Pós-Graduação em Ciência da Computação

DANIEL BION BARREIROS

AGRUPAMENTO DE DADOS INTERVALARES USANDO UMA ABORDAGEM NÃO LINEAR

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2016

Daniel Bion Barreiros

Agrupamento de dados Intervalares usando uma abordagem não linear

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

ORIENTADOR(A): **Profa. Dra. Renata Maria Cardoso Rodrigues de Souza**
CO-ORIENTADOR(A): **Prof. Dr. Marco Antonio de Oliveira Domingues**

RECIFE
2016

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

B271a Barreiros, Daniel Bion
Agrupamento de dados intervalares usando uma abordagem não linear / Daniel Bion Barreiros. – 2016.
48 f.: il., fig., tab.

Orientadora: Renata Maria Cardoso Rodrigues de Souza.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2016.
Inclui referências.

1. Inteligência computacional. 2. Análise de dados simbólicos. I. Souza, Renata Maria Cardoso Rodrigues de (orientadora). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2017-96

Daniel Bion Barreiros

Agrupamento de dados Intervalares usando uma abordagem não linear

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 24/08/2016

BANCA EXAMINADORA

Profa.Dra.Renata Maria Cardoso Rodrigues de Souza
Centro de Informática / UFPE
(**Orientadora**)

Profa.Dra.Roberta Andrade de Araújo Fagundes
Escola Politécnica

Profa.Dra.Caliteia Santana de Sousa
Departamento de Estatística/UFPE

Dedico esta dissertação a toda minha família,
companheiros de estudo e professores que me deram apoio
durante esta caminhada.

Agradecimentos

Gostaria de agradecer a todas as pessoas que mantive contato durante os últimos anos, e que de alguma forma contribuiram para o desenvolvimento deste trabalho.

Primeiramente agradeço ao Centro de Informática da Universidade Federal de Pernambuco pela oportunidade de poder participar do programa de mestrado.

Agradeço aos meus colegas da universidade Valter Eduardo da Silva Júnior e Bruno Almeida Pimeltel por toda troca de conhecimento durante esses últimos anos.

Agradeço aos meus orientadores Renata Maria Cardoso Rodrigues de Souza e Marco Antonio de Oliveira Domingues pelo auxílio no direcionamento da pesquisa, e por todo conhecimento que me passaram.

Agradeço aos meus pais Robson do Carmelo Santos Barreiros e Denise Castanha Bion, e meus familiares Esmeraldino José Xavier Bion, Victor Bion Cantinha Lima pelo apoio necessário.

Agradeço à minha noiva Eduarda Fernandes Custódio da Silva e sua família, por todo carinho e por estarem ao meu lado durante toda esta caminhada.

A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo.

—ALBERT EINSTEIN

Resumo

A Análise de Dados Simbólicos (ADS) é uma abordagem da área de inteligência computacional que visa desenvolver métodos para dados descritos por variáveis onde existem conjuntos de categorias, intervalos ou distribuições de probabilidade. O objetivo deste trabalho é estender um método probabilístico de agrupamento clássicos para dados simbólicos intervalares fazendo uso de funções de núcleo. A aplicação de funções de núcleo tem sido utilizada com sucesso no agrupamento para dados clássicos apresentando resultados positivos quando o conjunto de dados apresenta grupos não linearmente separáveis. No entanto, a literatura de ADS precisa de métodos probabilísticos para identificar grupos não linearmente separáveis. Para mostrar a eficácia do método proposto, foram realizados experimentos com conjuntos de dados intervalares reais, e conjuntos sintéticos fazendo uso de simulações Monte Carlo. Também se apresenta um estudo comparando o método proposto com diferentes algoritmos de agrupamento da literatura através de estatísticas que evidenciam o desempenho superior do método proposto em determinados casos.

Palavras-chave: *Expectation Maximization*. Funções de núcleo para intervalo. Análise de dados simbólicos. Agrupamento não linear

Abstract

Symbolic Data Analysis (SDA) is a domain in the computational intelligence area that aims to provide suitable methods for data described through multi-valued variables, where there are sets of categories, intervals, histograms, or weight (probability) distributions. This work aims to extend a probabilistic clustering method of classic data to symbolic interval data making use of kernel functions. The kernel functions application have been successfully used in classic data clustering showing positive results when the data set has non linearly separable groups. However, SDA literature needs more probabilistic methods to identify non linearly separable groups. To show the effectiveness of the proposed method, experiments were performed with real interval data sets, and synthetic interval data sets using Monte Carlo simulations. It is also presented a study comparing the proposed method with different clustering algorithms of the literature through statistics that demonstrate the superior performance of the proposed method in certain cases.

Keywords: Expectation Maximization. Kernel functions for interval. Symbolic Data Analysis. Nonlinear clustering

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

Sumário

1

Introdução

Com o grande crescimento da área de tecnologia nos últimos anos sabe-se que a maioria dos aplicativos possuem ferramentas de análise de dados para entender o que seus usuários consomem, assim como para sugerir novos produtos ou serviços para os mesmos. Os grandes desenvolvedores de jogos online, utilizam ferramentas para validar os dados enviados ao servidor, para analisar a fuga dos usuários, tipo de público que acessa o jogo, etc. Acontece que muitas informações são obtidas todo dia, e é caro guardar essas informações em um servidor, e muitas vezes os dados são analisados e descartados. Os dados simbólicos são interessantes pela sua capacidade de sumarizar grandes bases de dados clássicos em novos conjuntos de dados simbólicos de tamanho menor. A Análise de Dados Simbólicos (ADS) (DIDAY, 2003) é uma área, que nasceu da influência simultânea de vários campos de pesquisa como: análise de dados clássica, inteligência computacional, aprendizagem de máquina e banco de dados. O principal objetivo de SDA é desenvolver modelos para o tratamento de dados mais complexos, como intervalos, conjuntos e distribuições de probabilidades ou de pesos. Além disso, ADS é capaz de generalizar os métodos tradicionais com dados clássicos para métodos com dados simbólicos através do desenvolvimento exploratórios, estatísticos e representações gráficas para esses tipos de dados.

A análise dados simbólicos é bastante interessante no contexto de inteligência computacional, devido a sua capacidade de modelar imprecisão, incerteza e variabilidade presente nos dados. Os métodos desenvolvidos para essa área são adequados para lidar com dados imprecisos, resultantes de medidas com imprecisão relativa ou estimadas por intervalos de confiança, limites de um conjunto de possíveis valores de um item ou variação da extensão de uma variável através do tempo. Considere, por exemplo, um paciente tem sua taxa de glicose no sangue acompanhada pelo seu médico. Um paciente saudável pode ter o valor da sua taxa de glicose oscilando no intervalo [70, 110]. Um outro paciente também saudável, poderia ter o valor da sua taxa de glicose oscilando no intervalo [90, 105]. Uma análise clássica utilizando o ponto médio dos intervalos perderia a informação sobre a variação de glicose para cada paciente.

Em ADS o conhecimento extraído a partir dos conjuntos de dados, é representado por dados mais complexos. Os dados são descritos por variáveis multivaloradas que podem não

somente assumir um valor numérico ou categórico, mas um conjunto de categorias, intervalos ou distribuições de pesos. Por isso, se faz necessário o desenvolvimento e a adaptação de métodos de dados clássicos para dados simbólicos. Vários métodos já foram adaptados para a análise de dados simbólicos, como, k-vizinhos mais próximos (DOUX; LAURENT; NADA, 1997), *Support Vector Regression* (CARRIZOSA; GORDILLO, 2007), *MultiLayer Perceptrons* (ROSSI; CONAN-GUEZ, 2008), *Learning Vector Quantization* (FILHO, 2013) entre outros.

1.1 Motivação

É crescente a busca pela automatização da obtenção de conhecimento em diversos domínios, como reconhecimento de padrões, aprendizado de máquina, mineração de dados. Essas áreas costumam fazer uso de métodos de agrupamento. Um método de agrupamento consiste em alocar elementos similares em grupos, de forma que os elementos de um grupo tenham um alto grau de similaridade entre si e um alto grau de dissimilaridade com os elementos dos demais grupos.

Os métodos de agrupamento podem ser classificados em rígido (*hard*) e difuso (*fuzzy*). Um método de agrupamento rígido associa cada observação do conjunto de dados a uma única classe, enquanto que no agrupamento difuso cada observação está associada com todos os grupos da partição, ou seja, as observações não pertencem a um único grupo e, por esse motivo, se calcula os graus de pertinência. Um método difuso assume que os graus de pertinência devem ter valores entre 0 e 1 e a soma de todos os graus de pertinência de uma observação relativos aos grupos deve ser igual a 1 (THEODORIDIS; KOUTROUMBAS, 2006). Outro tipo de método de agrupamento é o método probabilístico. A abordagem probabilística considera o grau de pertinência de uma observação como sendo a probabilidade dessa observação pertencer a um determinado grupo.

Um dos grandes desafios em agrupamento está em realizar a separação dos grupos quando os dados estão distribuídos de maneira arbitrária. De uma modo geral, quando isso ocorre, costuma-se dizer que os dados são não linearmente separáveis. No trabalho de GIROLAMI (2001) foi desenvolvido um algoritmo capaz de produzir separações não lineares entre grupos, transformando o espaço de entradas em um espaço de alta dimensão e então executando o agrupamento neste novo espaço. Este mapeamento para o novo espaço de alta dimensão é realizado através de funções de núcleo (*kernel functions*) (CRISTIANINI; SHAEW-TAYLOR, 2000), cujo objetivo é aumentar o poder computacional de métodos lineares. Projeta-se os dados em um espaço de alta dimensão e encontra-se um hiperplano que separe este espaço linearmente, alocando cada indivíduo a um determinado grupo. Esta técnica é conhecida como Mercer *kernel trick*, uma alternativa muito utilizada em problemas de caráter não linear.

1.2 Justificativa

A presença de métodos de agrupamento probabilísticos na Análise de Dados Simbólicos é bastante escassa. O algoritmo *Expectation Maximization* para dados simbólicos intervalares introduzido por DOMINGUES (2010) embora funcione na resolução de problemas linearmente separáveis, ou seja, problemas em que indivíduos pertencentes a um determinado grupo podem ser separados de indivíduos pertencentes a outros grupos por um hiperplano, ele falha no domínio de problemas não linearmente separáveis. Por isso, se torna necessário o desenvolvimento de um método probabilístico de agrupamento não linear.

1.3 Objetivos

Este trabalho busca estender o método de agrupamento clássico probabilístico *Kernel Gaussian Mixture Model* (KGMM) (WANG; LEE; ZHANG, 2003), para o contexto de Análise de Dados Simbólicos fazendo uso de funções de núcleo para dados simbólicos intervalares definidas em COSTA; PIMENTEL; SOUZA (2013). Os objetivos específicos do trabalho são:

1. Estender o KGMM para dados simbólicos intervalares;
2. Utilizar o algoritmo como técnica de agrupamento;
3. Realizar simulações para verificar o desempenho do método proposto;
4. Aplicar a técnica proposta em bases de dados reais;
5. Comparar os resultados da técnica proposta com algoritmos existentes na literatura.

1.4 Estrutura do trabalho

Essa dissertação está organizada da seguinte forma: O Capítulo 2 mostra uma introdução aos dados simbólicos e algoritmos da literatura desenvolvidos para intervalos. O Capítulo 3 apresenta o IKEM, o método probabilístico de agrupamento de dados simbólicos intervalares não linear desenvolvido neste trabalho. O Capítulo 4 traz o estudo de desempenho do método proposto, comparando-o a outros métodos da literatura através de conjuntos de dados simbólicos intervalares sintéticos e reais. O Capítulo 5traz as considerações finais e trabalhos futuros.

2

Fundamentação Teórica

Neste capítulo é realizada uma revisão sobre dados simbólicos e métodos de agrupamento para dados simbólicos intervalares.

2.1 Dados Simbólicos

A análise de dados simbólicos (ADS) é uma extensão da análise de dados padrão onde tabelas de dados simbólicos são utilizados como entrada e objetos simbólicos são emitidos como resultado. As unidades de dados são chamados simbólicos, uma vez que são mais complexos do que os normais, pois não contêm apenas valores ou categorias, mas também incluem uma variação de estrutura (DIDAY, 2003). Um dos objetivos da ADS é prover técnicas para a redução de grandes bases de dados em bases de dados simbólicos e posterior análise exploratória dos dados através do emprego de técnicas de mineração de dados simbólicos já desenvolvidos. A representação de conhecimento através dos dados simbólicos permite a atribuição de múltiplos valores e regras para cada variável. Essas novas variáveis (conjuntos, intervalos e histogramas) tornam possível reter informações sobre a variabilidade intrínseca ou incerteza do conjunto de dados original (COSTA, 2011).

Os dados simbólicos foram propostos com o objetivo de introduzir uma descrição mais ampla das observações normalmente armazenadas usando valores pontuais, quantitativos e categóricos. Considere os cenários descritos abaixo:

- Considere que temos as seguintes variáveis de interesse: $v_1 = \{Cor\}, v_2 = \{Peso\}, v_3 = \{Partenocarpia\}$ e a população de interesse $\Omega = \{Espécies\ de\ Pupunhas\}$. A pupunha (*Bactris gasipaes*) é uma fruta muito saborosa e rica em nutrientes que ocorre em toda a Amazônia e faz parte da dieta dos nativos da região. O peso do fruto maduro da pupunha pode variar entre 1,0 e 97,5 gramas, pode apresentar coloração amarela, vermelha ou verde, e pode ocorrer ausência de semente no fruto (partenocarpia). Uma dada espécie $i \in \Omega$ sem semente, $v_3(i) = \{S\}$, tem peso menor do que 53 gramas, $v_2(i) = [1,0 : 53]$ e ocorre predominantemente na cor verde, $v_1(i) = \{Verde\}$. Não é tarefa simples inserir esse tipo de informação/regra em uma base de dados tradicional.

Este tipo de pupunha é melhor representada pela descrição:
$[Peso = [1,0 : 97,5]], [Cor = \{verde, amarelo, vermelho\}]$,
$[Partenocarpia = \{SIM(S), NÃO(N)\}]$ e $[se\{Partenocarpia = S\}]$ então
$\{Peso < 53 \; e \; Cor = verde\}$
Esta é uma descrição associada ao conceito de variável aleatória simbólica da pupunha descrita em (DOMINGUES, 2010). $(Peso, Cor, Partenocarpia) = ([1,0 : 97,5], \{amarelo, vermelho, verde\}, \{SIM, NÃO\})$.

- Suponha que uma empresa possua um banco de dados com os dados referentes às características de utilização de uma plataforma de jogos digitais de uma grande editora de livros. Nesta base são armazenadas informações relativas a todas as requisições feitas pelo usuário, assim como o tempo de sessão de cada jogo, pontuação, etc. Com o contínuo uso da plataforma a previsão é que a quantidade de informações cresça consideravelmente. Os dados simbólicos apresentam técnicas consistentes para sumarizar grandes bases de dados clássicos em novos conjuntos de dados simbólicos com tamanho reduzido, facilitando o gerenciamento e, em alguns casos, sem nenhuma perda de informação.

- Quando existe a necessidade de divulgação de informações de caráter sigiloso como faixas salariais, valores em investimentos de risco, percentuais de acidentes de trabalho. Podemos expressar esse tipo de dados através de dados simbólicos usando intervalos, distribuições de frequências ou distribuições de probabilidade.

Em determinadas situações, os dados intervalares representam melhor os valores de certas variáveis. Por exemplo, podemos utilizar intervalos para expressar a variação de temperatura durante um dia, ou intervalos em geral para um grupo de indivíduos. Os intervalos também podem ser usados quando se mede várias vezes uma mesma variável para um determinado indivíduo ao longo do tempo, e esta informação precisa ser resumida. Por exemplo, um paciente tem sua taxa de glicose no sangue acompanhada pelo seu médico ao longo de um mês, no final do mês os limites inferior e superior respectivamente da sua taxa de glicose oscilava no intervalo [70, 110]. Outros exemplos de dados de intervalo aparecem no caso de dados imprecisos, ou quando um certo parâmetro é estimado por um intervalo de confiança, e, em geral, sempre que a incerteza e imprecisão surge em um determinado problema (CARRIZOSA; GORDILLO, 2007). A representação de dados simbólicos do tipo intervalo será o objeto de estudo deste trabalho. A descrição dos outros tipos de dados simbólicos pode ser encontrada na literatura pertinente (BILLARD; DIDAY, 2007).

A Tabela 2.1 (BILLARD; DIDAY, 2007), contendo dados clássicos, será utilizada para ilustrar os conceitos necessários para a definição de dado simbólico intervalar, gerado a partir da aplicação do processo de generalização de bases clássicas. Essa tabela é representada por uma matriz $\mathbf{Y}$, $n \times p$ sendo composta por n registros médicos de indivíduos de uma típica companhia

Tabela 2.1: Tabela com dados clássicos

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}
1	Boston	M	24	M	S	2	0	74,9	68	120	79	183
2	Boston	M	56	M	C	1	2	84,4	84	130	90	164
3	Chicago	D	48	M	C	1	2	79,5	73	126	82	229
4	El Paso	M	47	F	C	0	1	64,0	78	121	86	239
5	Byron	D	79	F	C	0	4	69,0	84	150	88	187
6	Concord	M	12	M	S	2	0	33,1	69	126	85	109
7	Atlanta	M	67	F	C	1	0	75,4	81	134	89	190
8	Boston	O	73	F	C	0	4	75,0	77	121	81	181
9	Lindfield	D	29	M	C	2	2	103,1	62	124	81	214
10	Lindfield	D	44	M	C	1	3	98,1	71	125	79	218
11	Boston	D	54	M	S	1	0	96,7	57	118	88	189
12	Chicago	M	12	F	S	2	0	34,1	69	115	81	153
13	Macon	M	73	F	C	0	1	69,0	58	123	82	188
14	Boston	D	48	M	C	0	4	93,5	73	113	72	264
15	Peoria	O	79	F	C	0	3	69,5	72	106	78	118
16	Concord	D	20	M	S	2	1	121,7	79	123	80	205
17	Boston	D	20	F	S	2	0	71,3	75	116	87	180
18	Chicago	D	17	M	S	2	0	73,1	69	114	78	169
19	Stowe	D	31	M	C	1	2	83,1	81	118	84	185
20	Tara	M	83	M	C	0	1	58,1	80	108	80	224
21	Quincy	O	57	M	S	1	0	72,1	72	114	75	234
22	Atlanta	O	86	M	C	0	2	83,4	72	114	72	152

Tabela 2.2: Descrição dos dados da Tabela 2.1

	Descrição		Descrição
v_1	Domicílio	v_7	Número de filhos: ≥ 0
v_2	Tipo: Dental(D), Médico(M), Ótico(O)	v_8	Peso (em Kg): > 0
v_3	Idade (em Anos): ≥ 0	v_9	Pulso: > 0
v_4	Gênero: Masc. (M), Fem. (F)	v_{10}	Pressão sistólica: > 0
v_5	Estado civil: Solteiro (S), Casado (C)	v_{11}	Pressão diastólica: > 0
v_6	Número de pais vivos: 0, 1, 2	v_{12}	Colesterol Total: > 0

Tabela 2.3: Realizações das descrições das categorias com novas variáveis intervalares

	Tipo×Gênero	n_u	vi_1 Idade	vi_2 Est. Civil	vi_3 Peso	vi_4 Pulso	vi_5 Pais Vivos	vi_6 Colest.
w_1	Dental Masc.	8	[17:54]	{C,S}	[73,1:121,7]	[57:81]	{0,1,2}	[169:264]
w_2	Dental Fem.	2	[20:79]	{C,S}	[69,0:71,3]	[75:84]	{0,2}	[180:187]
w_3	Médico Masc.	4	[12:83]	{C,S}	[33,1:84,4]	[68:84]	{0,1,2}	[109:224]
w_4	Médico Fem.	4	[12:73]	{C,S}	[34,1:69,0]	[58:81]	{0,1,2}	[153:239]
w_5	Ótico Masc.	2	[57:86]	{C,S}	[72,1:83,4]	[72:72]	{0,1}	[152:234]
w_6	Ótico Fem.	2	[73:79]	{C}	[69,5:75,0]	[72:77]	{0}	[118:181]

de seguros, $n = 22$ o número de observações do conjunto de dados. A descrição dos campos da tabela é apresentada na Tabela 2.2. Para cada observação há um registro composto por informações familiares (estado civil, número de filhos, número de irmãos, etc.) e informações médicas (pressão, peso, colesterol, etc.), sendo $p = 12$ o número de variáveis aleatórias. Cada linha da Tabela 2.1 é composta por um conjunto de dados clássicos e representa uma realização para as variáveis aleatórias $(v_1, v_2, ..., v_p)$ para um determinado indivíduo. Para tabelas pequenas como essa, as técnicas estatísticas clássicas podem ser empregadas satisfatoriamente. Contudo, quando p e n são muito grandes, a análise pode tornar-se impraticável.

Os dados simbólicos podem ser extraídos a partir de bases de dados clássicas, como a apresentada na Tabela 2.1. Como exemplo, considere descrever as realizações da variável peso para a categoria "mulheres com seguro médico" (Tipo $\times$ Gênero). Aplicando essa regra à Tabela 2.1 resulta no vetor {34,1; 64,0; 69,0; 75,4}. Estes valores podem ser interpretados como realizações no intervalo [34,1:75,4]. A categoria "mulheres com seguro médico" ou (Tipo - v_2 $\times$ Gênero - v_4) é um exemplo de conceito simbólico. Como há 3 tipos (Dental (D), Médico(M) e Ótico(O)) e dois gêneros (Masc.(M) e Fem.(F)) na tabela, há portanto 6 possíveis categorias $(w_1, w_2, w_3, w_4, w_5, w_6)$, ou observações simbólicas, sendo cada observação um conjunto de indivíduos que satisfazem a descrição da categoria, gerando uma nova matriz $\mathbf{X}$. A tabela 2.3 apresenta um grupo de realizações simbólicas para as categorias (Tipo $\times$ Gênero).

Os dados simbólicos podem ser estruturados e podem registrar a variação dos valores. A Tabela 2.3 ilustra alguns exemplos do processo de extração de dados simbólicos a partir de bases clássicas (DIDAY, 2003).

Uma variável vi_j é do tipo intervalo se ela representa uma realização $x_{ij} = [a, b] \subset \Re$, com $a \leq b$ e a,b $\in \Re$ (conjunto dos números reais). Onde a ou x_{ij}^{inf} representa o limite inferior e b ou x_{ij}^{sup} representa o limite superior do intervalo. No exemplo da Tabela 2.3, os intervalos são gerados como resultado da agregação (generalização) de dados clássicos. Os valores dos intervalos de $\mathbf{X}$ referentes à categoria w_i na variável vi_j são dados por:

$$x_{ij}^{inf} = \min_{u \in \beta_i} y_{uj},$$

$$x_{ij}^{sup} = \max_{u \in \beta_i} y_{uj},$$

em que β_i é o vetor dos u-ésimos valores $(u \in \beta)$ que compõem a categoria w_i. Exemplos dessa definição podem ser obtidos do conjunto de dados simbólicos da Tabela 2.3. Considere a variável "Idade" para $i = 3$:

$vi_1(w_3) = Idade(w_3) = Idade(Médico\ x\ Masculino) = x_{3,1} = [12 : 83]$,

cujo resultado é um intervalo que cobre as idades dos homens com plano médico.

Merece destaque a variável "Pulso" da observação vi_5. Este exemplo ilustra o caso em que um intervalo representa um ponto clássico, cuja realização simbólica $x_{ij} = [a, a]$.

As variáveis simbólicas do tipo intervalo também podem serem empregadas quando

não é possível obter uma medida precisa das observações, como no caso dos instrumentos de medição. Em resumo, os dados simbólicos do tipo intervalo são bastante úteis como ferramenta para reduzir o tamanho de bases de dados através do processo de agregação dos dados, porém, essa representação simbólica tem a desvantagem de não apresentar as descrições das distribuições de frequências das observações originais.

2.2 Métodos de agrupamento para dados simbólicos intervalares

Um método de agrupamento consiste em alocar elementos similares em grupos, de forma que os elementos de um grupo tenham um alto grau de similaridade entre si e um alto grau de dissimilaridade com os elementos dos demais grupos. O processo de separar ou identificar grupos a partir de semelhanças entre objetos é comumente realizada por seres humanos e envolve um processo de aprendizagem. Por exemplo, uma criança pode associar as palavras cão e gato através da observação da diferença que existe entre estes dois animais. A análise de agrupamentos é normalmente considerada uma subárea de estudo de Reconhecimento de Padrões e Inteligência Artificial.

Diferentemente dos algoritmos de classificação, nos quais os objetos precisam ser rotulados por classe caracterizando um aprendizado supervisionado, os algoritmos de agrupamento não necessitam dessa rotulação, isto é, ocorre a aprendizagem não-supervisionada. Na aprendizagem supervisionada, a separação das classes é realizada por um supervisor, que mede o grau de desempenho desse algoritmo e realiza ajustes sobre o mesmo até que seja atingida alguma medida que seja considerada satisfatória, utilizando como recurso alguma informação externa sobre o domínio avaliado. Entretanto, essa abordagem possui uma limitação: é necessário um conhecimento preliminar do domínio estudado, o que não é possível em diversas situações. Na aprendizagem não-supervisionada, como ocorre nos métodos de agrupamento, não há a necessidade de informações a priori sobre o domínio avaliado, levando-se em consideração, apenas, a disposição dos dados e suas propriedades internas (ALMEIDA PIMENTEL, 2013).

Como esses métodos são executados de forma não-supervisionada, existem diversas técnicas para a estimação do número ideal de conjuntos finais que devem ser criados de forma a tornar a divisão dos dados mais representativa para o problema estudado, como apresentado em (THEODORIDIS; KOUTROUMBAS, 2006). Os métodos de agrupamento podem ser classificados de acordo com a forma que os mesmos interpretam os dados e a maneira com que esses objetos se organizam em agrupamentos. Os pricipais métodos de particionamento propostos para dados simbólicos são descritos nas seções a seguir.

2.2.1 K-médias rígido

Abordagens tradicionais de agrupamento geram partições, onde em uma partição, cada indivíduo pertence a um e somente um grupo. Esse tipo de agrupamento é conhecido rígido.

Consistem em obter uma partição a partir de um determinado conjunto de n elementos agrupados em um número pré-definido de k classes, onde $k \leq n$, de forma que cada classe possua pelo menos um elemento e cada elemento deve pertencer unicamente a uma classe, isto é, não admitem a existência de grupos vazios e que estes não tenham elementos em comum.

O funcionamento do algoritmo K-médias rígido para dados intervalares simbólicos definido em CARVALHO (2007) é descrito a seguir. Supondo que se tem um conjunto de dados intervalares simbólicos $\mathbf{X}$ como uma matriz $n \times p$ da qual cada linha é representada por $x_i = (x_{i1}, ..., x_{ip})$, onde $x_{ij} = [a_{ij}, b_{ij}], (j = 1, ..., p) \in \Im = \{[a,b] : a,b \in \Re, a \leq b\}$. Considere que $x_i^{inf} = (a_{i1}, ..., a_{ip})^T$ e $x_i^{sup} = (b_{i1}, ..., b_{ip})^T$ são dois vetores p-dimensionais associados aos limites inferior e superior respectivamente dos intervalos que descrevem a i-ésima observação de $\mathbf{X}$. O método de agrupamento k-médias rígido para dados intervalares simbólicos visa fornecer uma partição rígida de um conjunto de observações simbólicas em G aglomerados $C_1, ..., C_G$ e um conjunto correspondente de protótipos $g_1, ..., g_c$ tal que o critério J seja minimizado. Este critério é definido como:

$$J = \sum_{k=1}^{c} \sum_{i=1}^{n} \left[(x_i^{inf} - g_k^{inf})^2 + (x_i^{sup} - g_k^{sup})^2 \right], \tag{2.1}$$

Os centróides podem ser calculados da seguinte forma:

$$g_k^{inf} = \frac{1}{n_k} \sum_{i=1}^{n} x_i^{inf}, \tag{2.2}$$

$$g_k^{sup} = \frac{1}{n_k} \sum_{i=1}^{n} x_i^{sup},$$

onde n_k é o número de observações do grupo k.

A cada iteração do algoritmo, dois principais passos são executados: representação e alocação. No primeiro, os representantes de cada grupo são recalculados de acordo com os elementos presentes nesse grupo buscando minimizar a função objetivo. Enquanto no segundo passo, objetos são atribuídos às classes cuja distância entre o objeto e o representante da classe seja mínima dentre todas as distâncias para as demais classes da partição. O processo é repetido até que não haja mais alterações nas classes, isto é, elementos não sejam realocados, ou até que a diferença entre o critério atual e o calculado na iteração imediatamente anterior seja considerada pequena.

2.2.2 C-médias difuso

Em muitas aplicações é desejável que a similaridade de um indivíduo seja compartilhada entre os grupos. Isso permitiria uma melhor descrição de situações em que alguns indivíduos podem pertencer a grupos sobrepostos, ou no caso de alguns indivíduos não pertecerem a nenhum grupo, uma vez que são valores discrepantes. Agrupamentos difusos permitem associar um

indivíduo com todos os grupos através de um parâmetro que representa o grau de pertinência do indivíduo ao grupo. No C-médias difuso os elementos mais afastados do centróide possuem um menor grau de pertinência, enquanto aqueles mais próximos ao centróide têm uma pertinência maior. O centróide é obtido fazendo-se uma média ponderada do grau de todos os indivíduos para aquele grupo. (COSTA, 2011)

O funcionamento do algoritmo C-médias difuso para dados intervalares simbólicos definido em CARVALHO (2007) é descrito a seguir. Supondo que se tem um conjunto de dados intervalares simbólicos $\mathbf{X}$ como uma matriz $n \times p$ da qual cada linha é representada por $x_i = (x_{i1}, ..., x_{ip})$, onde $x_{ij} = [a_{ij}, b_{ij}], (j = 1, ..., p) \in \Im = \{[a,b] : a,b \in \Re, a \leq b\}$. Considere que $x_i^{inf} = (a_{i1}, ..., a_{ip})^T$ e $x_i^{sup} = (b_{i1}, ..., b_{ip})^T$ são dois vetores p-dimensionais associados aos limites inferior e superior respectivamente dos intervalos que descrevem a i-ésima observação de $\mathbf{X}$. O método de agrupamento C-médias difuso para dados intervalares simbólicos visa fornecer uma partição difusa de um conjunto de observações simbólicas em G aglomerados $C_1, ..., C_G$ e um conjunto correspondente de protótipos $g_1, ..., g_c$ tal que o critério J que avalia o encaixe entre os clusters e os seus protótipos seja minimizado. Este critério é definido como:

$$J = \sum_{k=1}^{c} \sum_{i=1}^{n} (u_{ik})^m \left[(x_i^{inf} - g_k^{inf})^2 + (x_i^{sup} - g_k^{sup})^2 \right], \tag{2.3}$$

onde u_{ik} é o parâmetro que representa o grau de pertinência da i-ésima observação de $\mathbf{X}$ ao grupo C_k, $(k = 1, ..., G)$ $m \in (1, +\infty)$ é um parâmetro que indica uma ponderação referente a pertinência das observações.

O algoritmo define um grau de pertinência inicial de cada observação para cada grupo, e alterna entre o passo de representação e o passo de alocação até que o critério J alcance um valor estacionário, representando um mínimo local.

O passo de representação consiste em definir os melhores protótipos utilizando o grau de pertinência atual.

$$g_k^{inf} = \frac{\sum_{i=1}^{n} (u_{ik})^m \, x_i^{inf}}{\sum_{i=1}^{n} (u_{ik})^m}, \tag{2.4}$$

$$g_k^{sup} = \frac{\sum_{i=1}^{n} (u_{ik})^m \, x_i^{sup}}{\sum_{i=1}^{n} (u_{ik})^m},$$

O passo de alocação consiste em atualizar o grau de pertinência utilizando os novos

protótipos.

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{\left(x_i^{inf} - g_k^{inf}\right)^2 + \left(x_i^{sup} - g_k^{sup}\right)^2}{\left(x_i^{inf} - g_j^{inf}\right)^2 + \left(x_i^{sup} - g_j^{sup}\right)^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{2.5}$$

Algorithm 1 C-médias difuso

Passo 1. Determine os valores iniciais de c tal que $2 \leq c <$ n; m tal que $1 < m < \infty$; T (limite de iterações); $\varepsilon > 0$; Inicialize $u_{ik}(i = 1,...,n \; e \; k = 1,...,c)$ de cada observação i para cada grupo k tal que $u_{ik} \geq 0$ e $\sum_{k=1}^{c} u_{ik} = 1$;

Passo 2. $t = 0$. Para cada iteração $t = t + 1$:

Passo 3. (*Passo de representação*): Calcule os protótipos g_k utilizando a Equação 2.4;

Passo 4. (*Passo de alocação*): Atualize o grau de pertinência u_{ik} de cada observação i para cada grupo Ck utilizando a Equação 2.5;

Passo 5. Se $t > T$ ou $|J_{t+1} - J_t| \leq \varepsilon$, então pare;

2.2.3 *Expectation Maximization*

O algoritmo *Interval Expectation Maximization* (IEM) tem se tornado uma ferramenta muito popular em análise estatística para estimação de parâmetros por máxima verossimilhança na presença de dados incompletos (faltantes), podendo também ser usado como técnica de agrupamento (DOMINGUES, 2010).

O agrupamento feito pelo IEM consiste em utilizar um Modelo de Mistura de Gaussianas, que pode ser visto como um modelo generativo já que ele assume que um conjunto de dados X é gerado por G subconjuntos a partir de uma função de probabilidade consistindo de componentes Gaussianos. Desta forma, podemos inferir os parâmetros do Modelo de Mistura de Gaussianas (MMG) através do algoritmo IEM buscando identificar e agrupar regiões densas no conjunto de dados.

O IEM é um procedimento iterativo, dividido em dois passos, que consiste em estimar novos parâmetros em termo de parâmetros antigos. No primeiro passo, *Expectation* (E-*step*) é calculada a probabilidade de cada elemento pertencer a um determinado grupo G. No segundo passo, *Maximization* (M-*step*) os parâmetros são atualizados de acordo com a probabilidade calculada no passo anterior, de forma que a cada iteração a verossimilhança aumenta e o algoritmo converge em um ponto específico.

Sejam $(C_1,...,C_G)$ uma partição em G grupos, em que cada grupo está relacionado a uma gaussiana, e $\hat{\theta}_k = (\hat{\tau}_k, \; \hat{\mu}_k, \; \hat{\Sigma}_k)(k = 1,...,G)$ um vetor de parâmetros associados a $k-$ésima classe onde $\hat{\mu}_k = [\hat{\mu}_k^{inf}, \hat{\mu}_k^{sup}]$ é o vetor de médias correspondente aos limites inferiores e superiores dos intervalos, respectivamente, $|\hat{\Sigma}_k|$ é o determinante da matriz de variâncias e covariâncias e $\hat{\tau}_k$ é o coeficiente da mistura que representa a proporção do grupo k no conjunto de dados.

Considerando v como o número de variáveis, podemos expressar a função densidade de probabilidade Gaussiana do grupo k por:

$$\hat{p}(x_i|C_k) = \frac{exp\left(-\frac{1}{2}\,[A+B]\right)}{(2\pi)^{v/2}|\hat{\Sigma}_k|^{1/2}}, \tag{2.6}$$

$$A = (x_i^{inf} - \hat{\mu}_k^{inf})^T \hat{\Sigma}_k^{-1} (x_i^{inf} - \hat{\mu}_k^{inf}),$$

$$B = (x_i^{sup} - \hat{\mu}_k^{sup})^T \hat{\Sigma}_k^{-1} (x_i^{sup} - \hat{\mu}_k^{sup})$$

A probabilidade da uma observação i pertencer ao grupo C_k é definida como:

$$\hat{p}(C_k|x_i) = \frac{\hat{\tau}_k\ \hat{p}(x_i|C_k)}{\sum_{j=1}^{G} \hat{\tau}_k\ \hat{p}(x_i|C_j))} \tag{2.7}$$

Os valores iniciais para os parâmetros θ podem ser obtidos a partir de uma partição aleatória ou por obtenção de k médias aleatórias, seguido de uma etapa de alocação dos indivíduos às k classes de acordo com a distância mínima computada. O passo E consiste em calcular a probabilidade de cada indivíduo pertencer a cada grupo através da Equação 2.7. No passo M atualizamos o vetor de parâmetros $\hat{\theta}$ através das Equações 2.8, 2.9 e 2.10 maximizando a função de log-verossimilhança na próxima iteração do algoritmo.

$$\hat{\tau}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{p}(C_k|x_i) \tag{2.8}$$

$$\hat{\mu}_k^{inf} = \frac{\sum_{i=1}^{n} \hat{p}(C_k|x_i) x_i^{inf}}{\sum_{i=1}^{n} \hat{p}(C_k|x_i)}, \tag{2.9}$$

$$\hat{\mu}_k^{sup} = \frac{\sum_{i=1}^{n} \hat{p}(C_k|x_i) x_i^{sup}}{\sum_{i=1}^{n} \hat{p}(C_k|x_i)}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} \hat{p}(C_k|x_i)(W+V)}{\sum_{i=1}^{n} \hat{p}(C_k|x_i)}, \tag{2.10}$$

$$W = (x_i^{inf} - \hat{\mu}_k^{inf})(x_i^{inf} - \hat{\mu}_k^{inf})^T,$$

$$V = (x_i^{sup} - \hat{\mu}_k^{sup})(x_i^{sup} - \hat{\mu}_k^{sup})^T$$

Os passos E e M do algoritmo são repetidos até que $||\theta^{t+1} - \theta^t||$ seja suficientemente pequeno.

Algorithm 2 Expectation Maximization

Passo 1. Determine os valores iniciais do vetor $\hat{\theta}(\hat{\tau}_k^0, \hat{\mu}_k^0, \hat{\Sigma}_k^0,)$ para cada grupo C_k;

Passo 2. $t = 0$. Para cada iteração $t = t + 1$:

Passo 3. (*E-Step*): Encontre a probabilidade de x_i pertencer a cada grupo $k(k = 1, ..., G)$ usando a Equação 2.7;

Passo 4. (*M-Step*): Atualize o vetor de parâmetros $\hat{\theta}_k$ de acordo com as Equações 2.8, 2.9 e 2.10;

Passo 5. Se $t > t_{max}$ ou $\sum_{k=1}^{G}\sum_{i=1}^{n}(\hat{p}(C_k|x_i)^t - \hat{p}(C_k|x_i)^{t-1})^2 < \varepsilon$, então pare;

3

Agrupamento não linear para dados simbólicos intervalares

Existem aplicações do mundo real que precisam de representações mais complexas que funções lineares. As funções de núcleo (*kernel functions*) são ferramentas poderosas para resolução desses problemas não lineares. Diversos algoritmos de aprendizado de máquina fazem uso de funções de núcleo, por exemplo, máquina de vetores de suporte, análise de componentes principais, redes neurais, entre outros.

As funções de núcleo projetam os dados em um espaço de alta dimensão para aumentar o poder computacional dos métodos lineares (CRISTIANINI; SHAEW-TAYLOR, 2000). O objetivo é encontrar um hiperplano que consiga separar o espaço linearmente nesse espaço de alta dimensão. Nos métodos de particionamento utilizando funções de núcleo, o produto interno entre as variáveis é substituído por uma função apropriada.

Neste capítulo é proposto o IKEM, um método de agrupamento para dados simbólicos intervalares utilizando funções de núcleo no espaço de características. Inicialmente, a seção 3.1 descreve as funções de núcleo para intervalo. A seção 3.2 explica o funcionamento do *EM* utilizando funções de núcleo no espaço de características.

3.1 Funções de núcleo para Intervalos

Considerando $\mathbf{X}$ uma matriz $n \times p$, a representação de intervalos no conjunto dos números reais $\Re$ é denotado pelo par ordenado de números reais $x_{ij} = [a,b]$, tal que $a \leq b$, a e $b \in \Re$, $(i = 1,...,n),(j = 1,...,p)$ (TAKAHASHI, 2012). Considere que $x_i^{inf} = (a_i^1,...,a_i^p)^T$ e $x_i^{sup} = (b_i^1,...,b_i^p)^T$ são dois vetores p-dimensionais associados aos limites inferior e superior respectivamente dos intervalos que descrevem a i-ésima observação de $\mathbf{X}$. Sendo $\phi : x_i \rightarrow \phi(x_i)$ uma função não linear que realiza um mapeamento do espaço de entrada para intervalos finitos $\mathbf{X}$ para um espaço de características de alta dimensão F (COSTA; PIMENTEL; SOUZA, 2013).

É possível construir novas funções de núcleo baseadas em outras funções de núcleo, apenas respeitando o Teorema de Mercer (CRISTIANINI; SHAEW-TAYLOR, 2000). Utilizando

a propriedade de que a soma de duas funções de núcleo sob um mesmo espaço de entrada representa uma outra função de núcleo válida, COSTA; PIMENTEL; SOUZA (2013) propõe que a função para dados intervalares definida por duas outras funções de núcleo expressa por:

$$IK(x_i,x_j) = \phi(x_i^{inf}).\phi(x_j^{inf}) + \phi(x_i^{sup}).\phi(x_j^{sup}), \tag{3.1}$$

é uma função de núcleo para todo x_i, $x_j \in \mathbf{X}$.

Seguindo a definição de funções de núcleo para dados clássicos (CRISTIANINI; SHAEW-TAYLOR, 2000), um função de núcleo para dados intervalares pode ser definida como:

$$K(x_i,x_j) = \phi(x_i).\phi(x_j), \tag{3.2}$$

para todo $x_i, x_j \in \mathbf{X}$.

Sendo $K1$ e $K2$ duas funções de núcleo definidas no espaço de entrada para intervalos, podemos afirmar que:

$$K(x_i,x_j) = K_1(x_i,x_j) + K_2(x_i,x_j), \tag{3.3}$$

para todo x_i, $x_j \in \mathbf{X}$.

Considerando que $\delta 1$ e $\delta 2$ são duas funções monótonas definidas sobre o espaço de entrada intervalar $\mathbf{X}$ tal que $\delta 1 : x_i \to x_i^{inf}$ e $\delta 2 : x_i \to x_i^{sup}$, podemos afirmar que:

$$K_1(x_i,x_j) = \phi(\delta 1(x_i)).\phi(\delta 1(x_j)) = \phi(x_i^{inf}).\phi(x_j^{inf})$$

$$K_2(x_i,x_j) = \phi(\delta 2(x_i)).\phi(\delta 2(x_j)) = \phi(x_i^{sup}).\phi(x_j^{sup})$$

Substituindo $K_1(x_i,x_j)$ e $K_2(x_i,x_j)$ em 3.3 temos que:

$$IK(x_i,x_j) = \phi(x_i^{inf}).\phi(x_j^{inf}) + \phi(x_i^{sup}).\phi(x_j^{sup}),$$

concluindo que 3.1 é uma função de núcleo para dados intervalares.

É importante ressaltar que a escolha da função de núcleo adequada é um fator importante para que o problema não linear no espaço de entrada se torne linear no espaço de características. COSTA; PIMENTEL; SOUZA (2013) definem duas funções de núcleo para intervalos da seguinte maneira:

- *Interval Radial Basis Function* (IRBF) definida como

$$IRBF(x_i,x_j) = exp\left(\frac{-\psi^2(x_i^{inf},x_j^{inf})}{2\sigma^2}\right) + exp\left(\frac{-\psi^2(x_i^{sup},x_j^{sup})}{2\sigma^2}\right), \tag{3.4}$$

onde $\psi^2(x_i^{inf},x_j^{inf}) = \sum_{f=1}^{p}(a_i^f - a_j^f)^2$ e $\psi^2(x_i^{sup},x_j^{sup}) = \sum_{f=1}^{p}(b_i^f - b_j^f)^2$ e σ é um parâmetro ajustável para determinar a largura da função de núcleo. Neste trabalho, o valor de σ é definido como 1.

- *Interval Polynomial Function* (IPF) definida como

$$IPF(x_i, x_j) = (x_i^{inf}.x_j^{inf} + c)^d + (x_i^{sup}.x_j^{sup} + c)^d, \tag{3.5}$$

onde d é o grau do polinômio e c é uma constante opcional.

3.2 *Interval Kernel Expectation Maximization*

O KGMM desenvolvido em WANG; LEE; ZHANG (2003) pode ser visto como a combinação do Modelo de Mistura de Gaussianas com funções de núcleo, utilizando o algoritmo *Expectation Maximization* (EM) para atingir a convergência. O *Interval Kernel Expectation Maximization* (IKEM) é o método desenvolvido neste trabalho buscando estender o KGMM para dados simbólicos intervalares. Para as n observações de treinamento $\{x_1, x_2, ..., x_n\}$ os pontos correspondentes no espaço de características são $\{\phi(x_1), \phi(x_2), ..., \phi(x_N)\}$. Podemos reescrever as equações do EM tradicional da seguinte forma:

E-step:

$$\hat{p}(\phi(x_i)|C_k) = \frac{1}{(2\pi)^{p/2}|\hat{\Sigma}_k|^{1/2}} exp\left(-\frac{1}{2}(\phi(x_i) - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1}(\phi(x_i) - \hat{\mu}_k)\right), \tag{3.6}$$

$$\hat{p}(C_k|\phi(x_i)) = \frac{\hat{\tau}_k\ \hat{p}(\phi(x_i)|C_k)}{\sum_{j=1}^{G} \hat{\tau}_k\ \hat{p}(\phi(x_i)|C_j))}, \tag{3.7}$$

M-step:

$$\hat{\tau}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{p}(C_k|\phi(x_i)) \tag{3.8}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{p}(C_k|\phi(x_i))\phi(x_i)}{\sum_{i=1}^{n} \hat{p}(C_k|\phi(x_i))}, \tag{3.9}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} \hat{p}(C_k|\phi(x_i))(\phi(x_i) - \hat{\mu}_k)(\phi(x_i) - \hat{\mu}_k)^T)}{\sum_{i=1}^{n} \hat{p}(C_k|\phi(x_i))} \tag{3.10}$$

Contudo, calcular diretamente os parâmetros das equações 3.9 e 3.10 é computacionalmente caro e algumas vezes impraticáveis. Podemos utilizar o Mercer *kernel trick* para contornar esta dificuldade e estimar os parâmetros no espaço de características.

3.2.1 Estimação de densidade no autoespaço

Podemos aproximar a função densidade de probabilidade Gaussiana dos dados de alta dimensões do IKEM descrita na Equação 3.6 empregando uma técnica proposta por WANG; LEE; ZHANG (2003) que utiliza decomposição em valores singulares que está descrita a seguir.

Primeiramente uma matriz $n \times n$ é definida como:

$$K_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{3.11}$$

onde $\langle \phi(x_i), \phi(x_j) \rangle$ representa o produto interno entre duas observações projetadas no espaço de características de alta dimensão. No contexto de dados simbólicos essa matriz *Kernel* inicial é preenchida utilizando uma função *Kernel* para intervalo.

A matriz *Kernel* centrada é calculada da seguinte forma:

$$\bar{K} = K - E_n K - K E_n + E_n K E_n \tag{3.12}$$

onde cada entrada da matriz $n \times n$, E_n é $\frac{1}{n}$. Desta forma o espaço original pode ser mapeado em um espaço de características de alta dimensão, como ilustrado na Figura 3.1.

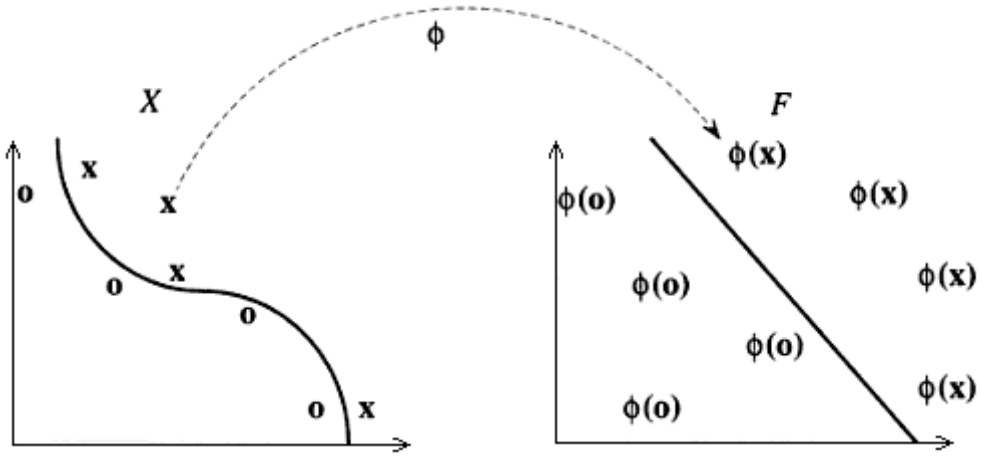

Figura 3.1: Um mapeamento em alta dimensão pode simplificar a tarefa de agrupamento (CRISTIANINI; SHAEW-TAYLOR, 2000)

Além disso, calculamos a matriz *Kernel* ponderada e a matriz *Kernel* ponderada projetada para o k-ésimo componente Gaussiano, respectivamente, utilizando as equações definidas em (YU, 2011):

$$(\tilde{K}_k)_{ij} = \omega_{k,i} \omega_{k,j} \bar{K}_{ij} \tag{3.13}$$

$$(\tilde{K}_k)'_{ij} = \omega_{k,j} \bar{K}_{ij} \tag{3.14}$$

onde

$$\omega_{k,i} = \left\{ \frac{p(C_k|\phi(x_i))}{\sum_{i=1}^{n} p(C_k|\phi(x_i))} \right\}^{1/2} \tag{3.15}$$

De acordo com MOGHADDAM; PENTLAND (1997) podemos aproximar a função

densidade de probabilidade Gaussiana 3.6 por:

$$\hat{p}(\phi(x_i)|C_k) = \frac{1}{(2\pi)^{\tilde{D}/2}\prod_{s=1}^{\tilde{D}}\{\lambda_{ks}\}^{1/2}} exp\left[-\frac{1}{2}\sum_{s=1}^{\tilde{D}}\frac{y_s^2}{\lambda_{ks}}\right] \tag{3.16}$$

onde $\tilde{D}$ é um parâmetro que denota o número de autovalores escolhidos, λ_{ks} é o s-ésimo autovalor da matriz *kernel* ponderada $\tilde{K}_k$ e y_s pode ser calculado da seguinte forma:

$$y_s = \beta_{ks}^T \Gamma_i, \tag{3.17}$$

sendo β_{ks} um vetor com os $\tilde{D}$ primeiros valores do autovetor da matriz *Kernel* ponderada $\tilde{K}_k$ e Γ_i representa a i-ésima coluna da matriz *Kernel* ponderada projetada $\tilde{K}_k'$. Notamos aqui que a matriz *Kernel* ponderada é usada para obter os autovalores λ_k e os autovetores β_k, enquanto que a matriz *Kernel* ponderada projetada é utilizada para calcular o y_s da equação 3.17.

Tendo em mãos a aproximação da função densidade de probabilidade Gaussiana 3.16, podemos calcular a probabilidade posteriori de acordo com a equação 3.7.

3.2.2 Inicialização e Convergência

Como todo método baseado em protótipos, o IKEM sofre do mal da inicialização, pois cada inicialização aleatória pode levar a resultados bastante distintos. Sendo assim, a etapa de inicialização precisa ser bem escolhida, assim como o método deve ser reiniciado aleatoriamente diversas vezes.

A etapa de inicialização do IKEM é baseada no IEM desenvolvido em (DOMINGUES, 2010). Escolha randomicamente duas observações g_1 e g_2 pertencentes ao conjunto de dados e atribua cada uma das outras observações a um grupo baseado na distância entre as observações e as outras duas escolhidas. O trabalho de (DOMINGUES, 2010) enfatiza que a distância de Hausdorff obteve resultados superiores a outras distâncias quanto ao número de iterações do algoritmo. Sendo x_i e x_j duas observações simbólicas intervalares, a distância normalizada de Hausdorff entre essas observações de p dimensões se dá por:

$$d(x_i,x_j) = \left\{\sum_{d=1}^{p}\left[\frac{Max[\,|x_{id}^{inf} - x_{jd}^{inf}|, |x_{id}^{sup} - x_{jd}^{sup}|\,]}{H_d}\right]^2\right\}^{1/2}, \tag{3.18}$$

sendo

$$H_d^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[Max[\,|x_{id}^{inf} - x_{jd}^{inf}|, |x_{id}^{sup} - x_{jd}^{sup}|\,]\right]^2.$$

Após atribuir cada observação a um grupo, se calcula a média e variância de cada grupo, utilizando-os para o cálculo da probabilidade a posteriori inicial, $\hat{p}(C_k|(x_i))$.

De acordo com COUVREUR (1997), uma das propriedades de convergência de algoritmo

baseado no EM é a função monotonicamente crescente da verossimilhança. Ao rodar o método diversas vezes podemos escolher os melhores resultados dos agrupamentos em função da máxima verossimilhança:

$$L = \sum_{i=1}^{n} log(max(\hat{p}(\phi(x_i)|C_k))) \tag{3.19}$$

Algorithm 3 Interval Kernel Expectation Maximization

Passo 1. Inicializar as probabilidades posteriori $\hat{p}(C_k|(x_i))(k = 1,...,G; i = 1,...,n)$;

Passo 2. Calcular as matrizes *kernel* K e $\bar{K}$;

Passo 3. $t = 0$. Para cada iteração $t = t + 1$:

Passo 4. Calcule $\omega_{k,i}, \tilde{K}_k$ e $\tilde{K}'_k$ de acordo com as equações 3.13 - 3.15 e $\hat{\tau}_k$ de acordo com a equação 3.8;

Passo 5. Faça a decomposição em valores singulares da matriz $\tilde{K}_k$ para obter os autovalores λ_k e autovetores β_k;

Passo 6. Estime a função densidade de probabilidade Gaussiana $\hat{p}(\phi(x_i)|C_k)$ através da equação 3.16;

Passo 7. Calcule a probabilidade a posteriori $\hat{p}(C_k|\phi(x_i))$ de acordo com a equação 3.7;

Passo 8. Se $t > t_{max}$ ou $\sum_{k=1}^{G}\sum_{i=1}^{n}(\hat{p}(C_k|\phi(x_i))^t - \hat{p}(C_k|\phi(x_i))^{t-1})^2 < \varepsilon$, então pare;

4

Apresentação e análise dos resultados

A seção 4.1 descreve o conceito básico de matriz de confusão, sendo suficiente para entender o conceito da Área da Curva ROC descrita na seção 4.2, que é utilizada para avaliação dos métodos de agrupamento. A seção 4.3 apresenta os conjuntos de dados utilizados na avaliação, os resultados e discussões sobre os experimentos realizados.

4.1 Matriz de confusão

Quando fazemos uma previsão, existem duas decisões corretas e duas erradas, gerando quatro tipos de resultados:

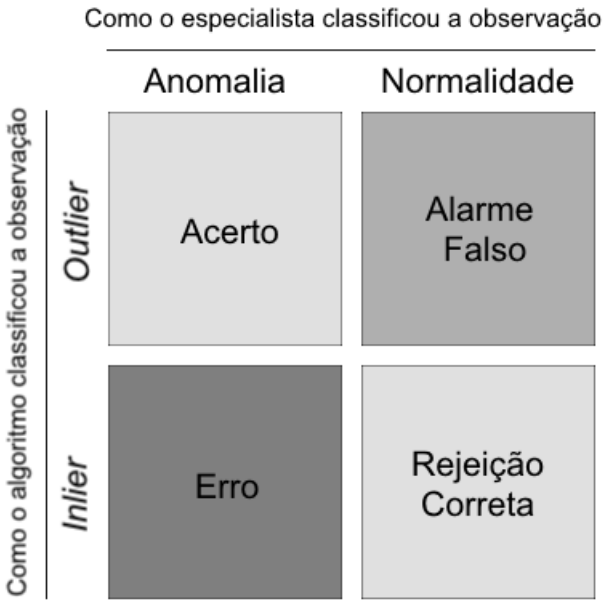

Figura 4.1: Matriz de confusão (JANSSENS, 2013)

A Figura 4.1 mostra uma matriz de confusão com os quatro resultados possíveis quando se compara a saída de um algoritmo com o rótulo do especialista do domínio. Por exemplo, ao rodar um antivírus em um computador podemos chegar aos seguintes resultados:

- Acerto: Um antivírus detectou um arquivo ofensivo como um vírus.
- Rejeição correta: Um antivírus não considerou um arquivo inofensivo como um vírus.
- Erro Tipo 1 ou Falso Positivo ou Alarme Falso: Um antivírus detectou um arquivo inofensivo como vírus.
- Erro Tipo 2 ou Falso Negativo ou Erro: Um antivírus não considerou um arquivo ofensivo como vírus.

O objetivo da matriz de confusão, ou matriz de correspondência, é permitir uma análise mais detalhada de uma classificação (FAWCETT, 2006).

4.2 Área da Curva ROC

A área da curva ROC,*Receiver Operating Characteristic* (ROC), é gerada a partir da combinação da taxa de falsos positivos com a taxa de acertos em uma única métrica. Sendo a sensibilidade a fração de acertos e 1-Especificidade a fração de falsos positivos como mostra a Figura 4.2.

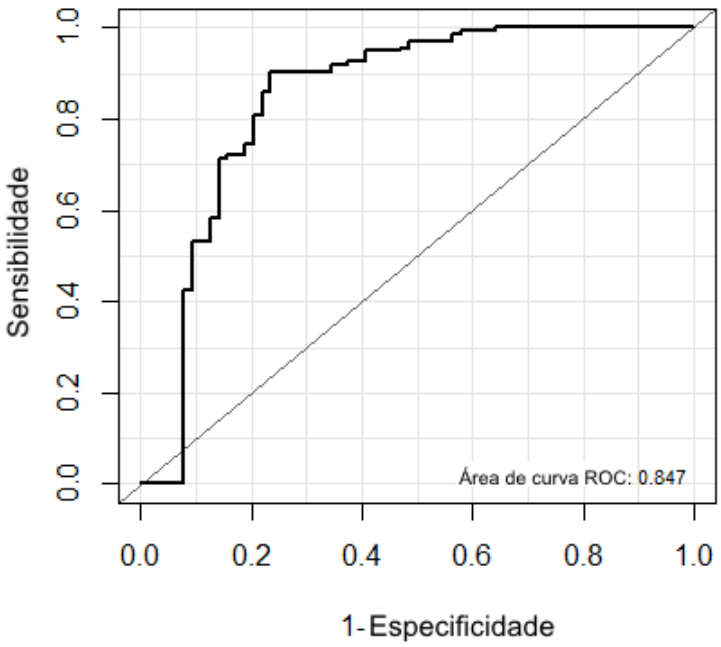

Figura 4.2: Curva ROC.

A área sob a curva ROC consiste em uma métrica de desempenho para avaliação de modelos, que permite estudar a variação para as medidas de sensibilidade e especificidade do modelo, para diferentes valores de ponto de corte. Essa curva confirma a boa capacidade de previsão do modelo, já que quanto maior a área entre a curva ROC e a diagonal principal, melhor

o desempenho do modelo (GÖNEN, 2007). A área da curva ROC é um valor sempre entre 0 e 1. A linha na diagonal representa a curva ROC de um preditor aleatório, que tem 0.5 de valor.

4.3 Experimentos e resultados

Esta seção apresenta os experimentos realizados juntamente com os resultados e discussões. Para cada conjunto de dados sintéticos, o índice de validação é estimado através de uma simulação Monte Carlo com 100 iterações. A finalidade da aplicação do método Monte Carlo é propiciar uma melhor avaliação quantitativa do desempenho dos métodos considerando situações com diferentes graus de dificuldades de agrupamento.

Para cada cenário, geramos 100 conjuntos de dados simbólicos intervalares provenientes da simulação Monte Carlo, baseados em uma distribuição gaussiana. Cada método é inicializado aleatoriamente 50 vezes, com configurações iniciais diferentes, e então é escolhido o melhor resultado para representar o método em cada conjunto de dados. Para os dados reais os métodos são rodados 100 vezes com 50 reinícios aleatórios.

O IKEM, método proposto no capítulo anterior e desenvolvido neste trabalho, é comparado com outros métodos da literatura. Os métodos IEM (DOMINGUES, 2010) e *Interval Kernel Fuzzy C-Means* (IKFCM) (COSTA, 2011), desenvolvidos para dados simbólicos intervalares, foram replicados e comparados com o método proposto.

No caso do IKFCM o critério J definido na Equação 2.3 serve como critério de validação, sendo escolhido o menor valor de J das 50 rodadas. Já para o IEM e IKEM, em geral se utiliza como critério a máxima verossimilhança da equação 3.19, sendo escolhido o maior valor das 50 rodadas. Testes de Friedman (FRIEDMAN, 1937) e Wilcoxon (WILCOXON, 1945) são aplicados aos valores obtidos da Área da Curva ROC, utilizando um nível de confiança de 5%.

4.3.1 Primeiro conjunto de dados intervalares sintéticos

O primeiro conjunto de dados sintético, mostrado nas Figuras 4.3 a 4.5, possui 200 retângulos distribuídos entre dois grupos com estruturas parecidas e tamanhos iguais: 100 retângulos para cada grupo gerados a partir de uma distribuição Gaussiana. Este conjunto foi desenvolvido apresentando características simples, para testar o funcionamento do algoritmo como método de agrupamento. Modificando a média, variância e a média da amplitude de cada grupo, conseguimos dificultar a convergência dos métodos, que acabam chegando em resultados diferentes.

Os grupos são representados por retângulos de cores diferentes, sendo preto para o grupo 1 e cinza para o grupo 2.

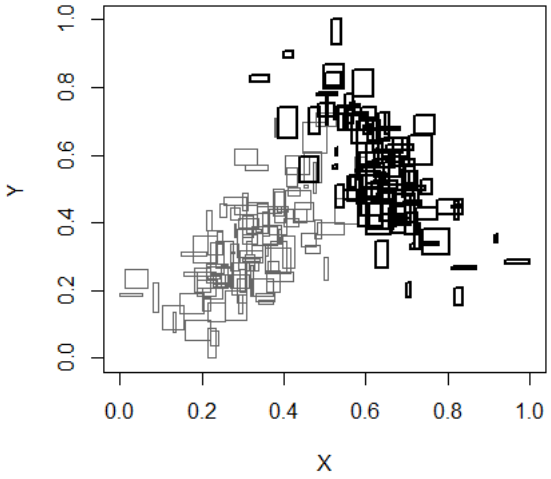

Figura 4.3: Cenário 1 do primeiro conjunto.

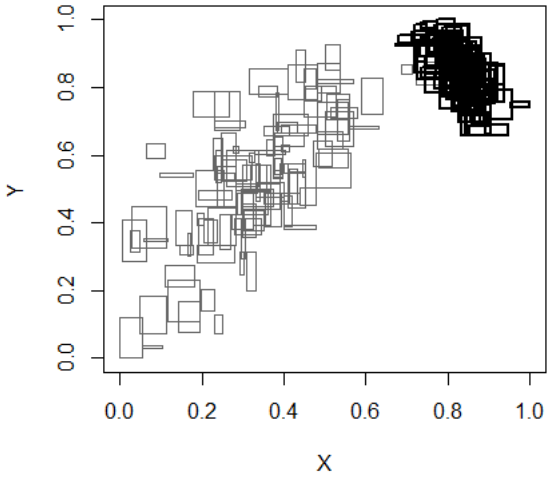

Figura 4.4: Cenário 2 do primeiro conjunto.

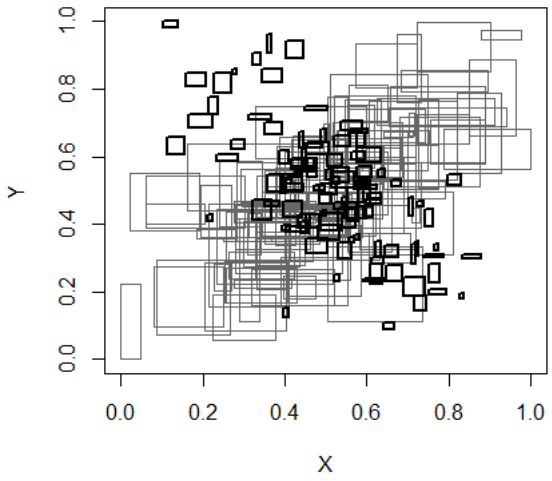

Figura 4.5: Cenário 3 do primeiro conjunto.

O cenário 1 apresenta um conjunto de dados representado por dois grupos linearmente separáveis com estruturas e dispersões semelhantes, que foi gerado para testar a funcionalidade dos experimentos. Os grupos possuem médias diferentes e matrizes de covariância semelhantes.

O cenário 2 é uma variação do primeiro cenário onde a dispersão de um grupo é maior que do outro grupo, foi desenvolvido para testar o comportamento das técnicas quando grupos tem matriz de covariância diferentes. Os grupos possuem médias e matrizes de covariância diferentes.

No cenário 3 os grupos foram sobrepostos, e a amplitude média dos intervalos de um grupo é maior que do outro. Aparentemente a dispersão dos grupos do terceiro cenário são

semelhantes, porém como a amplitude média dos intervalos de um grupo é maior do que a do outro, as matrizes de covariância estimadas acabam sendo diferentes. Os grupos possuem médias semelhantes e matrizes de covariância diferentes.

Tabela 4.1: Média e desvio padrão da área de curva ROC para o primeiro conjunto de dados sintéticos

Cenário	1. IEM	2. IKFCM	3. IKEM
1	0.9470 (0.0055)	0.8952 (0.0111)	0.8732 (0.0126)
2	0.9562 (0.0035)	0.9462 (0.0039)	0.8896 (0.0087)
3	0.6626 (0.0197)	0.5719 (0.0053)	0.7961 (0.0110)

A Tabela 4.1 mostra que nos cenários 1 e 2, todos os métodos possuem bons resultados em relação a média e desvio padrão, postos entre parênteses, da área da curva ROC, dando destaque ao método linear. Já no cenário 3, o método linear não obteve bons resultados, assim como o IKFCM. O interessante desse cenário é que o IKEM foi o único método que obteve um resultado bom.

A Tabela 4.2 contém os P-valores do teste estatístico de Friedman utilizados para comparar os resultados obtidos e mostrar se existe diferença significativa nos resultados dos métodos. A hipósese nula é de que os métodos possuem desempenhos médios semelhantes, a hipótese alternativa é de que os métodos possuem desempenhos diferentes. As médias μ_1, μ_2 e μ_3 correspondem aos métodos 1. IEM, 2. IKFCM e 3. IKEM respectivamente.

Tabela 4.2: P-valores do teste de Friedman da área da curva ROC no primeiro conjunto de dados sintéticos

Cenário	$H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$	Decisão
1	1.582e-06	Rejeita H_0
2	1.228e-08	Rejeita H_0
3	<2.2e-16	Rejeita H_0

O teste de Friedman mostra uma evidência muito forte contra a hipótese nula em todos os cenários, ou seja, os métodos não possuem um desempenho médio parecidos, por isso é necessário a aplicação de outro teste de hipótese dois a dois, para ratificar a eficiência do método proposto. A Tabela 4.3 mostra os resultados do teste de hipótese de Wilcoxon sobre a área de curva ROC dos métodos avaliados.

Tabela 4.3: P-valores do teste de Wilcoxon da área da curva ROC no primeiro conjunto de dados sintéticos

Cenário	$H_0 : \mu_3 = \mu_1$ $H_1 : \mu_3 > \mu_1$	Decisão	$H_0 : \mu_3 = \mu_2$ $H_1 : \mu_3 > \mu_2$	Decisão
1	1	Não Rejeita H_0	0.9635	Não Rejeita H_0
2	1	Não Rejeita H_0	1	Não Rejeita H_0
3	2.685e-10	Rejeita H_0	<2.2e-16	Rejeita H_0

O teste de Wilcoxon nos mostra que o método proposto teve resultados superiores aos métodos da literatura no Cenário 3. Já nos Cenários 1 e 2 os resultados obtidos pelo método não foram superiores. A direrença na amplitude média dos intervalos dos grupos foi suficiente para o método conseguir estimar as matrizes de covariância capaz de separar os grupos do Cenário 3, comprovando a superioridade do método proposto em problemas semelhantes a esse cenário.

4.3.2 Segundo conjunto de dados intervalares sintéticos

O segundo conjunto de dados sintético, mostrado nas Figuras 4.6 a 4.8 possui 180 retângulos distribuídos entre dois grupos: 100 retângulos no grupo centralizado e 80 retângulos no grupo mais disperso. Este conjunto foi desenvolvido para testar o funcionamento do método proposto em problemas onde os grupos são não linearmente separáveis.

Os grupos são representados por retângulos de cores diferentes, sendo preto para o grupo 1 e cinza para o grupo 2.

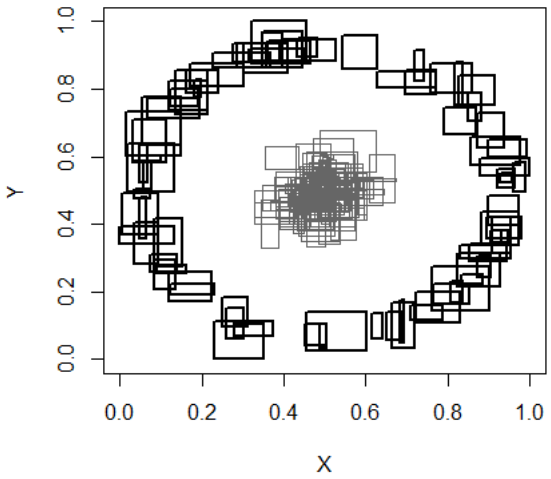

Figura 4.6: Cenário 1 do segundo conjunto.

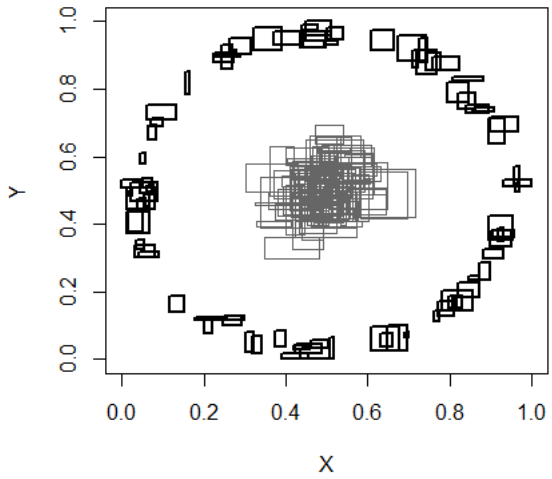

Figura 4.7: Cenário 2 do segundo conjunto.

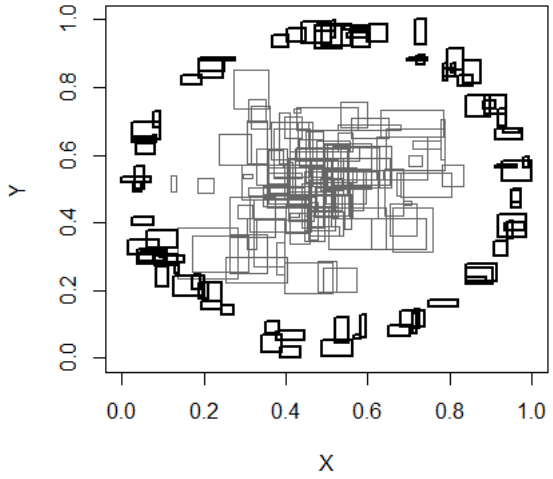

Figura 4.8: Cenário 3 do segundo conjunto.

O cenário 1 apresenta um conjunto de dados representado por dois grupos não linearmente separáveis. O cenário 2 é uma variação do primeiro cenário onde a amplitude média de um grupo foi diminuida, para testar o comportamento das técnicas na estimativa da matriz de covariância. O cenário 3 apresenta uma dispersão maior do grupo central, criado na tentativa de fazer os métodos errarem mais.

Tabela 4.4: Média e desvio padrão da área de curva ROC para o segundo conjunto de dados sintéticos

Cenário	1. IEM	2. IKFCM	3. IKEM
1	0.6019 (0.0070)	0.6547 (0.0151)	0.7197 (0.0185)
2	0.5508 (0.0233)	0.6511 (0.0139)	0.7413 (0.0095)
3	0.5312 (0.0092)	0.5885 (0.0046)	0.6084 (0.0196)

A Tabela 4.4 mostra que em todos os cenários o método linear não possui um bom desempenho.

A Tabela 4.5 contém os P-valores do teste estatístico de Friedman utilizados para comparar os resultados obtidos e mostrar se existe diferença significativa nos resultados dos métodos. A hipósese nula é de que os métodos possuem desempenhos médios semelhantes, a hipótese alternativa é de que os métodos possuem desempenhos diferentes. As médias μ_1, μ_2 e μ_3 correspondem aos métodos 1. IEM, 2. IKFCM e 3. IKEM respectivamente.

Tabela 4.5: P-valores do teste de Friedman da área da curva ROC no segundo conjunto de dados sintéticos

Cenário	$H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$	Decisão
1	7.468e-08	Rejeita H_0
2	8.725e-14	Rejeita H_0
3	3.264e-05	Rejeita H_0

O teste de Friedman mostra uma evidência muito forte contra a hipótese nula em todos os cenários, ou seja, os métodos não possuem um desempenho médio parecidos, por isso se torna necessário a aplicação de outro teste de hipótese dois a dois, para ratificar a eficiência do método proposto. A Tabela 4.5 mostra os resultados do teste de hipótese de Wilcoxon sobre a área de curva ROC dos métodos avaliados.

Tabela 4.6: P-valores do teste de Wilcoxon da área da curva ROC no segundo conjunto de dados sintéticos

Cenário	$H_0 : \mu_3 = \mu_1$ $H_1 : \mu_3 > \mu_1$	Decisão	$H_0 : \mu_3 = \mu_2$ $H_1 : \mu_3 > \mu_2$	Decisão
1	1.258e-09	Rejeita H_0	0.0004	Rejeita H_0
2	7.495e-14	Rejeita H_0	3.334e-08	Rejeita H_0
3	1.553e-05	Rejeita H_0	0.05786	Não Rejeita H_0

O teste de Wilcoxon nos mostra que o método proposto teve resultados superiores aos métodos da literatura nos Cenário 1 e 2. Já no Cenário 3 os resultados obtidos pelo método foram superiores apenas em relação ao método linear. Se espera que nesse tipo de conjunto os métodos não lineares obtenham um resultado superior ao método linear. A superioridade do IKEM em relação ao IKFCM nesse tipo de conjunto é que a distância utilizada pelo IKEM considera as dispersões dos grupos. Contudo, no cenário 3 como a dispersão do grupo centralizado é maior, o IKEM não consegue levar vantagem considerável em relação IKFCM.

4.3.3 Conjunto de dados: Agaricus

O conjunto de dados simbólicos real intervalar Agaricus, mostrado na Figura 4.9, foi utilizado no trabalho de COSTA (2011), suas informações foram extraídas a partir de espécies de fungos presentes no estado da Califórnia, Estados Unidos. Este conjunto possui 24 espécies do gênero *Agaricus*, onde cada uma é descrita por três variáveis do tipo intervalo: largura do píleo, espessura do estipe e largura dos esporos. Os dados são distribuídos entre dois grupos: comestível e não-comestível.

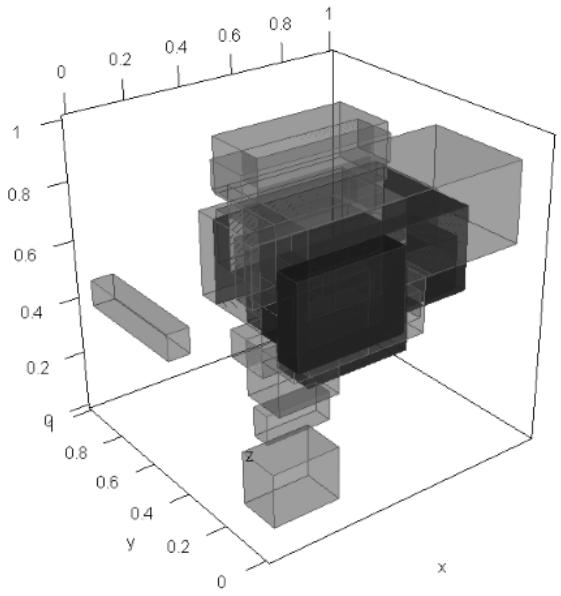

Figura 4.9: Conjunto de dados: *Agaricus*.

Tabela 4.7: Média e desvio padrão da área de curva ROC do conjunto de dados *Agaricus*

1. IEM	2. IKFCM	3. IKEM
0.6150 (0.0106)	0.5711 (0.0033)	0.6759 (0.0157)

De acordo com os valores da Tabela 4.7 todas os métodos apresentam bons resultados em relação em termos da média e desvio padrão, postos entre parênteses, da área da curva ROC.

A Tabela 4.8 contém o P-valor do teste estatístico de Friedman utilizado para comparar os resultados obtidos e mostrar se existe diferença significativa nos resultados dos métodos. A hipósese nula é de que os métodos possuem desempenhos médios semelhantes, a hipótese alternativa é de que os métodos possuem desempenhos diferentes. As médias μ_1, μ_2 e μ_3 correspondem aos métodos 1. IEM, 2. IKFCM e 3. IKEM respectivamente.

Tabela 4.8: P-valor do teste de Friedman da área da curva ROC do conjunto *Agaricus*

$H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$	Decisão
4.168e-07	Rejeita H_0

De acordo com o teste de Friedman, existe uma evidência muito forte contra a hipótese nula, logo, concluímos que os métodos não possuem um desempenho médio semelhante. A Tabela 4.9 mostra os resultados do teste de hipótese de Wilcoxon sobre a área de curva ROC dos métodos avaliados.

Tabela 4.9: P-valores do teste de Wilcoxon da área da curva ROC do conjunto *Agaricus*

$H_0 : \mu_3 = \mu_1$ $H_1 : \mu_3 > \mu_1$	Decisão	$H_0 : \mu_3 = \mu_2$ $H_1 : \mu_3 > \mu_2$	Decisão
0.0008	Rejeita H_0	6.697e-10	Rejeita H_0

O teste de Wilcoxon nos mostra que o método proposto obteve resultados superiores aos métodos da literatura neste conjunto de dados. Tendo em vista que o método IEM, mesmo sendo um método linear, teve a média da área de curva ROC maior que o IKFCM, podemos chegar à conclusão que a distância compartilhada pelo IEM e IKEM deram aos métodos superioridade em relação ao IKFCM. Contudo, o uso da função de núcleo do IKEM otimizou ainda mais seu resultado em relação ao IEM.

4.3.4 Conjunto de dados: Temperaturas das cidades

O conjunto de dados simbólicos real intervalar Temperaturas das cidades, mostrado na Figura 4.10, utilizado no trabalho de COSTA (2011), consiste de 35 cidades descritas por três variáveis do tipo intervalo que representam as temperaturas mínimas e máximas das cidades durante um ano, com cada variável representando um quadrimestre do ano: Janeiro a Abril, Maio a Agosto e Setembro a Dezembro.

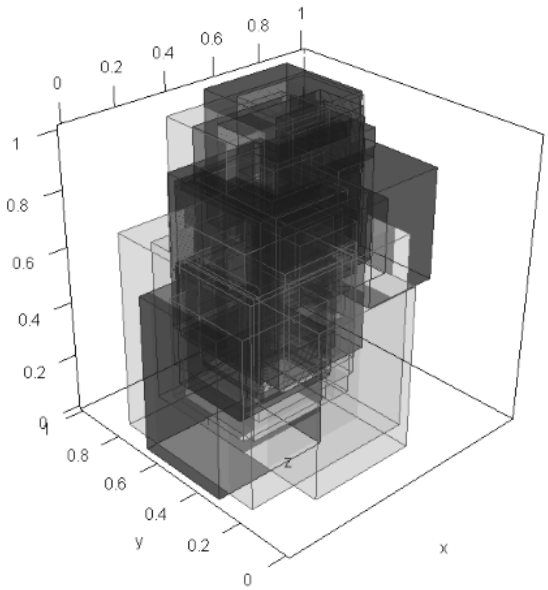

Figura 4.10: Conjunto de dados: Temperaturas de cidades.

Tabela 4.10: Média e desvio padrão da área de curva ROC do conjunto de dados Temperaturas

1. IEM	2. IKFCM	3. IKEM
0.5193 (0.0001)	0.5598 (0.0001)	0.6030 (0.0020)

De acordo com os valores da Tabela 4.10, a técnica proposta apresenta bons resultados em relação às demais técnicas da literatura, em termos da média da área da curva ROC.

A Tabela 4.11 contém o P-valor do teste estatístico de Friedman, utilizado para comparar os resultados obtidos e mostrar se existe diferença significativa nos resultados dos métodos. A hipósese nula é de que os métodos possuem desempenhos médios semelhantes, a hipótese alternativa é de que os métodos possuem desempenhos diferentes. As médias μ_1, μ_2 e μ_3 correspondem aos métodos 1. IEM, 2. IKFCM e 3. IKEM respectivamente.

Tabela 4.11: P-valor do teste de Friedman da área da curva ROC do conjunto Temperaturas

$H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$	Decisão
< 2.2e-16	Rejeita H_0

De acordo com o teste de Friedman mostrado na Tabela 4.11, existe uma evidência muito forte contra a hipótese nula, logo concluímos que os métodos não possuem um desempenho médio semelhantes. A Tabela 4.12 mostra os resultados do teste de hipótese de Wilcoxon sobre a área de curva ROC dos métodos avaliados.

Tabela 4.12: P-valores do teste de Wilcoxon da área da curva ROC do conjunto Temperaturas

$H_0 : \mu_3 = \mu_1$ $H_1 : \mu_3 > \mu_1$	Decisão	$H_0 : \mu_3 = \mu_2$ $H_1 : \mu_3 > \mu_2$	Decisão
< 2.2e-16	Rejeita H_0	2.353e-13	Rejeita H_0

O teste de Wilcoxon nos mostra que o método proposto obteve resultados superiores aos métodos da literatura neste conjunto de dados. Neste conjunto de dados reais, os métodos não lineares obtiveram resultados superior ao método linear IEM, por se tratar de um conjunto onde os grupos são misturados, dificultando o processo de agrupamento de um método linear. A IKEM apresentou uma ligeira vantagem sobre o IKFCM, apresentando resultados similares na maioria das rodadas, porém, conseguiu um resultado superior em algumas convergências, finalizando assim com uma média superior.

4.3.5 Conjunto de dados: Carros

O conjunto de dados simbólicos intervalares real Carros, mostrado na Figura 4.11, foi escolhido para a avaliação dos métodos em relação ao problema de não linearidade e sensibilidade a observações discrepantes, pois este conjunto apresenta algumas observações que podem ser consideradas como *outliers* candidatos. O conjunto de dados simbólicos Carros consiste de 33 modelos descritos por oito variáveis intervalares, duas variáveis categóricas e uma variável nominal. Este conjunto foi utilizado no trabalho de (FAGUNDES, 2010), de onde foram retiradas apenas três variáveis intervalares, sendo duas variáveis independentes descritas por: velocidade máxima e cilindrada do motor, e uma variável dependente: preço.

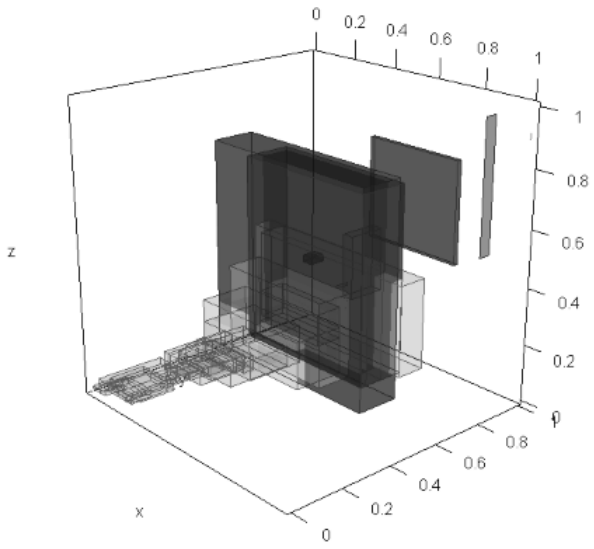

Figura 4.11: Conjunto de dados: Carros.

O grupo utilizado como alvo neste conjunto foi retirado da tese de (FAGUNDES, 2010), onde se aponta as observações que podem ser consideradas como *outliers* candidatos.

Tabela 4.13: Média e desvio padrão da área de curva ROC do conjunto de dados Carros

1. IEM	2. IKFCM	3. IKEM
0.9027 (0.0020)	0.9089 (0.0001)	0.9136 (0.0013)

De acordo com os valores da Tabela 4.13, a técnica proposta apresenta bons resultados em relação às demais técnicas da literatura em termos da média da área da curva ROC.

A Tabela 4.14, contém o P-valor do teste estatístico de Friedman utilizados para comparar os resultados obtidos e mostrar se existe diferença significativa nos resultados dos métodos. As médias μ_1, μ_2 e μ_3 correspondem aos métodos 1. IEM, 2. IKFCM e 3. IKEM respectivamente.

Tabela 4.14: P-valor do teste de Friedman da área da curva ROC do conjunto Carros

$H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$	Decisão
6.983e-08	Rejeita H_0

De acordo com o teste de Friedman, mostrado na Tabela 4.14, existe uma evidência muito forte contra a hipótese nula, logo, concluímos que os métodos não possuem um desempenho médio semelhante. Apesar de apresentarem resultados bastante parecidos, como a variância dos resultados neste conjunto foi muito pequena, a hipótese nula do teste foi rejeitada.

A Tabela 4.15 mostra os resultados do teste de hipótese de Wilcoxon sobre a área de curva ROC dos métodos avaliados.

Tabela 4.15: P-valores do teste de Wilcoxon da área da curva ROC do conjunto Carros

$H_0 : \mu_3 = \mu_1$ $H_1 : \mu_3 > \mu_1$	Decisão	$H_0 : \mu_3 = \mu_2$ $H_1 : \mu_3 > \mu_2$	Decisão
0.00209	Rejeita H_0	8.42e-08	Rejeita H_0

O teste de Wilcoxon nos mostra que o método proposto obteve resultados superiores aos métodos da literatura, para este conjunto de dados. Todos os métodos se saíram bem neste conjunto de dados, porém, apesar dos resultados serem bastante similares, a variância dos resultados neste conjunto foi muito pequena, fazendo com que a hipótese nula dos testes fosse rejeitada.

5

Conclusão

O trabalho mostra que o algoritmo proposto apresenta resultados superiores em relação ao método linear IEM, e, em algum casos, superior ao método não linear IKFCM. Foi notado que o algoritmo IKEM possui vantagem quanto ao IKFCM em bases de dados simbólicos intervalares, quando a matriz de covariância dos grupos são significativamente diferentes. Um bom exemplo disso é quando os grupos possuem dispersões do centro dos intervalos bem diferentes ou quando a amplitude média de um grupo é maior que a amplitude média do outro. Isso por ser explicado através da análise da medida de dissimilaridade utilizada pelos algoritmos. O IKEM utiliza a distância de Mahalanobis no cálculo da função de densidade, que considera a dissimilaridade entre duas observações baseada na estimativa das médias e matrizes de covariância dos grupos, enquanto o IKFCM utiliza uma distância Euclidiana que apenas calcula a dissimilaridade entre duas observações baseada apenas no protótipo dos grupos.

A principal vantagem do algoritmo proposto é a simplicidade de implementação e o tempo de convergência, pois, ao utilizar aproximações, evitamos realizar alguns cálculos muito caros ou até impraticáveis, dependendo do número de variáveis e da quantidade de observações do problema apresentado. Já a principal desvantagem desse método é que, apesar de ser paramétrico, não conseguimos visualizar os parâmetros como a média ou a matriz de covariância, já que o algoritmo não calcula diretamente os mesmos.

A distância de Mahalanobis também pode ser uma faca de dois gumes por utilizar a matriz de covariância estimada. Dependendo da inicialização das médias dos grupos, essa matriz de covariância pode ser estimada de forma errada, atrapalhando a convergência do método. Como todos os métodos baseados em protótipos a inicialização dos parâmetros tem grande influência no resultado final, pois inicializações diferentes podem gerar resultados bem distintos. Portanto, é interessante buscar a otimização da inicialização e se recomenda rodar o método diversas vezes com configurações iniciais diferentes.

5.1 Trabalhos Futuros

Um ponto a ser explorado é a utilização de outras funções de núcleo que, neste trabalho, ficou limitado ao uso da *Radial Basis Function* (RBF). Será interessante a geração de vários conjuntos sintéticos para um estudo de caso aprofundado, adicionando a utilização de outras funções de núcleo e o ajuste de parâmetros de cada função. Apesar dos bons resultados neste trabalho, a técnica proposta foi utilizada apenas para particionar dois grupos, será preciso testar seu funcionamento em uma quantidade maior de grupos. Também se vê necessário a comparação com outras técnicas de agrupamento de dados simbólicos intervalares da literatura, como por exemplo o algoritmo c-médias difuso para dados simbólicos intervalares baseado na distância de Mahalanobis, que apesar de não utilizar funções de núcleo, compartilha a mesma distância que a técnica proposta neste trabalho. Além de tudo, utilizar outros índices de validação de desempenho, como o Índice Corrigido de Rand (ICR) e o Índice de Davies e Bouldin (DB).

Referências

ALMEIDA PIMENTEL, B. de. **Agrupamento de Dados Simbolicos usando abordagem *Possibilistic***. 2013.

BILLARD, L.; DIDAY, E. **An introduction to Support Vector Machines and other kernel-based learning methods**. [S.l.]: Wiley, 2007.

CARRIZOSA, E.; GORDILLO, J. **Support Vector Regression for imprecise data**. 2007.

CARVALHO, F. de A.T. de. Fuzzy c-means clustering methods for symbolic interval data. **Pattern Recognition Letters**, [S.l.], p.423–437, 2007.

COSTA, A. F. B. F. da. **Agrupamento de Dados Simbolicos Intervalares usando funções de Kernel**. 2011.

COSTA, A. F. B. F. da; PIMENTEL, B. A.; SOUZA, R. M. C. R. de. Clustering interval data through kernel-induced feature space. **J Intell Inf Syst**, [S.l.], p.109–140, 2013.

COUVREUR, C. The EM Algorithm: a guided tour. **Computer Intensive Methods in Control and Signal Processing**, [S.l.], p.209–222, 1997.

CRISTIANINI, N.; SHAEW-TAYLOR, J. **An introduction to Support Vector Machines and other kernel-based learning methods**. [S.l.]: Cambridge University Press, 2000.

DIDAY, E. An Introduction to Symbolic Data Analysis and the Sodas Software. **Intelligent Data Analysis**, [S.l.], p.583–601, 2003.

DOMINGUES, M. A. O. **Métodos Robustos em Regressão Linear para Dados Simbólicos do Tipo Intervalo**. 2010.

DOUX, A.-C.; LAURENT, J.-P.; NADA, J.-P. **Symbolic Data Analysis With the K-Means Algorithm for User Profiling**. 1997.

FAGUNDES, R. A. A. **Métodos de Regressão Robusta e Kernel para Dados Intervalares**. 2010.

FAWCETT, T. **An introduction to ROC analysis**. 2006.

FILHO, T. D. M. E. S. **UMA ABORDAGEM ADAPTATIVA DE LEARNING VECTOR QUANTIZATION PARA CLASSIFICAÇÃO DE DADOS INTERVALARES**. 2013.

FRIEDMAN, M. **The use of ranks to avoid the assumption of normality implicit in the analysis of variance**. 1937.

GIROLAMI, M. **Mercer Kernel Based Clustering in Feature Space**. 2001.

GÖNEN, M. **Analyzing Receiver Operating Characteristic curves with SAS, (Sas Press Series). Cary, NC**: sas publishing. 2007.

JANSSENS, J. H. **Outlier Selection and One-Class Classification**. 2013.

MOGHADDAM, B.; PENTLAND, A. Probabilistic Visual Learning for Object Representation. **Pattern Analysis and Machine Intelligence**, [S.l.], v.19, p.696–710, 1997.

ROSSI, F.; CONAN-GUEZ, B. **Multi-Layer Perceptrons and Symbolic Data**. 2008.

TAKAHASHI, A. **Máquina de Vetores-Suporte Intervalar**. 2012.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. [S.l.]: Academic Press, 2006.

WANG, J.; LEE, J.; ZHANG, C. Kernel Trick Embedded Gaussian Mixture Model. **Lecture Notes in Computer Science**, [S.l.], p.159–174, 2003.

WILCOXON, F. **Individual comparisons by ranking methods**. 1945.

YU, J. A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. **Chemical Engineering Science**, [S.l.], p.506–519, 2011.