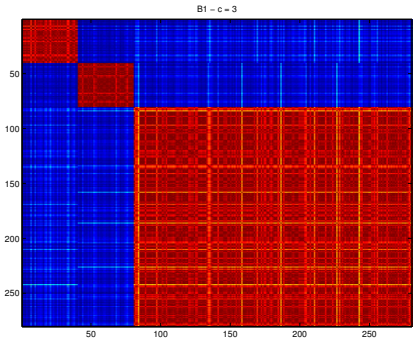


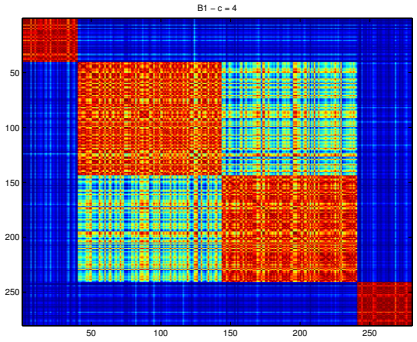
Figura 5.6: Histograma dos Resultados para a base de dados *B1*. A métrica PE mostrou-se pouco robusta em relação a agrupamentos desbalanceados.



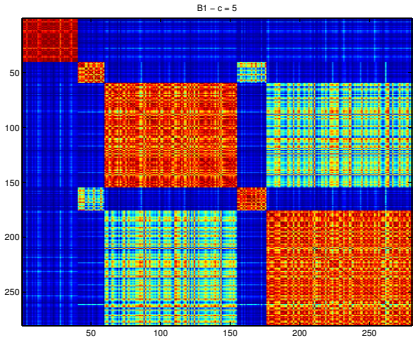
(a) Matriz de Proximidade para a base de dados *B1* - $c = 2$.



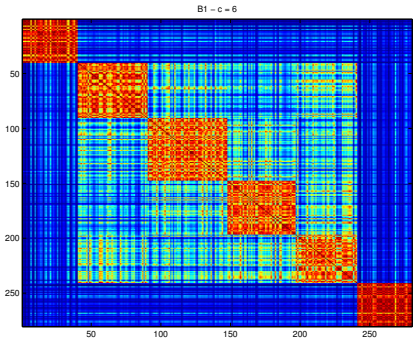
(b) Matriz de Proximidade para a base de dados *B1* - $c = 3$.



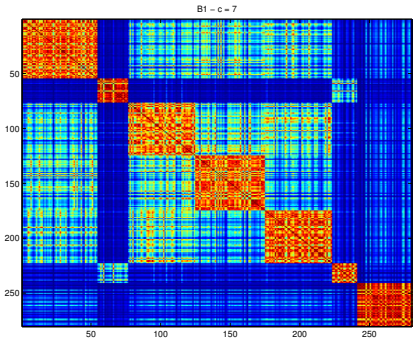
(c) Matriz de Proximidade para a base de dados *B1* - $c = 4$.



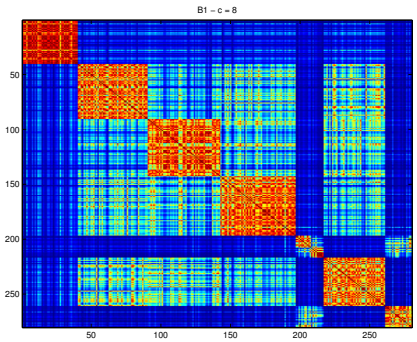
(d) Matriz de Proximidade para a base de dados *B1* - $c = 5$.



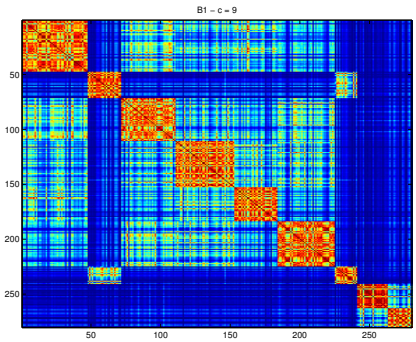
(e) Matriz de Proximidade para a base de dados *B1* - $c = 6$.



(f) Matriz de Proximidade para a base de dados *B1* - $c = 7$.

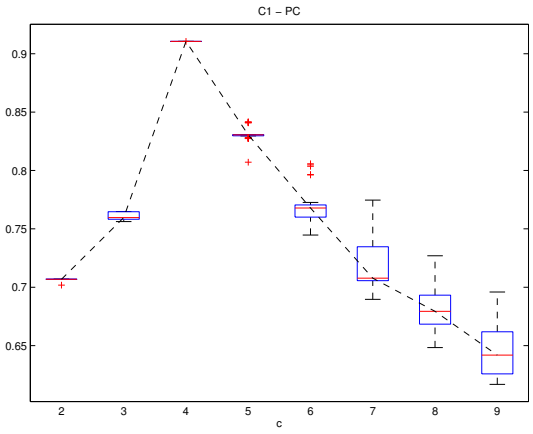


(g) Matriz de Proximidade para a base de dados *B1* - $c = 8$.

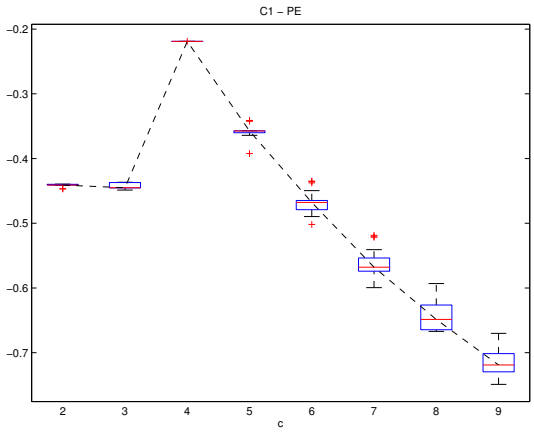


(h) Matriz de Proximidade para a base de dados *B1* - $c = 9$.

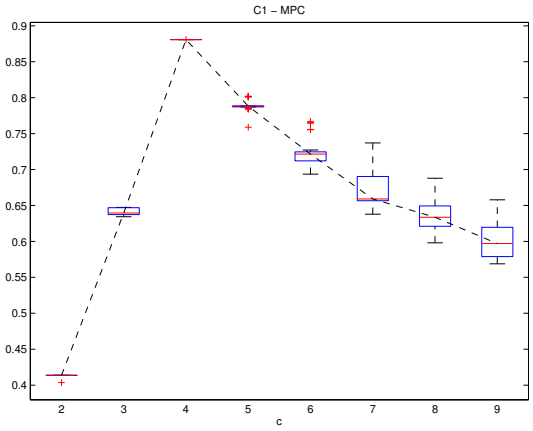
Figura 5.7: Matrizes de Proximidade para a base de dados *B1*.



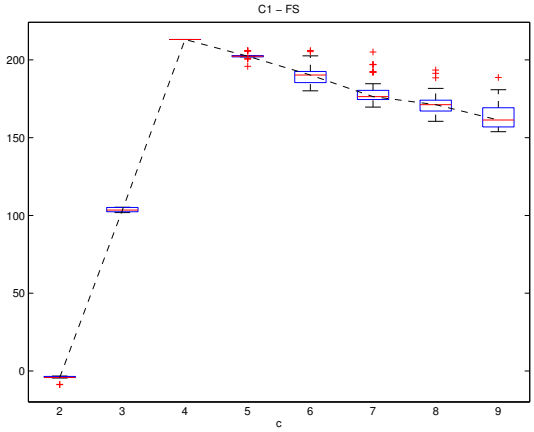
(a) Resultado para a métrica PC na base de dados *C1*.



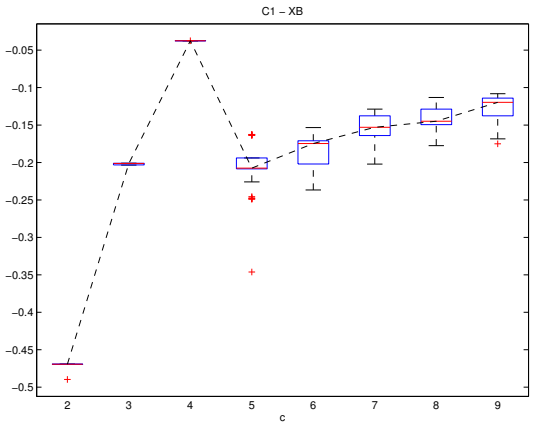
(b) Resultado para a métrica PE na base de dados *C1*.



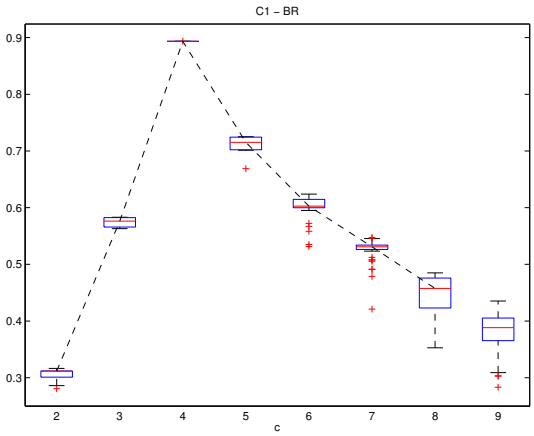
(c) Resultado para a métrica MPC na base de dados *C1*.



(d) Resultado para a métrica FS na base de dados *C1*.



(e) Resultado para a métrica XB na base de dados *C1*.



(f) Resultado para a métrica BR na base de dados *C1*.

Figura 5.8: Resultados das métricas para a base de dados *C1*.

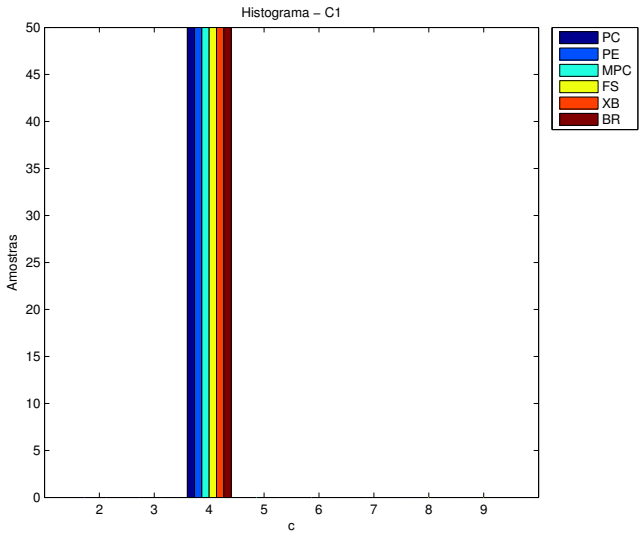
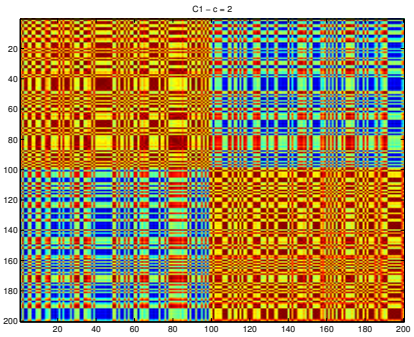
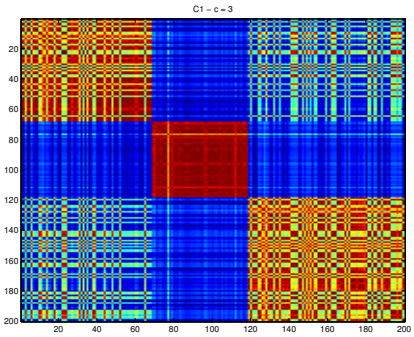


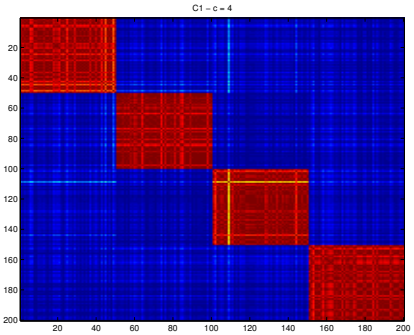
Figura 5.9: Histograma dos resultados para a base de dados *C1*. Todas as métricas são capazes de encontar o número de grupos $c = 4$.



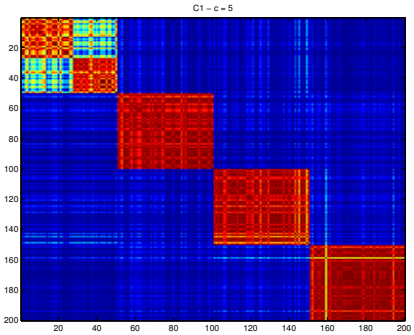
(a) Matriz de Proximidade para a base de dados *C1* - $c = 2$.



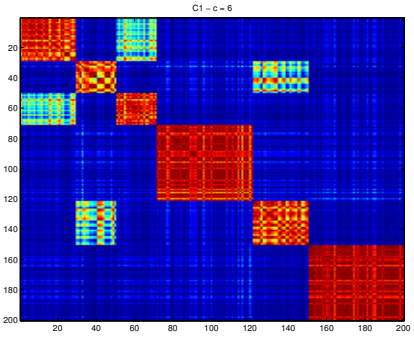
(b) Matriz de Proximidade para a base de dados *C1* - $c = 3$.



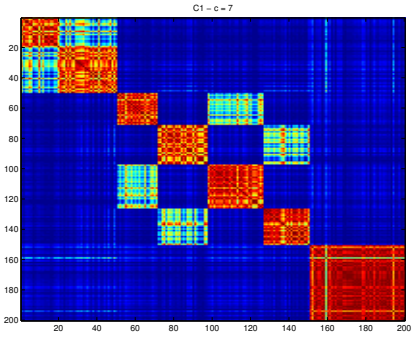
(c) Matriz de Proximidade para a base de dados *C1* - $c = 4$.



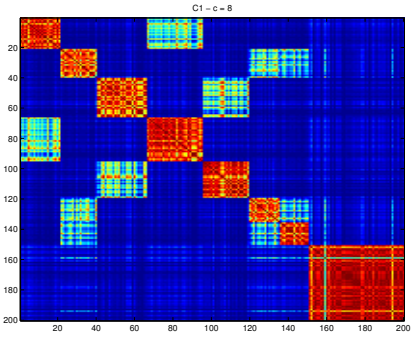
(d) Matriz de Proximidade para a base de dados *C1* - $c = 5$.



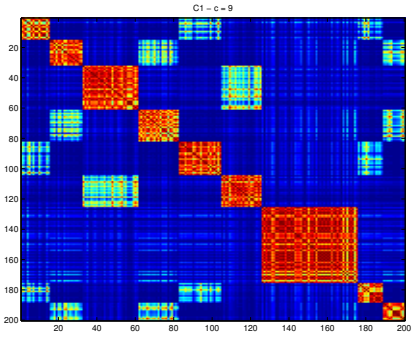
(e) Matriz de Proximidade para a base de dados *C1* - $c = 6$.



(f) Matriz de Proximidade para a base de dados *C1* - $c = 7$.

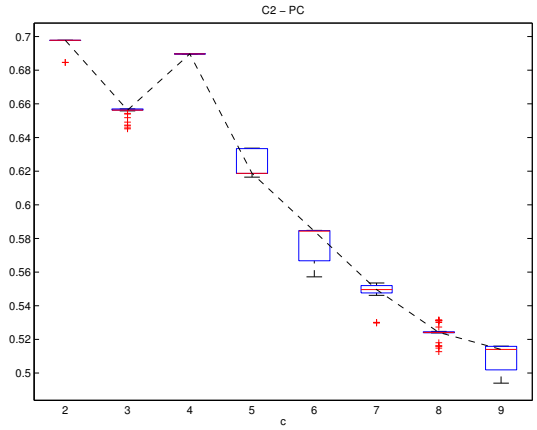


(g) Matriz de Proximidade para a base de dados *C1* - $c = 8$.

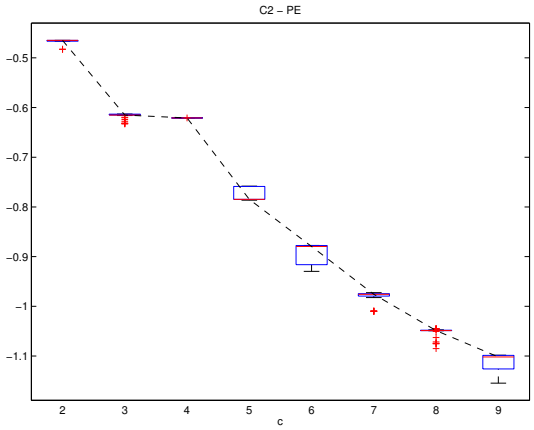


(h) Matriz de Proximidade para a base de dados *C1* - $c = 9$.

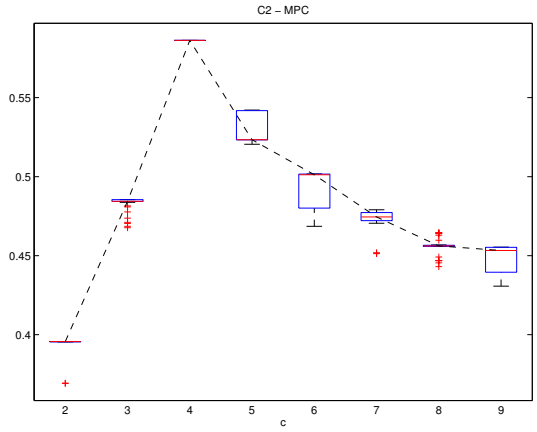
Figura 5.10: Matrizes de Proximidade para a base de dados *C1*.



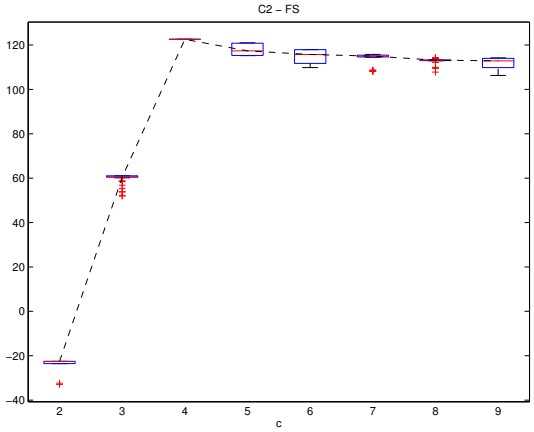
(a) Resultado para a métrica PC na base de dados *C2*.



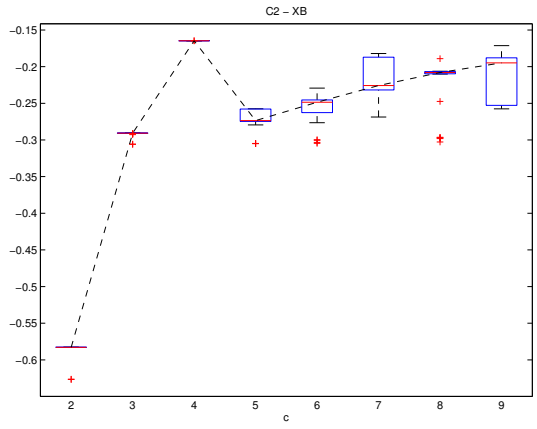
(b) Resultado para a métrica PE na base de dados *C2*.



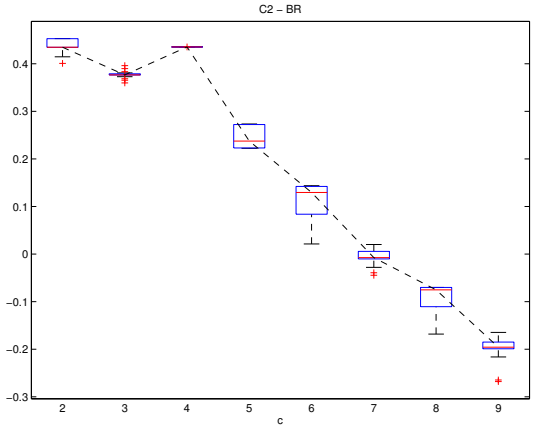
(c) Resultado para a métrica MPC na base de dados *C2*.



(d) Resultado para a métrica FS na base de dados *C2*.



(e) Resultado para a métrica XB na base de dados *C2*.



(f) Resultado para a métrica BR na base de dados *C2*.

Figura 5.11: Resultados das métricas para a base de dados *C2*.

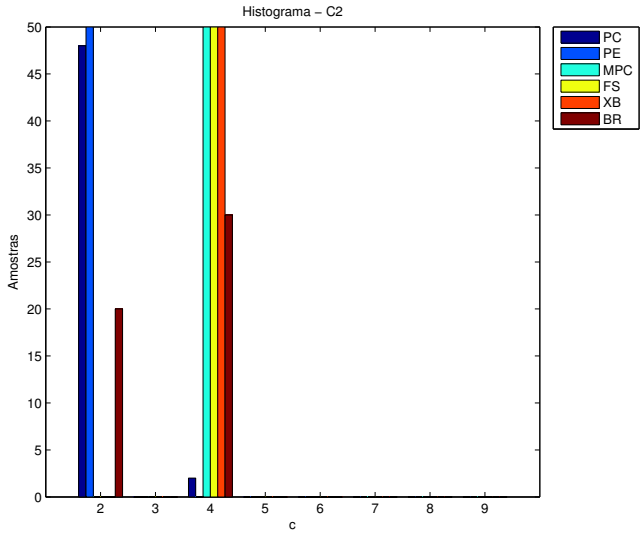
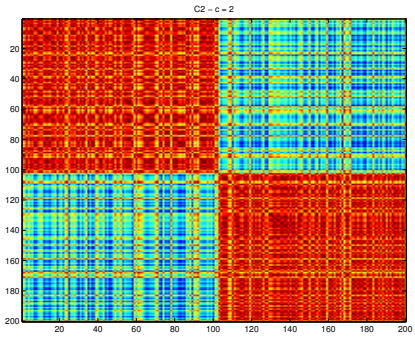
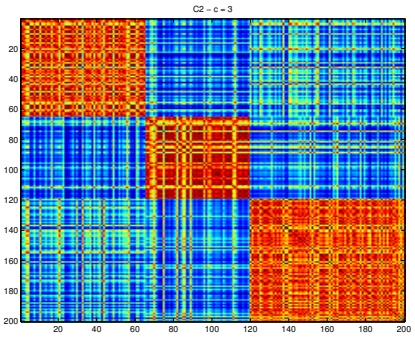


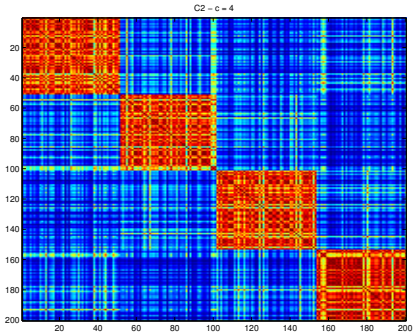
Figura 5.12: Histograma dos resultados para a base de dados *C2*. As métricas MPC. FS e XB se mostram robustas em relação à superposição dos agrupamentos. Já as métricas PC e PE não conseguem mais identificar o número de grupos $c = 4$. A métrica proposta BR apresenta resultados divididos entre $c = 2$ e $c = 4$.



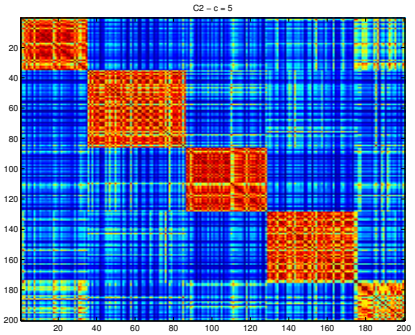
(a) Matriz de Proximidade para a base de dados *C2* - $c = 2$.



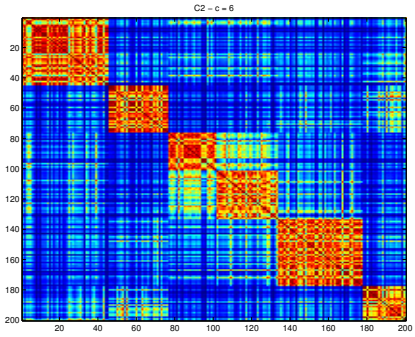
(b) Matriz de Proximidade para a base de dados *C2* - $c = 3$.



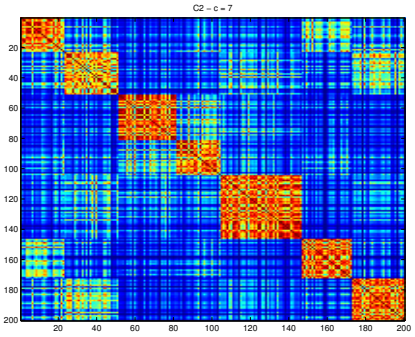
(c) Matriz de Proximidade para a base de dados *C2* - $c = 4$.



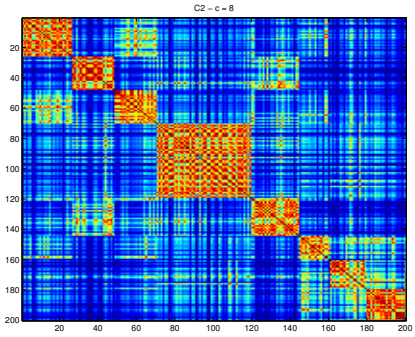
(d) Matriz de Proximidade para a base de dados *C2* - $c = 5$.



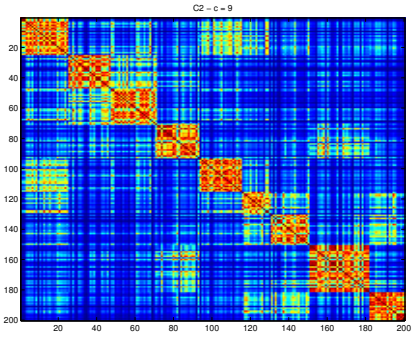
(e) Matriz de Proximidade para a base de dados *C2* - $c = 6$.



(f) Matriz de Proximidade para a base de dados *C2* - $c = 7$.



(g) Matriz de Proximidade para a base de dados *C2* - $c = 8$.



(h) Matriz de Proximidade para a base de dados *C2* - $c = 9$.

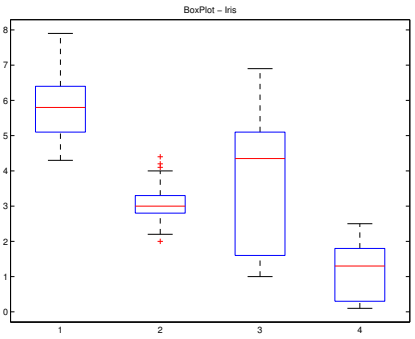
Figura 5.13: Matrizes de Proximidade para a base de dados *C2*.

5.3 Experimentos Com Bases de Dados Reais

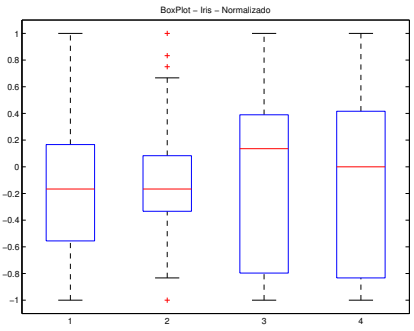
Para os experimentos com bases reais, a função geradora dos dados $P(\mathbf{x})$ é desconhecida, não sendo possível realizar afirmações sobre o número de grupos dos conjuntos. Além disto, também não é possível afirmar se as bases são formadas por agrupamentos compactos, premissa assumida para o funcionamento do método proposto e as outras métricas. Sendo assim, não é objetivo desta seção fornecer uma resposta convicta de um número "correto" de grupos para as bases, e sim fornecer um mecanismo de avaliação comparativo que seja capaz de validar se a partição está conforme os critérios adotados para cada caso.

As bases escolhidas para realizar os testes experimentais desta seção são denominadas *Iris*, *Wine*, *Glass* e *Wdbc*, todas retidadas do repositório UCI [Bache & Lichman, 2013]. Para estes conjuntos, todas as características são reais, portanto a métrica de dissimilaridade utilizada pelo FCM é a distância euclidiana. Além disto, é considerado que não existe nenhuma informação à priori a não ser os próprios dados e o domínio de negócios trabalhado é desconhecido. Desta forma, a fim de evitar o desbalanceamento entre os atributos de cada base, é realizada uma análise das características para as bases escolhidas através de gráficos *box-plot*, conforme as Figuras 5.14a, 5.14c, 5.14e e 5.14g.

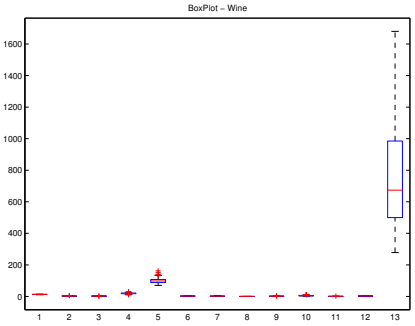
Para a base *Iris* é possível perceber que, apesar das magnitudes não serem muito distantes, a primeira característica poderia dominar o processo de agrupamento. Os dados são então normalizados e passam a apresentar as características apresentadas na Figura 5.14b. Em relação à base *Wine*, é possível observar que a característica 13 possui ordem de grandeza muito discrepante em relação aos outros atributos. Visto que não existem informações sobre o domínio do problema em questão, a base é também normalizada para a realização dos experimentos, obtendo um perfil de características mais uniforme, conforme a Figura 5.14b. A base de dados *Glass* também apresenta uma característica muito discrepante das demais. Portanto, é realizada a normalização e os dados passam a apresentar o perfil de características mostrado na Figura 5.14f. Por fim, a base *Wdbc* também é normalizada devido à alta discrepância de suas características 4 e 24. O novo perfil de características pode ser visualizado na Figura 5.14h.



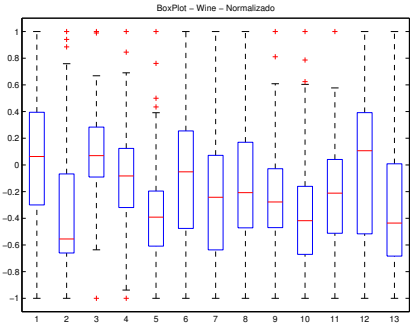
(a) *Boxplot* da base de dados *Iris*.



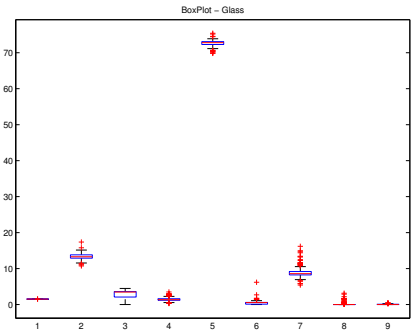
(b) *Boxplot* da base de dados *Iris* normalizada.



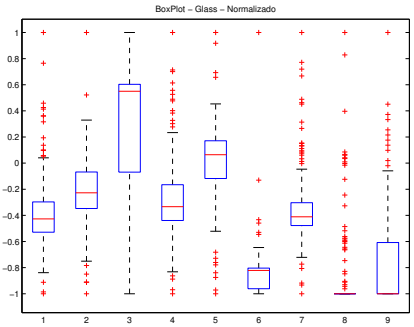
(c) *Boxplot* da base de dados *Wine*.



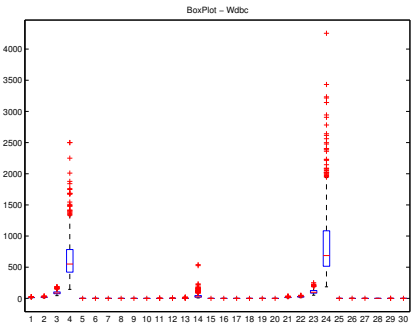
(d) *Boxplot* da base de dados *Wine* normalizada.



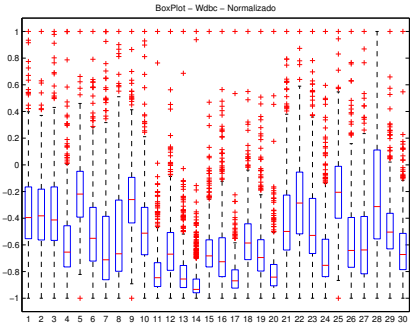
(e) *Boxplot* da base de dados *Glass*.



(f) *Boxplot* da base de dados *Glass* normalizada.



(g) *Boxplot* da base de dados *Wdbc*.



(h) *Boxplot* da base de dados *Wdbc* normalizada.

Figura 5.14: *Boxplot* das bases de dados reais *Iris*, *Wine*, *Glass* e *Wdbc*.

5.3.1 Resultados e Discussão

As Figuras 5.15a, 5.15b, 5.15c, 5.15d, 5.15e e 5.15f ilustram os resultados obtidos para as métricas para a base *Iris*. Através das imagens é possível verificar que o perfil da métrica proposta BR é parecido com o das métricas PC, PE e MPC, que não dependem da informação de distâncias entre grupos. Pode-se observar também que as métricas FS e XB sofrem grande influência da reamostragem dos dados, resultando em índices que se aproximam da aleatoriedade. O histrograma da Figura 5.16 demonstra estes resultados obtidos, onde fica claro a quase uniformidade dos resultados para as métricas FS e XB, assim como a convergência em $c = 2$ para as métricas BR, PC, PE e MPC. Esta convergência é também coerente com as matrizes de proximidade da Figura 5.17, onde a matriz de proximidade para $c = 2$ se destaca pela uniformidade dos agrupamentos obtidos.

Para a base de dados *Wine*, é possível observar os resultados das métricas nas Figuras 5.18a, 5.18b, 5.18c, 5.18d, 5.18e, 5.18f. O histograma da Figura 5.19 mostra que a métrica proposta BR, juntamente com as métricas PC e PE, encontra $c = 2$ em todas as rodadas. A métrica MPC fica dividida entre $c = 2$ e $c = 3$, enquanto a métrica XB encontra $c = 3$ para quase todos os casos. Já a métrica FS novamente não é capaz de encontrar uma resposta definitiva, apesar de se perceber uma tendência para o alto número de grupos. Apesar dos resultados encontrados pelas métricas, ao visualizar as matrizes de proximidade para a base *Wine* na Figura 5.20, não é possível se verificar a existência de agrupamentos convincentes para nenhum valor de c. Este é um dos casos em que, apesar das métricas apontarem para um determinado valor de c, a visualização da matriz de proximidade revela a verdadeira estrutura dos dados, o que leva a concluir que a base de dados, com os atributos e métricas utilizadas, não possui agrupamentos baseados em centróides, não obedecendo a premissa de grupos compactos.

Em relação à base *Glass* , as métricas estão na Figura 5.21. É possível observar que o método proposto BR se assemelha aos perfis de PC e PE, enquanto MPC, FS e XB apresentam grandes variações. Através do histograma dos resultados da Figura 5.22 observa-se que BR, PC, PE e MPC concordam em todas as 50 repetições com o valor $c = 2$, enquanto FS e XB novamente apresentam vários valores de c como resultado. É possível confirmar estes resultados através da matriz de proximidade da Figura 5.23a, que apesar de apresentar uma certa interação entre os grupos apresenta-se como uma matriz de proximidade coerente com os resultados obtidos pelas métricas.

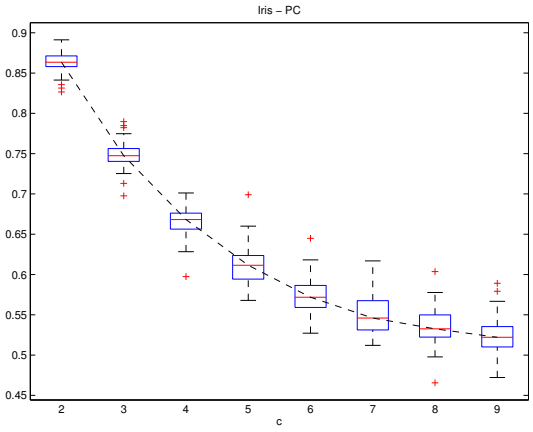
Para a última base analisada *Wdbc*, os resultados das métricas se encontram na Figura 5.24. É possível observar uma menor variação entre as séries de cada métrica,

efeito que pode ser explicado pelo fato da base *Wdbc* possuir mais observações que as demais. O histograma dos resultados pode ser visualizado na Figura 5.25. A métrica proposta BR e as métricas PC, PE, MPC e XB concordam com o valor $c = 2$ para todas as repetições. Já a métrica FS resulta em $c = 9$ para a maioria dos casos. Pela matriz de proximidade da Figura 5.26a é possível concluir que os resultados obtidos são coerentes.

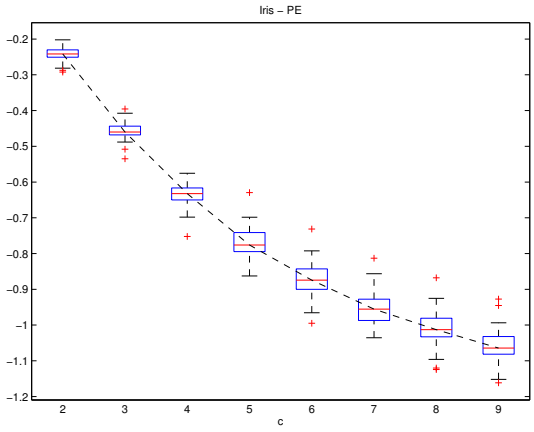
5.4 Conclusões do Capítulo

Neste capítulo, uma série de experimentos foram conduzidos para avaliar a eficácia do método proposto para encontrar o número de grupos compactos em bases de dados reais e sintéticos. Os resultados para as bases sintéticas e reais mostraram que a métrica desenvolvida possui desempenho similar a outras métricas da literatura, sendo inclusive robusta em relação a problemas de agrupamentos desbalanceados.

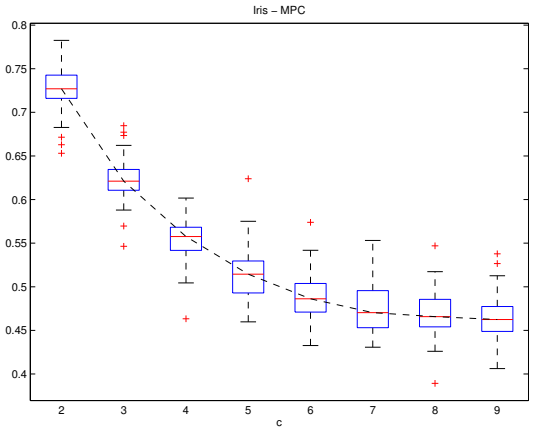
Adicionalmente, foi também mostrada a eficácia da etapa de visualização da matriz de proximidade para validar os resultados encontrados pelas métricas. Tem-se, por exemplo, o caso da base *Wine* onde apesar de existir coerência entre as métricas, foi possível observar em suas matrizes de proximidade que os dados não possuem estrutura de agrupamentos compactos, invalidando o resultado encontrado de $c = 2$.



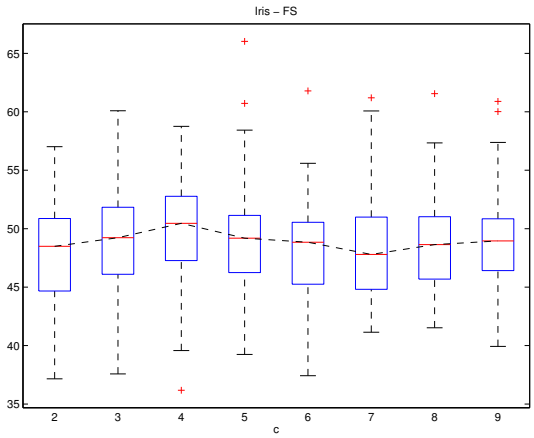
(a) Resultado para a métrica PC na base de dados *Iris*.



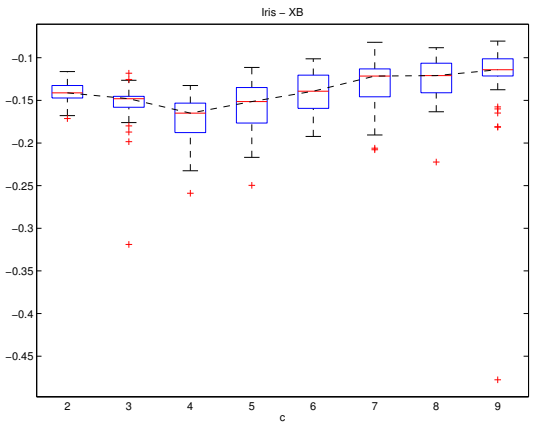
(b) Resultado para a métrica PE na base de dados *Iris*.



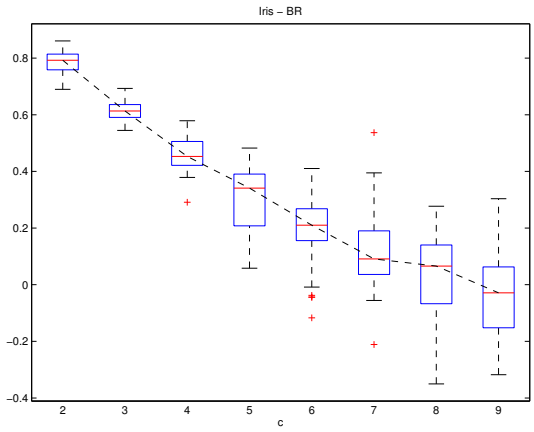
(c) Resultado para a métrica MPC na base de dados *Iris*.



(d) Resultado para a métrica FS na base de dados *Iris*.



(e) Resultado para a métrica XB na base de dados *Iris*.



(f) Resultado para a métrica BR na base de dados *Iris*.

Figura 5.15: Resultados das métricas para a base de dados Iris.

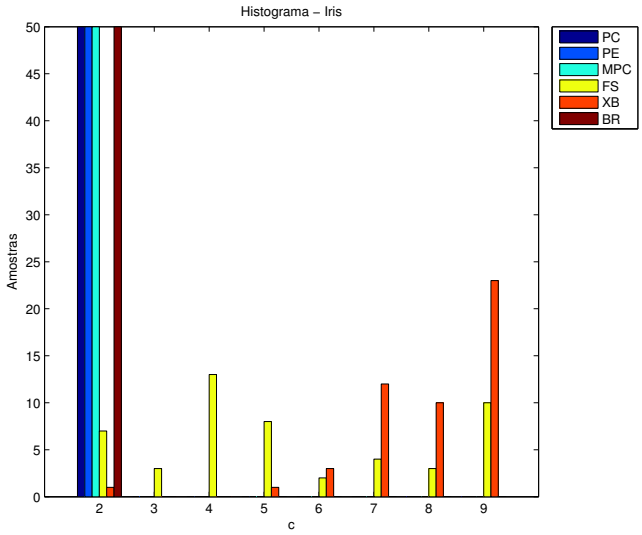
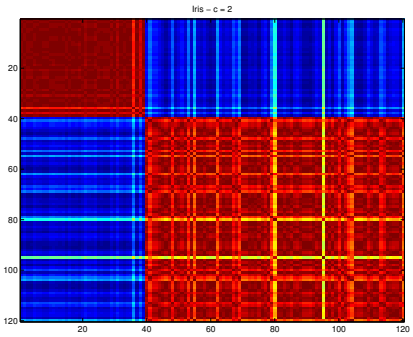
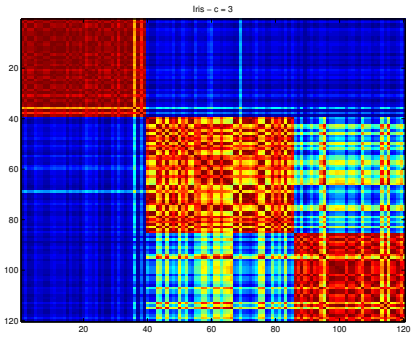


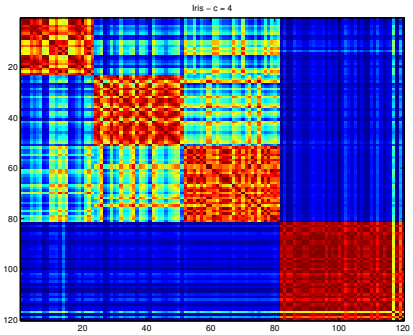
Figura 5.16: Histograma dos resultados para a base de dados *Iris*.



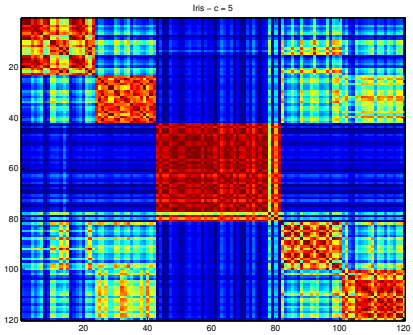
(a) Matriz de Proximidade para a base de dados *Iris* - $c = 2$.



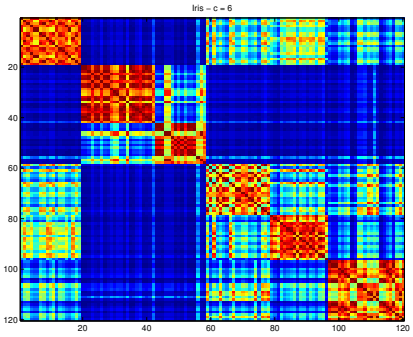
(b) Matriz de Proximidade para a base de dados *Iris* - $c = 3$.



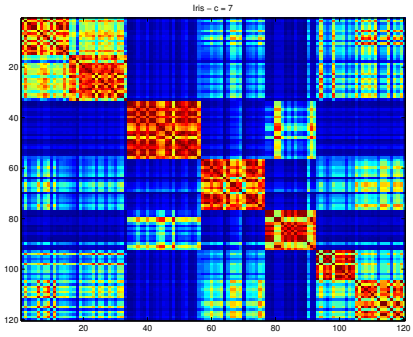
(c) Matriz de Proximidade para a base de dados *Iris* - $c = 4$.



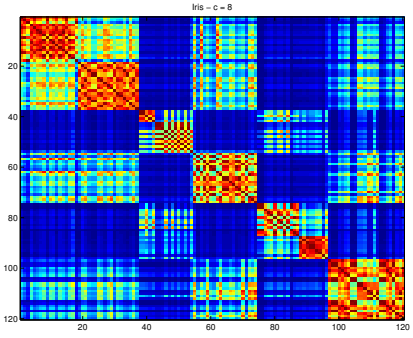
(d) Matriz de Proximidade para a base de dados *Iris* - $c = 5$.



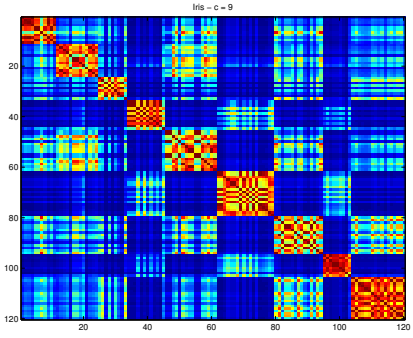
(e) Matriz de Proximidade para a base de dados *Iris* - $c = 6$.



(f) Matriz de Proximidade para a base de dados *Iris* - $c = 7$.

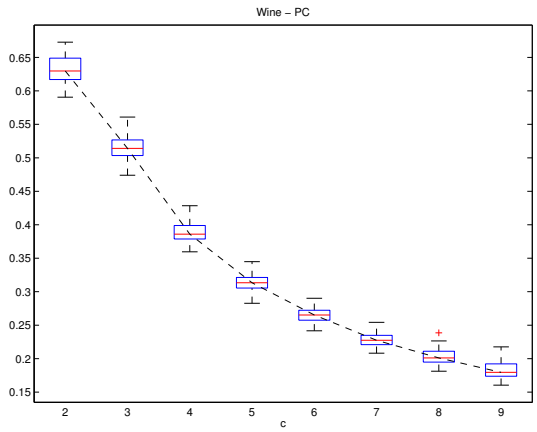


(g) Matriz de Proximidade para a base de dados *Iris* - $c = 8$.

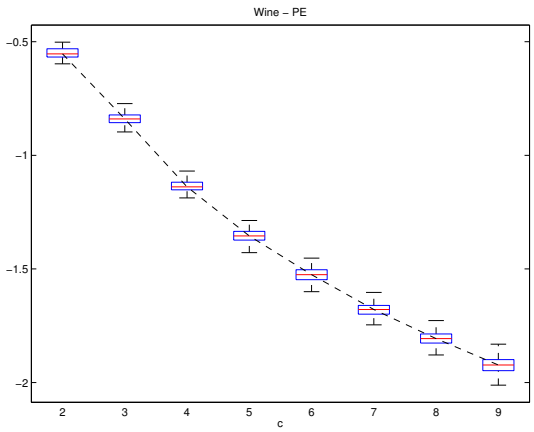


(h) Matriz de Proximidade para a base de dados *Iris* - $c = 9$.

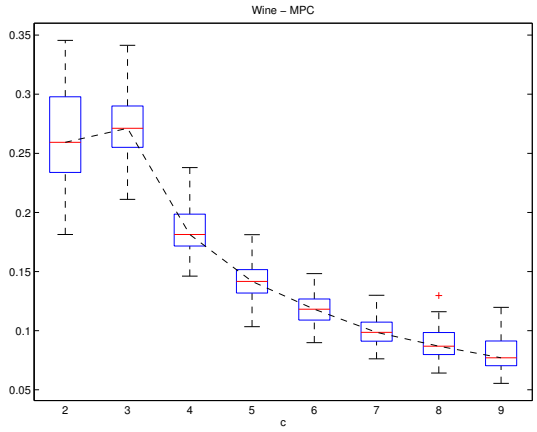
Figura 5.17: Matrizes de Proximidade para a base de dados *Iris*.



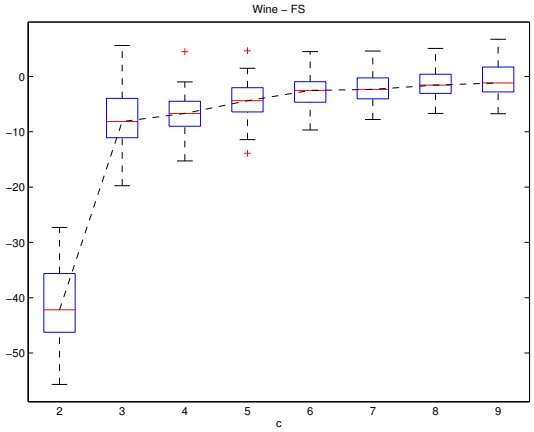
(a) Resultado para a métrica PC na base de dados *Wine*.



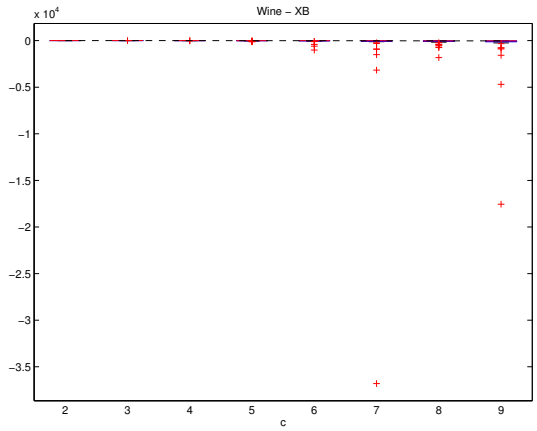
(b) Resultado para a métrica PE na base de dados *Wine*.



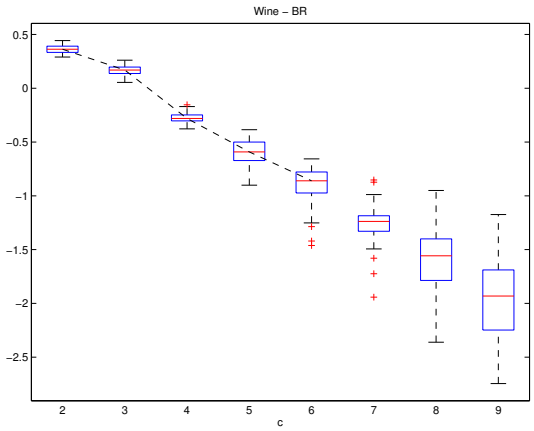
(c) Resultado para a métrica MPC na base de dados *Wine*.



(d) Resultado para a métrica FS na base de dados *Wine*.



(e) Resultado para a métrica XB na base de dados *Wine*.



(f) Resultado para a métrica BR na base de dados *Wine*.

Figura 5.18: Resultados das métricas para a base de dados *Wine*.

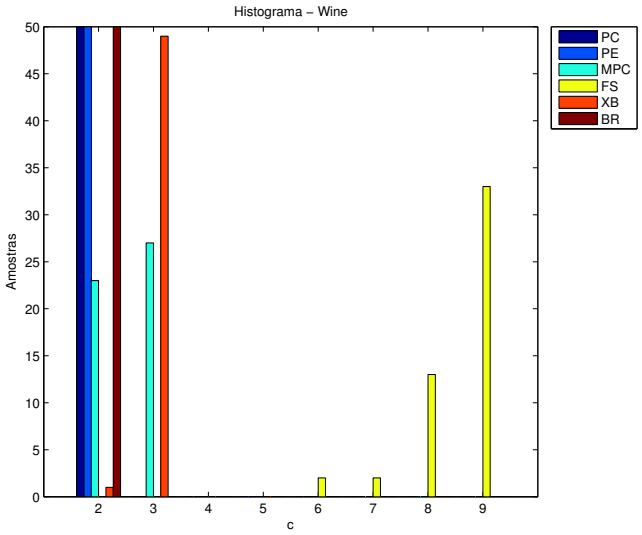
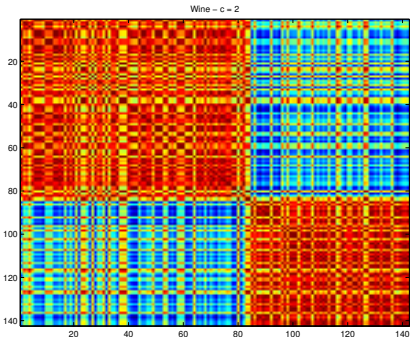
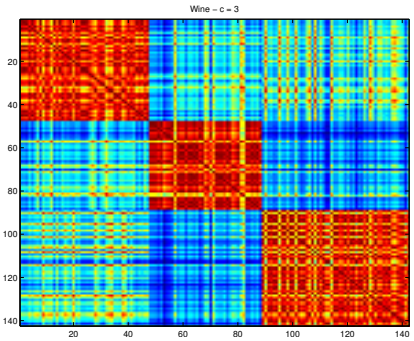


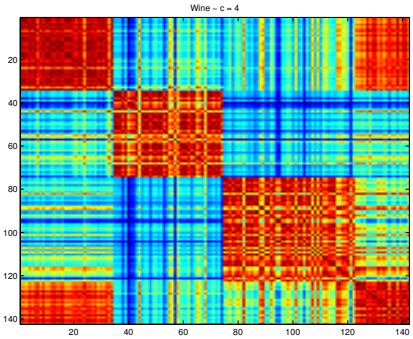
Figura 5.19: Histograma dos resultados para a base de dados *Wine*.



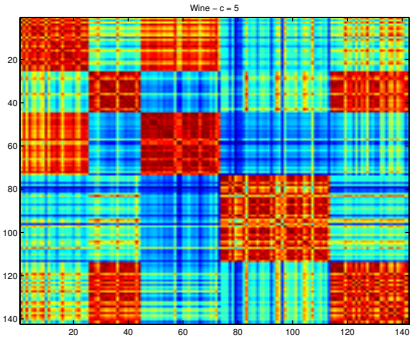
(a) Matriz de Proximidade para a base de dados *Wine* - $c = 2$.



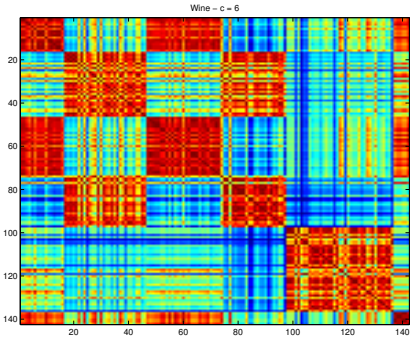
(b) Matriz de Proximidade para a base de dados *Wine* - $c = 3$.



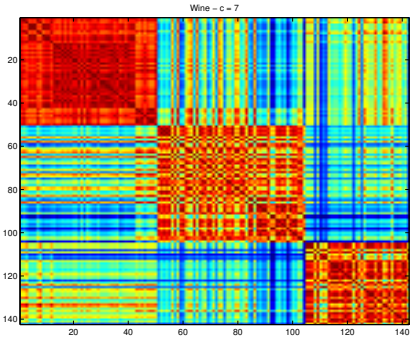
(c) Matriz de Proximidade para a base de dados *Wine* - $c = 4$.



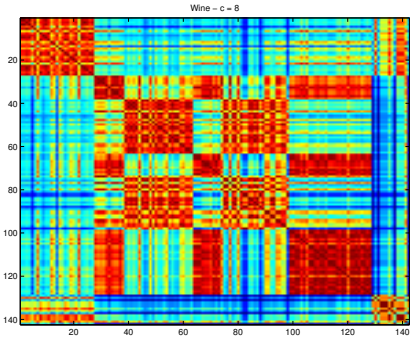
(d) Matriz de Proximidade para a base de dados *Wine* - $c = 5$.



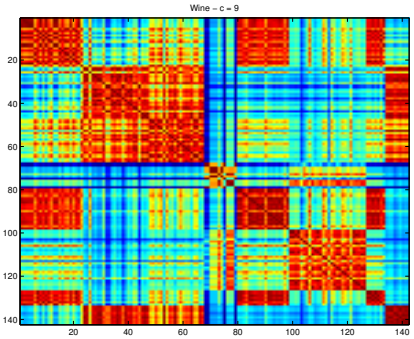
(e) Matriz de Proximidade para a base de dados *Wine* - $c = 6$.



(f) Matriz de Proximidade para a base de dados *Wine* - $c = 7$.

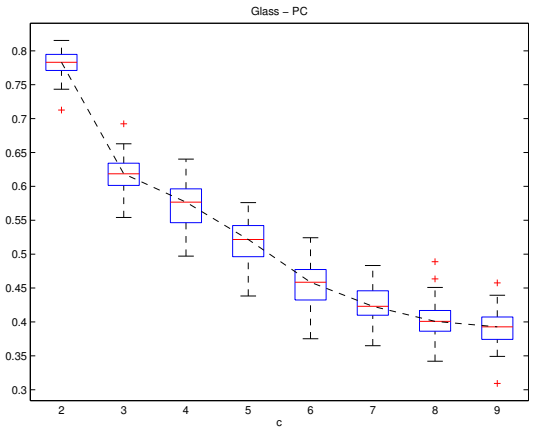


(g) Matriz de Proximidade para a base de dados *Wine* - $c = 8$.

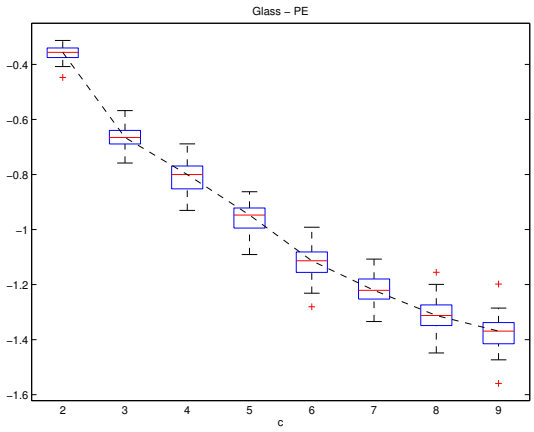


(h) Matriz de Proximidade para a base de dados *Wine* - $c = 9$.

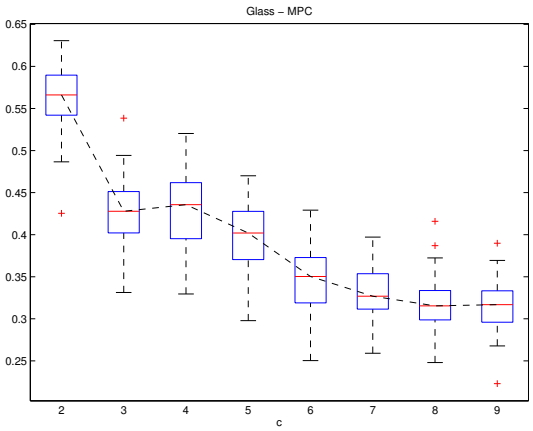
Figura 5.20: Matrizes de Proximidade para a base de dados *Wine*.



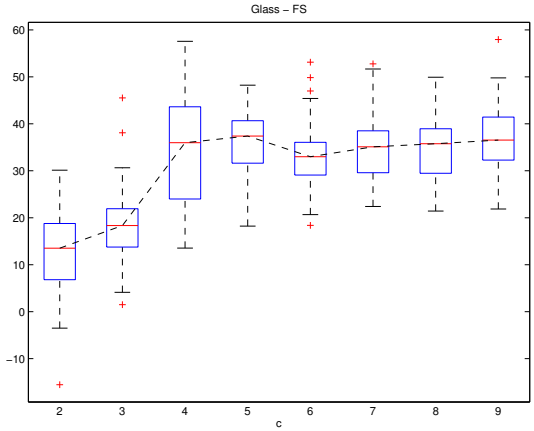
(a) Resultado para a métrica PC na base de dados *Glass*.



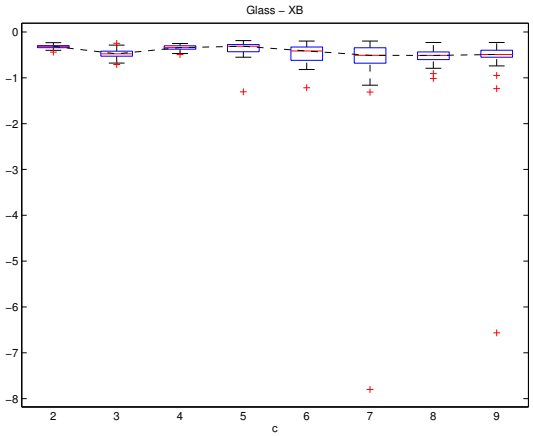
(b) Resultado para a métrica PE na base de dados *Glass*.



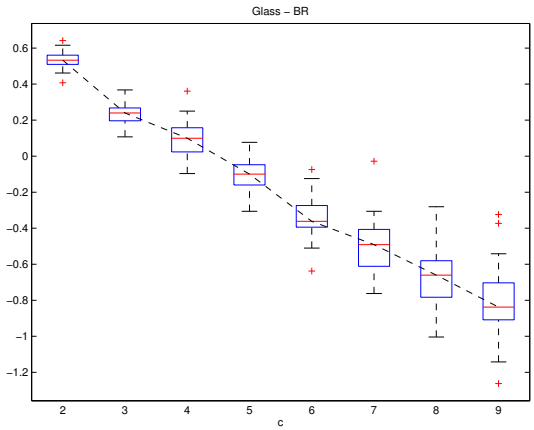
(c) Resultado para a métrica MPC na base de dados *Glass*.



(d) Resultado para a métrica FS na base de dados *Glass*.



(e) Resultado para a métrica XB na base de dados *Glass*.



(f) Resultado para a métrica BR na base de dados *Glass*.

Figura 5.21: Resultados das métricas para a base de dados *Glass*.

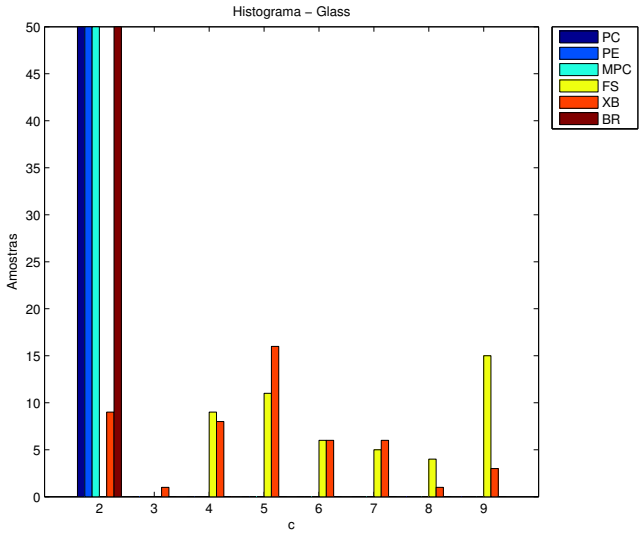
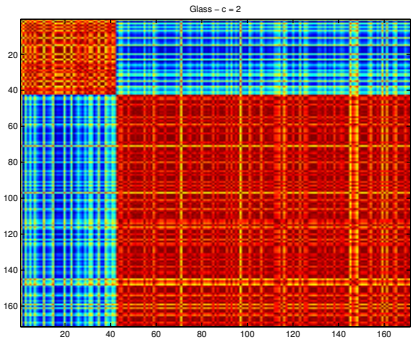
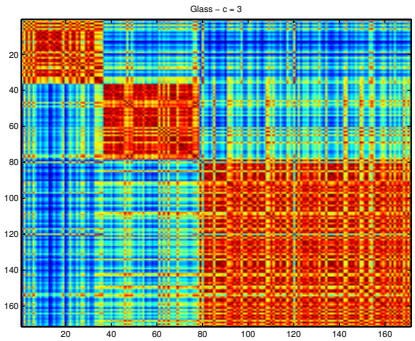


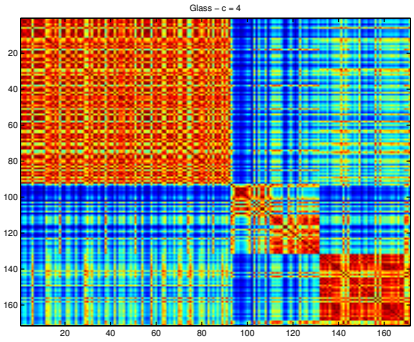
Figura 5.22: Histograma dos resultados para a base de dados *Glass*.



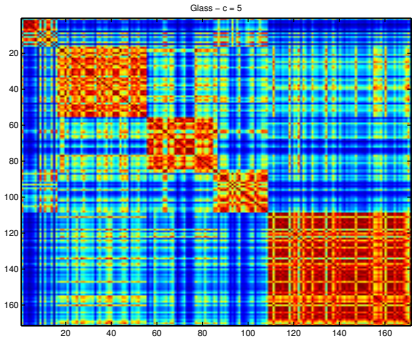
(a) Matriz de Proximidade para a base de dados *Glass* - $c = 2$.



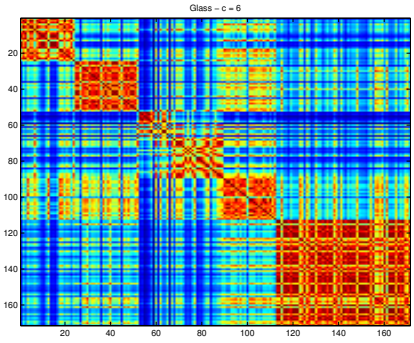
(b) Matriz de Proximidade para a base de dados *Glass* - $c = 3$.



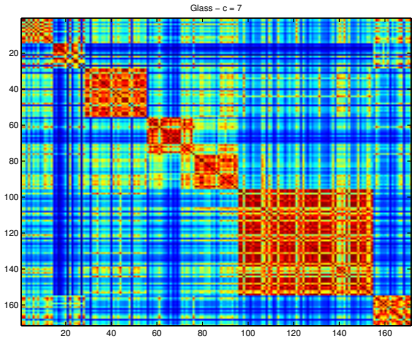
(c) Matriz de Proximidade para a base de dados *Glass* - $c = 4$.



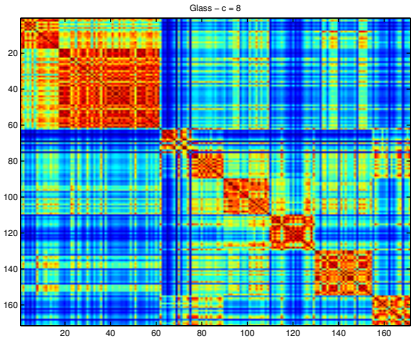
(d) Matriz de Proximidade para a base de dados *Glass* - $c = 5$.



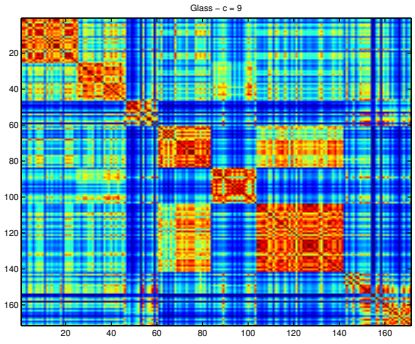
(e) Matriz de Proximidade para a base de dados *Glass* - $c = 6$.



(f) Matriz de Proximidade para a base de dados *Glass* - $c = 7$.

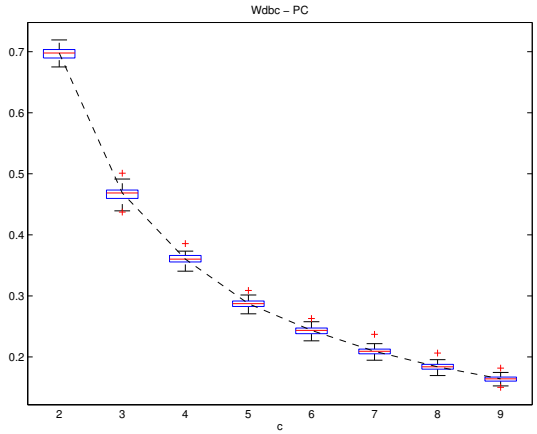


(g) Matriz de Proximidade para a base de dados *Glass* - $c = 8$.

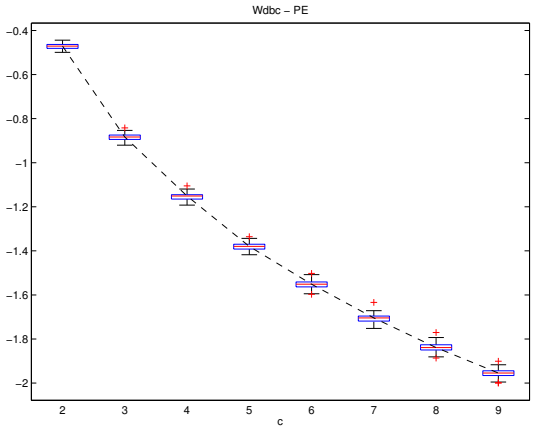


(h) Matriz de Proximidade para a base de dados *Glass* - $c = 9$.

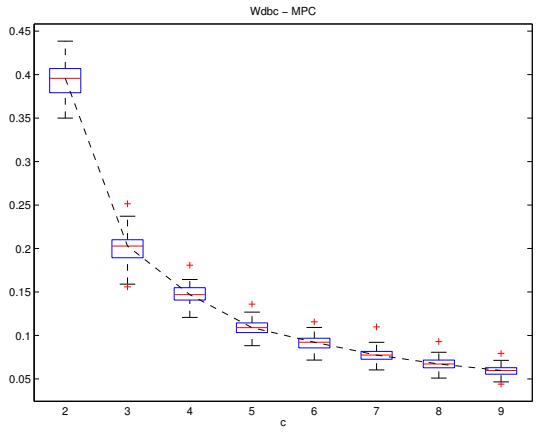
Figura 5.23: Matrizes de Proximidade para a base de dados *Glass*.



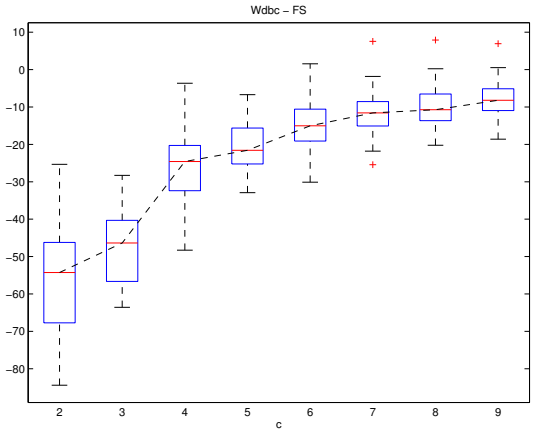
(a) Resultado para a métrica PC na base de dados *Wdbc*.



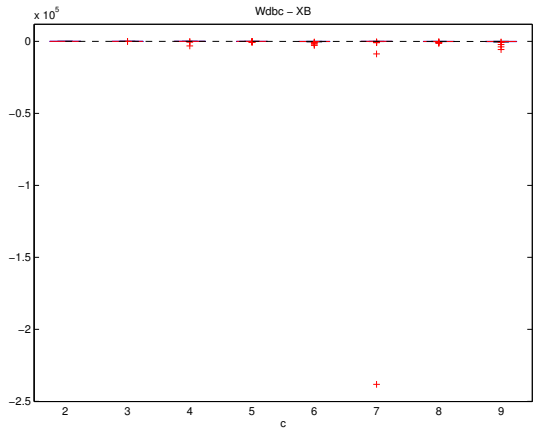
(b) Resultado para a métrica PE na base de dados *Wdbc*.



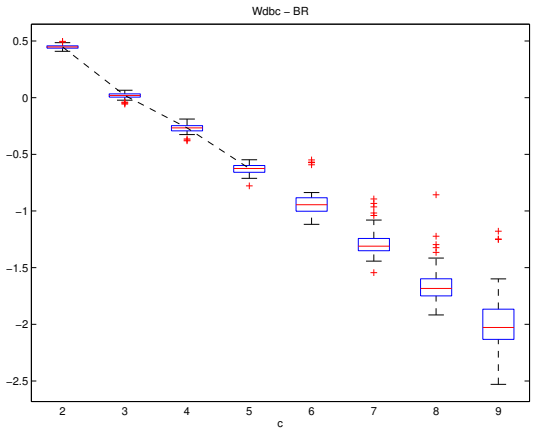
(c) Resultado para a métrica MPC na base de dados *Wdbc*.



(d) Resultado para a métrica FS na base de dados *Wdbc*.



(e) Resultado para a métrica XB na base de dados *Wdbc*.



(f) Resultado para a métrica BR na base de dados *Wdbc*.

Figura 5.24: Resultados das Métricas para a base de dados *Wdbc*.

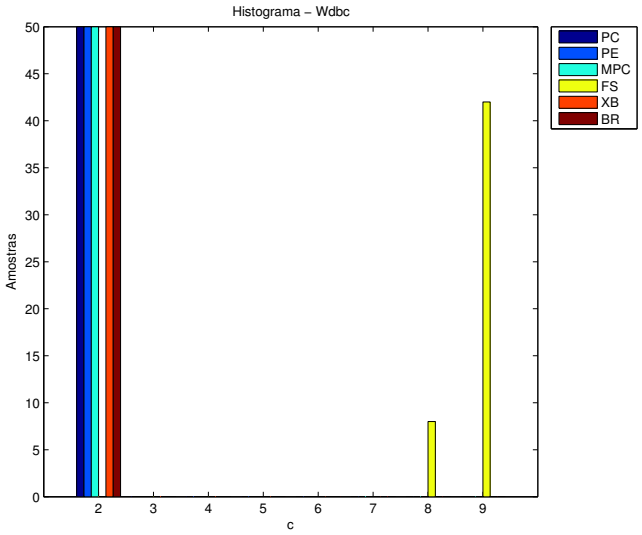
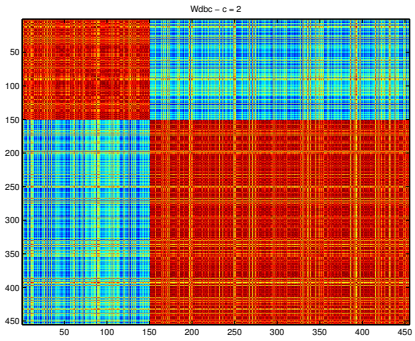
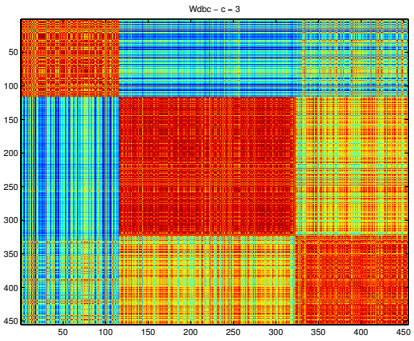


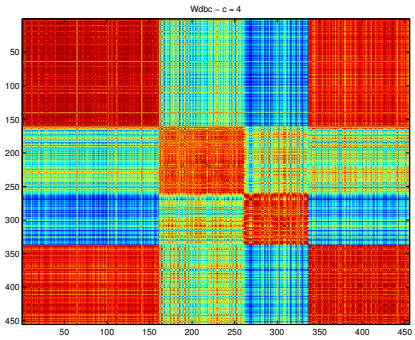
Figura 5.25: Histograma dos resultados para a base de dados *Wdbc*.



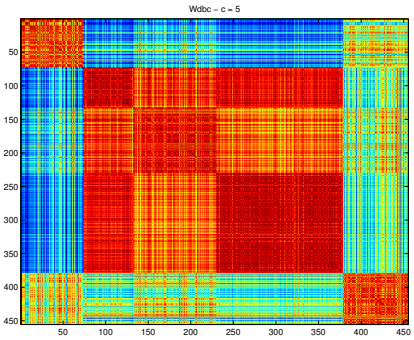
(a) Matriz de Proximidade para a base de dados *Wdbc* - $c = 2$.



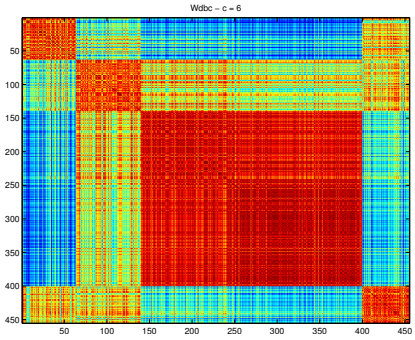
(b) Matriz de Proximidade para a base de dados *Wdbc* - $c = 3$.



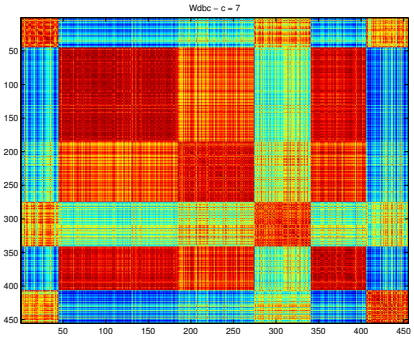
(c) Matriz de Proximidade para a base de dados *Wdbc* - $c = 4$.



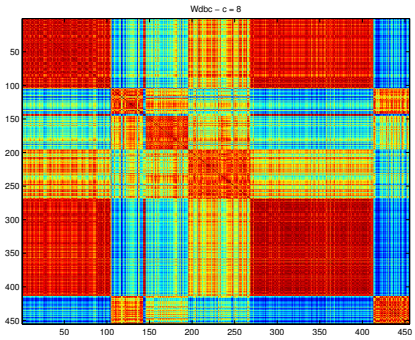
(d) Matriz de Proximidade para a base de dados *Wdbc* - $c = 5$.



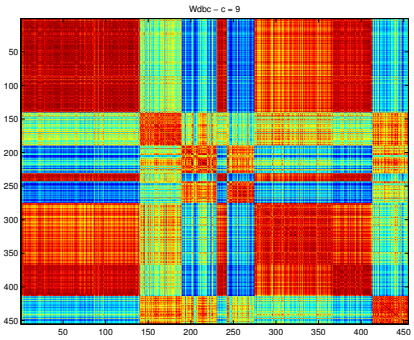
(e) Matriz de Proximidade para a base de dados *Wdbc* - $c = 6$.



(f) Matriz de Proximidade para a base de dados *Wdbc* - $c = 7$.



(g) Matriz de Proximidade para a base de dados *Wdbc* - $c = 8$.



(h) Matriz de Proximidade para a base de dados *Wdbc* - $c = 9$.

Figura 5.26: Matrizes de Proximidade para a base de dados *Wdbc*.

Capítulo 6

Conclusões e Propostas de Continuidade

Esta dissertação abordou aspectos teóricos e práticos da análise de agrupamentos, particularmente, o problema de encontrar o número de grupos em bases de dados não rotuladas. Primeiramente, foi realizada uma revisão do processo geral de análise de agrupamentos, considerando as etapas de representação dos dados, escolha das métricas de proximidade, tipos de algoritmos e validação. Em seguida, foram apresentados os métodos existentes na literatura para lidar com a seleção do número de grupos. Entre eles foram discutidas as abordagens que utilizam teoria da informação, construção de *ensembles*, estatística e grafos. Foi mostrado que a função objetivo do algoritmo FCM J_{FCM} apresenta um comportamento descendente com o aumento do número de grupos c, o que a torna insuficiente para a escolha da quantidade de agrupamentos. Para ilustrar esta questão foi utilizado um problema sintético para traçar a curva dos valores obtidos para várias repetições do algoritmo. Através desta análise foi discutido o paralelo existente entre o conflito do número de grupos e a função objetivo J_{FCM} com o dilema viés-variância dos problemas de aprendizado supervisionado.

As ideias conceituais provenientes da análise do comportamento da função objetivo J_{FCM} foram então utilizadas para a formulação de um novo método para a validação de agrupamentos. A essência da nova abordagem baseia-se na noção intuitiva de que os elementos de um mesmo grupo devem possuir alta magnitude de proximidade, enquanto as similaridades dos elementos de diferentes grupos devem ser de baixa magnitude. A métrica proposta é construída através de medidas estatísticas calculadas da matriz de proximidade *fuzzy*.

Através de experimentos com bases de dados sintéticas e reais foi possível demonstrar a validade do método proposto. Nos experimentos controlados, onde a função

geradora dos dados $P(\mathbf{X})$ era conhecida, a métrica proposta mostrou-se coerente com outras métricas da literatura, sendo inclusive superior para o caso de grupos desbalanceados. Mas para o caso de agrupamentos com superposição, o método não foi capaz de detectar corretamente o número de funções geradoras. Para os experimentos em bases reais, os resultados também foram consistentes em relação às outras métricas da literatura.

Além disto, a análise qualitativa das matrizes de proximidade mostrou-se como uma maneira prática de inferir a qualidade de uma determinada partição para um conjunto de dados. Ela mostrou-se particularmente útil quando utilizada em dados reais e complexos, situações nas quais as funções geradoras $P(\mathbf{X})$ são desconhecidas. A análise visual das matrizes de proximidade apresentada é útil também para identificar padrões desconhecidos na estrutura espacial dos dados.

Por fim, espera-se que os resultados do presente estudo, em termos dos conceitos teóricos e práticos apresentados, possam ser aplicados em problemas reais de análise de agrupamentos, bem como possam servir como base para o desenvolvimento de novas métricas de validação de agrupamentos.

6.1 Propostas de Continuidade

Segere-se como propostas de continuidade deste trabalho, investir nos seguintes problemas relacionados ao tema:

- A visualização das matrizes de proximidade se mostrou uma ferramenta poderosa para verificar a qualidade das partições geradas pelo processo de agrupamento. Porém, as ideias apresentadas em [Tsafrir et al., 2005] e [Võhandu et al., 2006] para reordenação de linhas e colunas de matrizes de proximidade, poderiam ser utilizadas para reordenar as submatrizes do método proposto, possibilitando ao usuário final a visualização de outros detalhes.

- O desenvolvimento do método proposto nesta dissertação se baseou no cálculo de medidas estatísticas baseadas nas matrizes de proximidade. Estas funções foram modeladas como sendo uma representação da homogeinização média das magnitudes de similaridade dos elementos de um mesmo grupos e elementos de diferentes grupos. Conforme descrito em [Duda et al., 2000], a variabilidade das magnitudes também pode ser utilizada como critério em tarefas de agrupamento. Assim, torna-se interessante e promissora a ideia de investigar os efeitos da utilização da informação de variabilidade das magnitudes.

Referências Bibliográficas

Ayad, H. G. & Kamel, M. S. (2008). Cumulative voting consensus method for partitions with variable number of clusters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(1):160--173.

Ayad, H. G. & Kamel, M. S. (2010). On voting-based consensus of cluster ensembles. *Pattern Recognition*, 43(5):1943--1953.

Bache, K. & Lichman, M. (2013). Uci machine learning repository.

Barzily, Z.; Volkovich, Z.; Akteke-Öztürk, B. & Weber, G.-W. (2009). On a minimal spanning tree approach in the cluster validation problem. *Informatica*, 20(2):187--202.

Ben-Hur, A.; Horn, D.; Siegelmann, H. T. & Vapnik, V. (2002). Support vector clustering. *The Journal of Machine Learning Research*, 2:125--137.

Bezdek, J. C. (1973). Cluster validity with fuzzy sets.

Bezdek, J. C. (1975). Mathematical models for systematics and taxonomy. Em *Proceedings of Eighth International Conference on Numerical Taxonomy*, volume 3, pp. 143--166.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers.

Bezdek, J. C. & Pal, N. R. (1995). Cluster validation with generalized dunn's indices. Em *Artificial Neural Networks and Expert Systems, 1995. Proceedings., Second New Zealand International Two-Stream Conference on*, pp. 190--193. IEEE.

Bezdek, J. C. & Pal, N. R. (1998). Some new indexes of cluster validity. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 28(3):301--315.

Boudraa, A.-O. (1999). Dynamic estimation of number of clusters in data sets. *Electronics Letters*, 35(19):1606--1608.

Bouguessa, M.; Wang, S. & Sun, H. (2006). An objective approach to cluster validation. *Pattern Recognition Letters*, 27(13):1419--1430.

Celeux, G. & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13(2):195--212.

Chen, K. & Liu, L. (2003). A visual framework invites human into the clustering process. Em *Scientific and Statistical Database Management, 2003. 15th International Conference on*, pp. 97--106. IEEE.

Chen, K. & Liu, L. (2004). Vista: Validating and refining clusters via visualization. *Information Visualization*, 3(4):257--270.

Dave, R. N. (1996). Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters*, 17(6):613--623.

Duda, R. O.; Hart, P. E. & Stork, D. G. (2000). *Pattern Classification (2nd Edition)*. Wiley-Interscience. ISBN 0471056693.

Ester, M.; peter Kriegel, H.; S, J. & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. pp. 226--231. AAAI Press.

Faloutsos, C. & Lin, K.-I. (1995). *FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets*, volume 24. ACM.

Filippone, M.; Camastra, F.; Masulli, F. & Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern recognition*, 41(1):176--190.

Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611--631.

Fred, A. (2001). Finding consistent clusters in data partitions. Em *Multiple classifier systems*, pp. 309--318. Springer.

Fukuyama, Y. & Sugeno, M. (1989). A new method of choosing the number of clusters for the fuzzy c-means method. Em *Proc. 5th Fuzzy Syst. Symp*, volume 247.

Gan, G.; Ma, C. & Wu, J. (2007). *Data clustering: theory, algorithms, and applications*, volume 20. Siam.

Geman, S.; Bienenstock, E. & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1--58.

Ghosh, J. & Acharya, A. (2011). Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):305--315.

Gokcay, E. & Principe, J. C. (2002). Information theoretic clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(2):158--171.

Gordon, A. (1999). Classification. 1999. *Chapman&Hall, CRC, Boca Raton, FL*.

Gower, J. C. & Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of classification*, 3(1):5--48.

Günter, S. & Bunke, H. (2003). Validation indices for graph clustering. *Pattern Recognition Letters*, 24(8):1107--1113.

Guyon, I. (2006). *Feature extraction: foundations and applications*, volume 207. Springer.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157--1182.

Halkidi, M.; Batistakis, Y. & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107--145.

Hamerly, Y. F. G. (2007). Pg-means: learning the number of clusters in data. Em *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, p. 393. MIT Press.

Huang, Z.; Cheung, D. W. & Ng, M. K. (2001). An empirical study on the visual cluster validation method with fastmap. Em *Database Systems for Advanced Applications, 2001. Proceedings. Seventh International Conference on*, pp. 84--91. IEEE.

Huang, Z. & Lin, T. (2000). A visual method of cluster validation with fastmap. Em *Knowledge Discovery and Data Mining. Current Issues and New Applications*, pp. 153--164. Springer.

Ichino, M. & Yaguchi, H. (1994). Generalized minkowski metrics for mixed feature-type data analysis. *Systems, Man and Cybernetics, IEEE Transactions on*, 24(4):698--708.

Izakian, H. & Pedrycz, W. (2013). Agreement-based fuzzy c-means for clustering data with blocks of features. *Neurocomputing*, (0):–. ISSN 0925-2312.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651--666.

Jain, A. K.; Murty, M. N. & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264--323.

Jenssen, R.; Hild, K.; Erdogmus, D.; Principe, J. C.; Eltoft, T. et al. (2003). Clustering using renyi's entropy. Em *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 1, pp. 523--528. IEEE.

Jolliffe, I. (2005). *Principal component analysis*. Wiley Online Library.

Kaufman, L. & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis.

Kaufman, L. & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley-Interscience.

Kim, D.-W.; Lee, K. H. & Lee, D. (2003). Fuzzy cluster validation index based on inter-cluster proximity. *Pattern Recognition Letters*, 24(15):2561--2574.

Kim, M.; Yoo, H. & Ramakrishna, R. (2004a). Cluster validation for high-dimensional datasets. Em *Artificial Intelligence: Methodology, Systems, and Applications*, pp. 178--187. Springer.

Kim, Y.-I.; Kim, D.-W.; Lee, D. & Lee, K. H. (2004b). A cluster validation index for gk cluster analysis based on relative degree of sharing. *Information Sciences*, 168(1):225--242.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464--1480.

Kothari, R. & Pitts, D. (1999). On finding the number of clusters. *Pattern Recognition Letters*, 20(4):405--416.

Kwon, S. H. (1998). Cluster validity index for fuzzy clustering. *Electronics Letters*, 34(22):2176--2177.

Lange, T.; Roth, V.; Braun, M. L. & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299--1323.

Li, M. J.; Ng, M. K.; Cheung, Y.-M. & Huang, J. Z. (2008). Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *Knowledge and Data Engineering, IEEE Transactions on*, 20(11):1519--1534.

Lindsten, F.; Ohlsson, H. & Ljung, L. (2011). Just relax and come clustering! a convexification of k-means clustering. Relatório técnico, Tech. Rep. LiTH-ISY.

Loia, V.; Pedrycz, W. & Senatore, S. (2003). Proximity fuzzy clustering for web context analysis. Em *EUSFLAT Conf.*, pp. 59--62.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. Em *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, p. 14. California, USA.

Marler, R. T. & Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6):369--395.

McLachlan, G. & Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159--179.

Ng, A. Y.; Jordan, M. I.; Weiss, Y. et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849--856.

Ng, M. & Huang, J. (2002). M-fastmap: A modified fastmap algorithm for visual cluster validation in data mining. Em *Advances in Knowledge Discovery and Data Mining*, pp. 224--236. Springer.

Pakhira, M. K.; Bandyopadhyay, S. & Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern recognition*, 37(3):487--501.

Pal, N. R. & Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. *Fuzzy Systems, IEEE Transactions on*, 3(3):370--379.

Pal, N. R. & Biswas, J. (1997). Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6):847--857.

Pascual, D.; Pla, F. & Sánchez, J. S. (2010). Cluster validation using information stability measures. *Pattern Recognition Letters*, 31(6):454--461.

Pękalska, E. & Duin, R. P. (2005). *The dissimilarity representation for pattern recognition: foundations and applications*. Number 64. World Scientific.

Queiroz, F. A.; Braga, A. P. & Pedrycz, W. (2009). Sorted kernel matrices as cluster validity indexes. Em *2009 IFSA World Congress*, volume 1.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846--850.

Rendón, E.; Abundez, I.; Arizmendi, A. & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27--34.

Roth, V.; Lange, T.; Braun, M. & Buhmann, J. (2002). A resampling approach to cluster validation. Em *Compstat*, pp. 123--128. Springer.

Rousseeuw, L. & Kaufman, L. (1987). Clustering by means of medoids. *Statistical data analysis based on the L1-norm and related methods*, 405.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53--65.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27--64.

Shannon, C. E. & Weaver, W. (1949). The mathematical theory of communication (urbana, il. *University of Illinois Press*, 19(7):1.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Sun, H.; Wang, S. & Jiang, Q. (2004). Fcm-based model selection algorithms for determining the number of clusters. *Pattern recognition*, 37(10):2027--2037.

Tang, Y.; Sun, F. & Sun, Z. (2005). Improved validation index for fuzzy clustering. Em *American Control Conference, 2005. Proceedings of the 2005*, pp. 1120--1125. IEEE.

Tibshirani, R. & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511--528.

Tibshirani, R.; Walther, G. & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411--423.

Tsafrir, D.; Tsafrir, I.; Ein-Dor, L.; Zuk, O.; Notterman, D. A. & Domany, E. (2005). Sorting points into neighborhoods (spin): data analysis and visualization by ordering distance matrices. *Bioinformatics*, 21(10):2301--2308.

Valente, R. X.; Braga, A. P. & Pedrycz, W. (2013). A new fuzzy clustering validity index based on fuzzy proximity matrices. Brasil, Porto de Galhinhas. BRICS CCI, Conference Publishing Services (CPS).

Vapnik, V. N. (1998). Statistical learning theory.

Vega-Pons, S. & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337--372.

Võhandu, L.; Kuusik, R.; Torim, A.; Aab, E. & Lind, G. (2006). Some algorithms for data table (re) ordering using monotone systems. Em *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases: 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'06)*, pp. 417--422.

Wang, W. & Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy sets and systems*, 158(19):2095--2117.

Wu, K.-L. & Yang, M.-S. (2005). A cluster validity index for fuzzy clustering. *Pattern Recognition Letters*, 26(9):1275--1291.

Xie, X. L. & Beni, G. (1991). A validity measure for fuzzy clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(8):841--847.

Xu, R. & Wunsch, D. (2008). *Clustering*, volume 10. Wiley. com.

Xu, R.; Wunsch, D. et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645--678.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338--353.