Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Modelagem Matemática e Computacional

AGRUPAMENTO SUBTRATIVO BASEADO EM KERNEL PARA DADOS SIMBÓLICOS DE NATUREZA INTERVALAR

Camila Ravena de Oliveira

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional, UFPB, da Universidade Federal da Paraíba, como parte dos requisitos necessários à obtenção do título de Mestre em Modelagem Matemática e Computacional.

Orientadores: Marcelo Rodrigo P. Ferreira
Sérgio de Carvalho Bezerra

João Pessoa
Maio de 2018

Catalogação na publicação
Seção de Catalogação e Classificação

O48a Oliveira, Camila Ravena de.
Agrupamento subtrativo baseado em kernel para dados simbólicos de natureza intervalar / Camila Ravena de Oliveira. - João Pessoa, 2018.
86 f.

Orientação: Marcelo Rodrigo Portela Ferreira.
Coorientação: Sérgio de Carvalho Bezerra.
Dissertação (Mestrado) - UFPB/CI.

1. Dados Simbólicos. 2. Agrupamento Subtrativo. 3. Agrupamento Kernel. 4. Variável Intervalar. I. Ferreira, Marcelo Rodrigo Portela. II. Bezerra, Sérgio de Carvalho. III. Título.

UFPB/BC

AGRUPAMENTO SUBTRATIVO BASEADO EM KERNEL PARA DADOS SIMBÓLICOS DE NATUREZA INTERVALAR

Camila Ravena de Oliveira

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL (PPGMMC) DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DA PARAÍBA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM MODELAGEM MATEMÁTICA E COMPUTACIONAL.

Examinada por:

Prof. Marcelo Rodrigo Portela Ferreira, D.Sc.

Prof. Sérgio de Carvalho Bezerra, D.Sc.

Prof. Eufrásio de Andrade Lima Neto, D.Sc.

Prof. Hemílio Fernandes Campos Coêlho, D.Sc.

JOÃO PESSOA, PB – BRASIL
MAIO DE 2018

A Josineide Maria, minha mãe,
companheira, confidente e amiga,
fonte de minha persistência e de-
dicação diante das dificuldades.

Agradecimentos

A Deus

Pela sua infinita bondade e misericórdia, pois sem ele eu não teria forças para essa longa jornada. Por ter iluminado meu caminho, para que a conclusão desta dissertação fosse possível.

A família

Em especial a minha mãe, Josineide, meu exemplo de vida, pelo apoio emocional, sentimental, físico e principalmente pelo amor incondicional que me foi dado, também por acreditar que esse sonho seria possível e tê-lo sonhado junto comigo. Aos meus avós maternos, José Maurício (em memória) e Maria Francisca, que sempre foram meu espelho e fonte de inspiração. Ao meu irmão, José Maurício, pelo amor e apoio. As tias e madrinhas, Josélia e Maria do Carmo, pelo apoio incondicional a todo momento. Assim como à todos os tios(as) e primos(as), em especial à Vilma, Vamberg, Vanessa, Ana, Yasmim, Bianca, Brawne e Gabriel, que estando estes longe ou perto sempre me encorajaram. A Rony, pelo companheirismo e apoio.

Aos amigos

A Izabele, Fabrício, Tainã e Rayanna, amigos de todas as horas, pelo carinho, atenção e amor. A Allisson e André, amigos e companheiros da UFPB.

Aos professores

Ao meu orientador, Marcelo Ferreira, pelo apoio e incentivo durante essa jornada. A Andrea Rocha, Hemílio Coêlho, Roberto Quirino e Sérgio de Carvalho, professores que tenho imenso carinho e respeito, pelos conselhos e incentivo em todos os momentos, dedico a eles também esta dissertação.

"Que os vossos esforços desafiem as impossibilidades, lembrai-vos de que as grandes coisas do homem foram conquistadas do que parecia impossível."

Charles Chaplin

Resumo da Dissertação apresentada ao PPGMMC/CI/UFPB como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AGRUPAMENTO SUBTRATIVO BASEADO EM KERNEL PARA DADOS SIMBÓLICOS DE NATUREZA INTERVALAR

Camila Ravena de Oliveira

Maio/2018

Orientadores: Marcelo Rodrigo P. Ferreira
Sérgio de Carvalho Bezerra

Programa: Modelagem Matemática e Computacional

Apresenta-se, nesta dissertação, extensões de métodos de agrupamento subtrativo conhecidos. O método de agrupamento subrativo para dados simbólicos de natureza intervalar (iSBC) é uma extensão do método de agrupamento subtrativo desenvolvido por Chiu (1994), já os métodos de agrupamento subtrativo baseados em *kernel* definidos por uma ou duas componentes para dados simbólicos de natureza intervalar (iKSBC1C e iKSBC2C, respectivamente) são extensões do método de agrupamento subtrativo baseado em *kernel* proposto por Kim et al. (2005). Além disso, serão propostas seis estratégias: os centróides dos métodos propostos serão dados como entradas para os métodos K-médias para dados do tipo intervalo baseado em distância L_2 proposto por De Carvalho, Brito e Bock (2006) (iKM+iSBC, iKM+iKSBC1C e iKM+iKSBC2C) e *kernel* K-médias para dados simbólicos do tipo intervalo, desenvolvido por Costa (2011) (iKKM+iSBC, iKKM+iKSBC1C e iKKM+iKSBC2C), como forma de minimizar a sensibilidade que esses métodos tem em relação a escolha do centróide para definição da partição inicial. Experimentos utilizando dados reais mostraram que os métodos subtrativos baseados em *kernel* propostos (iKSBC1C e iKSBC2C) obtiveram melhor desempenho que o método iSBC, além disso os métodos K-médias (iKM+iSBC, iKM+iKSBC1C e iKM+iKSBC2C) e *kernel* K-médias (iKKM+iSBC, iKKM+iKSBC1C e iKKM+iKSBC2C), ambos para dados simbólicos do tipo intervalo, utilizando os centróides dos métodos propostos como entradas obtiveram melhor desempenho que os métodos iKM e iKKM.

Palavras-chave: Dados Simbólicos, Agrupamento Subtrativo, Agrupamento *Kernel*, Variável Intervalar

Abstract of Dissertation presented to PPGMMC/CI/UFPB as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

KERNEL-BASED SUBTRACTIVE CLUSTERING FOR SYMBOLIC INTERVAL DATA

Camila Ravena de Oliveira

May/2018

Advisors: Marcelo Rodrigo P. Ferreira
Sérgio de Carvalho Bezerra

Program: Computational Mathematical Modelling

In this work, we present extensions for known subtractive clustering methods. The subtractive clustering method for symbolic interval data (iSBC) as an extension of the subtractive clustering method developed by Chiu (1994), as well as the *kernel*-based subtractive clustering methods defined by one or two components for symbolic interval data (iKSBC1C and iKSBC2C, respectively) as extensions of a *kernel*-based subtractive clustering method proposed by Kim et al . (2005). In addition, six strategies will be proposed: the centroids of the proposed methods will be given as inputs to the methods K-means for interval data based on L_2 distance proposed by De Carvalho, Brito and Bock (2006) (iKM+iSBC, iKM+iKSBC1C and iKM+iKSBC2C) and *kernel* K-means for symbolic data of the interval-valued developed by Costa (2011) (iKKM+iSBC, iKKM+iKSBC1C and iKKM+iKSBC2C) as a way to minimize the sensitivity of these methods to the choice of the centroid for definition of the initial partition. Experiments using real data showed that the proposed *kernel*-based subtractive clustering methods (iSBC1C and iSBC2C) obtained better performance than the iSBC method, as well as the K-means (iKM+iSBC, iKM+iKSBC1C and iKM+iKSBC2C) and *kernel* K-means (iKKM+iSBC, iKKM+iKSBC1C and iKKM+iKSBC2C) methods, both for symbolic data interval-valued, using the centroids of methods proposed as inputs for them also obtained better performance that the iKM and iKKM methods.

Keywords: Symbolic Data, Subtractive Clustering, *Kernel* Clustering, Interval-valued Variables.

Sumário

Lista de Tabelas

Capítulo 1

Introdução

Desde que a tecnologia da informação tornou-se indispensável para diversas atividades da vida moderna proporcionando um rápido crescimento do volume das informações armazenadas e dando origem a gigantescos bancos de dados (tanto em tamanho quanto em dimensionalidade e complexidade), a mineração de dados ou data mining (HASTIE, TIBSHIRANI, FRIEDMAN, 2001; HAND, MANNILA, SMYTH, 2001), passou a figurar como um dos mais importantes campos de pesquisa.

De uma maneira geral, a mineração de dados pode ser vista como um processo de extração não trivial de informações implicitamente contidas em grandes conjuntos de dados, com o objetivo de auxiliar no processo de tomada de decisões. Ou seja, a tarefa é extrair importantes padrões e tendências e entender o que os dados dizem (HASTIE, TIBSHIRANI, FRIEDMAN, 2001). Trata-se então de um campo de pesquisa interdisciplinar, contemplando métodos estatísticos (por exemplo, análise discriminante, análise de agrupamentos, regressão linear múltipla) e computacionais (por exemplo, aprendizado da máquina, inteligência artificial, tecnologia de banco de dados), com aplicações em inúmeras áreas do conhecimento, tais como astronomia, biologia, engenharia, finanças, marketing e medicina.

Tipicamente, os dados são armazenados em matrizes $n \times p$, onde n corresponde ao número de registros e p corresponde ao número de variáveis estudadas. Em muitas situações, n pode ser da ordem de milhões de registros e p, por sua vez, pode atingir a ordem de milhares de variáveis, fazendo com que cálculos como, por exemplo, a inversão de uma matriz $n \times n$, se tornem computacionalmente muito custosos. Dessa forma, tornou-se tarefa primordial sumarizar esses enormes conjuntos de dados em termos de seus conceitos subjacentes, a fim de extrair deles novos conhecimentos.

Uma alternativa é representar os dados através de listas, intervalos, distribuições e afins. Estas representações são exemplos de um tipo de dado denominado dado simbólico, que tem sido tratado principalmente pela Análise de Dados Simbólicos (ADS) — *Symbolic Data Analysis (SDA)* —, um campo de pesquisa relacionado à análise multivariada, reconhecimento de padrões e inteligência artificial. O objetivo

principal da Análise de Dados Simbólicos é desenvolver métodos adequados à análise de dados agregados, descritos por variáveis multivaloradas, onde as células das matrizes de dados podem conter conjuntos de categorias, intervalos, distribuições de freqüência, histogramas, distribuições de probabilidade, etc. (BILLARD, DIDAY, 2003; BOCK, DIDAY, 2000).

Métodos de agrupamento são técnicas utilizadas na mineração de dados com objetivo de agrupar de forma não supervisionada os casos da base de dados em grupos. Esses métodos vêm sendo aplicados em várias áreas da ciência, como: taxonomia, processamento de imagens, recuperação de informação, dentre outros. Exemplos de aplicações incluem: segmentação de mercado, segmentação de imagens, agrupar resultados de buscas na internet, agrupar páginas da internet com base em seu conteúdo, agrupar pessoas em redes sociais com base em suas preferências ou características.

Componentes cruciais no agrupamento são a representação de padrões e a medida de similaridade. Todo agrupamento é feito com objetivo de maximizar a homogeneidade dentro de cada grupo e maximizar a heterogeneidade entre os grupos. Ou seja, elementos que fazem parte de um mesmo grupo devem apresentar alta similaridade (similaridade intra-grupo), isto é, seguem um padrão similar, mas devem ser muito dissimilares de objetos de outros grupos (similaridade inter-grupos). Segundo Filippone (2008):

> "Cada padrão é normalmente representado por um conjunto de características do sistema em estudo. É muito importante notar que uma boa escolha de representação de padrões pode levar a melhorias no desempenho do grupo. Se é possível escolher um conjunto adequado de características depende do sistema em estudo. Uma vez que uma representação é fixa, é possível escolher uma medida de similaridade apropriada entre os padrões."

As técnicas de agrupamento mais populares podem ser divididas em métodos hierárquicos e métodos particionais. Os métodos hierárquicos produzem uma resposta representada por uma estrutura completa de hierarquia, i.e., uma sequência aninhada de partições do conjunto de padrões de entrada; sua saída é uma estrutura hierárquica de grupos conhecida como *dendrograma*. Por outro lado, nos métodos particionais o objetivo é obter uma partição única do conjunto de padrões de entrada em um número fixo de grupos, tipicamente através da otimização (geralmente local) de uma função objetivo; o resultado é a criação de hipersuperfícies de separação entre os grupos. Os métodos de agrupamento particionais foram desenvolvidos sob dois diferentes paradigmas: agrupamento rígido (*hard*) Jain (2010) e agrupamento difuso (*fuzzy*) Höppner et al. (1999). Nos métodos de agrupamento do tipo rígido,

os grupos são naturalmente disjuntos e não se sobrepõem. Nesse caso, cada padrão pode pertencer a um, e somente um, grupo. No caso dos métodos de agrupamento do tipo difuso, um padrão pode pertencer a todos os grupos com um certo grau de pertinência. Uma exposição detalhada dos principais métodos de agrupamento difuso pode ser encontrada em Höppner et al. (1999) ou Everitt et al. (2011). Adicionalmente, uma boa revisão sobre os vários métodos de agrupamento pode ser encontrada, por exemplo, em Jain (2010) ou em Jain, Murty e Flynn (1999).

Steinhaus (1956) desenvolveu o método de agrupamento particional mais conhecido e largamente utilizado, o K-médias, que promove o particionamento de um conjunto de dados, em K grupos disjuntos, onde cada grupo é caracterizado por um ponto central, denominado centróide, que otimiza o valor da função para o conjunto de dados através da média aritmética dos vetores pertencentes a cada grupo. Este método apresenta limitações, tais como: as soluções encontradas convergem para ótimos locais, o que o torna extremamente sensível à escolha inicial dos centróides; e ele só apresenta boas soluções para conjuntos de dados que sejam linearmente separáveis (JAIN, 2010). De Carvalho, Brito e Bock (2006) propuseram o método K-médias para dados do tipo intervalo baseado em distância L_2.

Yager e Filev (1994) desenvolveram o método de agrupamento de montanha, um método de agrupamento simples que estima os centróides do grupo construindo e modificando a função de montanha em um espaço da grade. Tal método é baseado em aumentar o espaço de dados e computar um valor potencial para cada ponto de grade com base em suas distâncias para os pontos de dados reais (CHIU, 1997). No entanto, embora o método de montanha seja eficaz para conjuntos de dados de baixa dimensão, torna-se ineficaz quando aplicado a dados de alta dimensão. Para reduzir a complexidade computacional deste método, Chiu (1994) sugeriu calcular a função potencial nos pontos de dados em vez dos pontos de grade, uma abordagem conhecida como método de agrupamento subtrativo. Usando este método, o número de "pontos de grade" a serem avaliados é igual ao número de pontos de dados, independentemente da dimensão do problema. Outra vantagem deste método é que elimina a necessidade de especificar uma resolução de grade, onde precisão e complexidade computacional devem ser considerados. Neste trabalho será proposta uma extensão desse método, para dados simbólicos de natureza intervalar.

Medidas de distância são exemplos básicos de medidas de dissimilaridade e a distância Euclidiana é a mais comumente utilizada em métodos de agrupamento particionais (rígido e difuso). Métodos de agrupamento baseados na distância Euclidiana apresentam bom desempenho quando aplicados a conjuntos de dados nos quais os grupos são aproximadamente hiperesféricos e aproximadamente linearmente separáveis. Contudo, quando a estrutura dos dados é complexa, i.e., grupos com formas não hiperesféricas e/ou padrões não-linearmente separáveis, esses métodos

podem não apresentar desempenho satisfatório. Por causa dessa limitação, diversos métodos capazes de lidar com dados cuja estrutura é complexa têm sido propostos, dentre os quais, métodos de agrupamento baseados em funções *kernel*.

A essência dos métodos baseados em *kernel* envolve a realização de um mapeamento não-linear arbitrário Φ do espaço original p-dimensional $X \subset \mathbb{R}^p$ para um espaço de dimensão mais alta (possivelmente infinita), chamado espaço de características, $\mathcal{F}$. A razão para passarmos a dimensões mais altas é que em tais dimensões pode ser possível obter grupos bem definidos e linearmente separáveis. Métodos baseados em *kernel* possuem a vantagem de que produtos internos no espaço de características podem ser expressos por um *kernel* $\mathcal{K}$.

Girolami (2002) desenvolveu um algoritmo que mapeia o espaço de entrada em um espaço de alta dimensão e executa o agrupamento nesse novo espaço, e este mapeamento para o espaço de alta dimensão é realizado através de funções de *kernel*. Diversos métodos de agrupamento baseados em *kernel* têm sido propostos modificando abordagens já existentes, tais como K-médias, fuzzy C-médias e SOM, que passaram a incorporar funções *kernel* em suas soluções (FILLIPONE et al., 2008).

O método de agrupamento *kernel* K-médias desenvolvido por Schölkopf, Smola e Müller (1998), uma extensão do K-médias, mapeia os dados originais em um espaço $\mathbb{R}^p$ de dimensão mais alta, de forma que esta representação se torne linearmente separável. Costa (2011) propôs o método *kernel* K-médias para dados simbólicos do tipo intervalo, onde cada objeto é descrito por um vetor de intervalos e diferentes funções *kernel* são adaptadas para tratar intervalos.

Kim et al. (2005) propuseram um método de agrupamento subtrativo baseado em *kernel* e neste trabalho serão propostas duas extensões desse método para dados simbólicos do tipo intervalo. Para os métodos K-médias para dados do tipo intervalo baseado em distância L_2 e *kernel* K-médias para dados simbólicos do tipo intervalo, serão propostas estratégias utilizando os centróides obtidos dos métodos subtrativos propostos nesta dissertação, como forma de minimizar a sensibilidade que esses métodos tem em relação a esolha do centróide para definição da partição inicial. Experimentos utilizando dados reais ilustram a utilidade dos métodos e estratégias propostas.

1.1 Objetivos

Os objetivos deste trabalho são:

- Propor uma extensão para o método de agrupamento subtrativo, utilizando dados simbólicos de natureza intervalar;

- Propor duas extensões, utilizando também dados simbólicos do tipo intervalo, para o método de agrupamento subtrativo baseado em kernel, utilizando:

    1. Função de *kernel* Gaussiana definida por uma componente;
    2. Função de *kernel* Gaussiana definida por duas componentes;

- Propor três estratégias, utilizando os centróides das três extensões propostas como entrada, para o método de agrupamento K-médias;
- Propor três estratégias, também utilizando os centróides das três extensões propostas como entrada, para o método de agrupamento *kernel* K-médias;
- Implementar os algoritmos das extensões e estratégias propostas;
- Avaliar os métodos propostos através do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA), utilizando conjuntos de dados reais.

1.2 Organização da dissertação

Além deste capítulo, esta dissertação de mestrado está organizada da seguinte maneira:

- **Capítulo 2:** este capítulo se divide em quatro seções, a primeira apresenta o histórico dos métodos de agrupamento; a segunda apresenta como surgiu, onde é aplicada a Análise de Dados Simbólicos (ADS) e como as variáveis simbólicas dividem-se; a terceira apresenta como os dados simbólicos do tipo intervalo são apresentados e a quarta descreve o método de agrupamento de montanha;
- **Capítulo 3:** apresenta os métodos para agrupamento de dados utilizados como base para as extensões e estratégias propostas neste trabalho e uma breve revisão sobre agrupamento baseado em *kernel* para dados simbólicos e dados simbólicos intervalares;
- **Capítulo 4:** introduz os métodos de agrupamento subtrativo para dados simbólicos de natureza intervalar propostos (iSBC, iKSBC1C e iKSBC2C) e as estratégias utilizando os centróides dos métodos propostos como entrada para os métodos K-médias (iKM+iSBC, iKM+iKSBC1C e iKM+iKSBC2C) e *kernel* K-médias (iKKM+iSBC, iKKM+iKSBC1C e iKKM+iKSBC2C), bem como os esquemas dos algoritmos utilizados para cada método;

- **Capítulo 5:** mostra que para comparar e avaliar os métodos de agrupamento considerados nesta dissertação serão utilizados o Índice de Rand Ajustado (IRA) e a taxa total de erro de alocação (TEA), além de expor como esses índices são obtidos;

- **Capítulo 6:** descreve os conjuntos de dados reais utilizados e expõe os resultados obtidos com a execução dos métodos e estratégias desenvolvidas no Capítulo 4;

- **Capítulo 7:** apresenta as conclusões desta dissertação e as sugestões para trabalhos futuros;

- Esquemas dos algoritmos dos métodos para agrupamento de dados utilizados como base para as extensões e estratégias propostas e os conjuntos de dados reais utilizados nesta dissertação.

Capítulo 2

Revisão Bibliográfica

2.1 Histórico dos Métodos de Agrupamento

O método K-médias foi proposto por Steinhaus (1956). Logo depois foram lançados o FORGY, em Forgy (1965) e o ISODATA, em Ball e Hall (1965), dois métodos variantes do K-médias, que realizam uma realocação iterativa e auto-organizável dos dados, respectivamente. Em 1973 surgiu uma extensão do K-médias baseada em lógica fuzzy, o Fuzzy C-médias, proposto por Dunn (1973). Backer (1978) propôs uma outra abordagem de agrupamento baseada em lógica fuzzy, o método Backer. Linde, Buzo e Gray (1980) propuseram uma variante do K-médias, utilizando a distância Itakura-Saito para quantização vetorial no processamento da fala. Depois, Bezdek (1981) desenvolveu o Método Fuzzy C-médias Melhorado.

No ano de 1989, um método baseado na técnica de rede neural artificial foi proposto por Kohonen (1989), o SOM (Mapas Auto Organizáveis). O K-medoids foi proposto em Kauffman e Rousseeuw (1990), onde os grupos são representados usando a mediana dos dados em vez da média. Yager e Filev (1994) propuseram o método de agrupamento de montanha, um algoritmo fuzzy a fim de estimar os centróides dos grupos. Ainda em 1994, Chiu (1994) propôs um extensão do método de agrupamento da montanha, conhecido como método de agrupamento subtrativo.

Chinrungrueng and Séquin (1995), propuseram o Opt- K-médias - Optimal adaptative K-means, que tem ajuste dinâmico da taxa de aprendizagem e pode ser ajustado, sem envolver quaisquer atividades do usuário. K-médias com métrica de distância Mahalanobis foi utilizada para detectar grupos hiperelipsoidais em Mao e Jain (1996).

Schölkopf, Smola e Müller (1998) desenvolveram o *Kernel* K-médias, que é uma proposta baseada em *kernel* para detectar grupos de formatos arbitrários. E Bradlay, Fayyad e Reina (1998) apresentou o Fast K-médias, uma versão do K-médias, que não necessita de todos os dados para se encaixar na memória, ao mesmo tempo e

opera em uma única passagem sobre pontos de dados.

Um dos primeiros métodos de agrupamento baseado em Computação Evolucionária foi o GKA (Genetic k-means Algorithm), proposto por Krishna e Murtyg (1999). Pelleg e Moore (1999) apresentaram o Kd-tree K-médias, onde uma árvore k-dimensional é usada para identificar de forma eficiente os centros de grupos mais próximos para todos os pontos de dados, no K-médias. Steinbach, Karypis e Kumar (2000) propuseram uma versão hierárquica divisiva do K-médias, o K-médias Divisivo, que recursivamente particiona os dados em dois grupos em cada etapa. Recentes avanços no K-médias e outros algoritmos de agrupamento baseados no erro médio quadrático com suas aplicações foram propostas por Hansen e Mladenović (2001) com o J-médias, uma nova heurística de busca local para a soma mínimos quadrados.

Cheung (2003) apresentou o K*-médias, uma outra generalização do algoritmo K-médias, que é aplicável a grupos com formatos de elipse e esfera, e não pré determina o número de grupos. Likas, Vlassis and Verbeek (2003) propuseram um algoritmo K-médias global, constituído por uma série de processos de agrupamento K-médias com o número de grupos que variam de 1 a K, independente das partições iniciais e fornece aceleração computacional.

Har-Peled e Mazumdar (2004) propuseram o método coreset K-médias, que primeiro resume um grande conjunto em um subconjunto relativamente pequeno de dados, e em seguida, aplica os algoritmos de agrupamento ao resumido conjunto de dados. Banerjee et al. (2005) utilizaram a família de distâncias de Bregman para K-médias.

No ano de 2005, Mota Filho (2005) lançaram o método C-médias com GA - Genetic Algorithm, que aplica o método Fuzzy C-médias na fase de avaliação do cromossomo em Algoritmo Genético. Camastra e Verri (2005) apresentam NKM-Novel Kernel Method, um método baseado em *kernel* que usa SVM-Support Vector Machine e tem inspiração no K-médias para obter superfícies de separação naturalmente não-lineares dos dados. Kim et al. (2005) apresentaram uma versão kernelizada do método de agrupamento subtrativo.

De Carvalho, Brito e Bock (2006) propuseram o método K-médias para dados do tipo intervalo baseado em distância L_2. Arthur e Vassilvitskii (2007) lançaram o K-médias^{++}. Em 2008, uma variante do K-médias utilizando distância L_1 foi proposta por Kashima et al. (2008). Costa (2011) apresentou o *Kernel* K-médias para dados simbólicos do tipo intervalo.

2.2 Dados Simbólicos

A Análise de Dados Simbólicos (ADS) - Symbolic Data Analysis (SDA) foi desenvolvida para lidar com grandes conjuntos de dados e surgiu, simultaneamente, da influência das áreas de Análise Exploratória de Dados (TUKEY, 1977; DIDAY et al., 1982; LEBART, MORINEAU, WARWICK, 1984), Inteligência Artificial (MICHALSKI, 1973; RUSSEL, NORVIG, 2010) e Taxonomia Numérica (SNEATH, SOKAL, 1973; HAYES-ROTH, MCDERMOTT, 1978). Diday (1986) foi um dos primeiros trabalhos com os princípios básicos da abordagem simbólica.

Um conjunto de dados pode ser estruturado como um conjunto de dados simbólico, quando agregado para ser estabelecido de forma mais gerenciável (BILLARD; DIDAY, 2003), no entanto o máximo de informações sobre os dados devem ser preservadas. A principal diferença entre dados simbólicos e convencionais é que as células das matrizes de dados simbólicos podem conter conjuntos de categorias, intervalos, distribuições de freqüência, histogramas, distribuições de probabilidade, dentre outros (DIDAY, 1987; BOCK, DIDAY, 2000; BILLARD, DIDAY, 2003). Algum procedimento estatístico para a análise de dados convencional deve ser adaptado em conformidade à análise de dados simbólicos, de acordo com o tipo de dados simbólicos utilizados. Bock e Diday (2000), Billard e Diday (2003, 2006) e Diday e Noirhomme-Fraiture (2008) fornecem descrições para essas adaptações.

As variáveis simbólicas dividem-se basicamente em variáveis multi-valoradas, modais e do tipo intervalo. Os dados do tipo intervalo são o tipo de dados simbólicos mais popular na literatura e os mais comumente encontrados na prática; e os hipercubos bidimensionais e tridimensionais vieram naturalmente para a visualização deste tipo de dados (KAO et al., 2014). Foram propostos muitos métodos para dados do tipo intervalo, incluindo a análise de componentes principais (CHOUAKRIA, DIDAY, CAZES, 1998; PALUMBO, LAURO, 2003; GIOIA, LAURO, 2006; HAMADA, MINAMI e MIZUTA, 2008), análise de agrupamento (BOCK, 2002; BRITO, 2002; SOUZA, de CARVALHO, 2004; EL GOLLI, CONAN-GUEZ, ROSSI, 2004; DE CARVALHO, BRITO, BOCK, 2006; CHAVENT et al., 2006; BOCK, 2008), análise discriminante (LAURO, VERDE e PALUMBO, 2000; DUARTE SILVA e BRITO, 2006), modelos de regressão (BILLARD, DIDAY, 2000; LIMA NETO, de CARVALHO, 2008, 2010), dentre outros.

Usando dados simbólicos do tipo de intervalo, cada linha da matriz de dados convencionais contém um vetor de valores numéricos, enquanto cada linha na matriz de dados simbólicos contém um vetor de intervalos, que descrevem o comportamento de um grupo de amostras.

2.2.1 Dados Simbólicos do Tipo Intervalo

Segundo Billard e Diday (2003) suponha que estamos interessados na variável intervalar Z, e que o valor de observação para o objeto u seja o intervalo $Z(u) = [a_u, b_u]$, para $u \in E = \{1, ..., m\}$, $E \in \mathbb{R}^p$, $a_u = \{a_1, ..., a_m\}$ e $b_u = \{b_1, ..., b_m\}$.

Assumimos que os vetores de descrição individuais $x \in \mathbb{R}^p$ estejam uniformemente distribuídos ao longo do intervalo $Z(u)$. Além disso, cada objeto é considerado igualmente provável de ser observado com probabilidade $\frac{1}{m}$. Portanto, a função de distribuição empírica, $F_Z(\xi)$, é uma mistura de m distribuições uniforme $\{Z(u), u = 1, \ldots, m\}$. Portanto, de acordo com Bertrand e Goupil (2000) pode-se mostrar que a função de densidade empírica de Z é

$$f(\xi) = \frac{1}{m} \sum_{u \in E} \frac{I_u(\xi)}{||Z(u)||}, \xi \in \Re \tag{2.1}$$

em que $I_u(.)$ é a função de indicador de ξ que está ou não está no intervalo $Z(u)$ e $||Z(u)||$ é tamanho desse intervalo. Observe que o somatório em (2.1) é apenas sobre os objetos u para os quais $\xi \in Z(u)$.

Para construir um histograma, faça $I = [\min_{u \in E}(a_u), \max_{u \in E}(b_u)]$ o intervalo que abrange todos os valores observados de Z em X, e suponha que partimos I em r subintervalos $I_g = [\xi_{g-1}, \xi_g], g = 1, \ldots, r-1$, e $I_r = [\xi_{r-1}, \xi_r]$. Então o histograma para Z é a representação gráfica da distribuição de freqüência $\{(I_g, p_g), g = 1, \ldots, r\}$, em que

$$p_g = \frac{1}{m} \sum_{u \in E} \frac{||Z(u) \cap I_g||}{||Z(u)||} \tag{2.2}$$

isto é, p_g é a probabilidade de que um vetor arbitrário de descrição individual x esteja no intervalo I_g. Se quisermos plotar o histograma com altura f_g no intervalo I_g, de modo que a "área" seja p_g, então $p_g = (\xi_g - \xi_{g-1}) f_g$.

De acordo com Billard e Diday (2003), a média amostral simbólica dos centróides para uma variável do tipo intervalo Z é dada por:

$$\bar{Z} = \frac{1}{2m} \sum_{u \in E} (b_u + a_u) \tag{2.3}$$

e a variância amostral simbólica dos centróides é dada por:

$$S^2 = \frac{1}{3m} \sum_{u \in E} (b_u^2 + 2 b_u a_u + a_u^2) - \frac{1}{4m^2} [\sum_{u \in E} (b_u + a_u)]^2 \tag{2.4}$$

2.3 Método de Agrupamento de Montanha

Yager e Filev (1994) propuseram um algoritmo simples e efetivo, chamado método de montanha, para estimar o número e a localização inicial dos centróides dos grupos. O método deles é baseado em aumentar o espaço de dados e computar um valor potencial para cada ponto de grade com base em suas distâncias para os pontos de dados reais (CHIU, 1997). Um ponto de grade com muitos pontos de dados nas proximidades terá um alto valor potencial. O ponto de grade com o maior valor potencial é escolhido como primeiro centróide do grupo. A idéia-chave em seu método é que, uma vez escolhido o primeiro centróide de grupo, o potencial de todos os pontos de grade é reduzido de acordo com a distância do centróide do grupo. Pontos de grade próximos do primeiro centróide de grupo terão alto valor potencial. O próximo centróide do grupo é então colocado no ponto de grade com o maior valor potencial restante. Este procedimento de aquisição do novo centróide do grupo e a redução do potencial dos pontos de grade circundantes repete-se até que o potencial de todos os pontos de grade seja menor do que um dado limite δ. O δ se torna um fator importante, se δ for muito grande podemos perder alguns grupos importantes e se for muito pequeno podemos ter muitos grupos. É difícil especificar δ como uma constante que funcione bem para todos os exemplos (YANG; WU, 2005).

Yager e Filev (1994) consideram um conjunto de dados $X = \{\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n\}$ no espaço p-dimensional $\mathbb{R}^p$. Seja $\underset{\sim}{x}_{jk}$ a coordenada k-ésima do j-ésimo ponto de dados para $1 \leq j \leq n$ e $1 \leq k \leq p$. O espaço p-dimensional $\mathbb{R}^p$ é restrito a um hipercubo p-dimensional I_1 x I_2 x ... x I_p onde os intervalos I_k, $1 \leq k \leq p$ são definidos pelos intervalos das coordenadas $\underset{\sim}{x}_{jk}$, isto é:

$$I_k = [\min_{1 \leq j \leq n} (\underset{\sim}{x}_{jk}), \max_{1 \leq j \leq n} (\underset{\sim}{x}_{jk})] \tag{2.5}$$

Evidentemente, o hipercubo contém o conjunto de dados X. Então os intervalos I_k são subdivididos em r_k pontos equidistantes. Essa discretização forma uma grade p-dimensional no hipercubo com os pontos $N = \{(y_1, ..., y_p), y_1 \in \{1, ..., r_1\}, ..., y_p \in \{1, ..., r_p\}\}$. Devemos denotar as coordenadas equidistantes dos pontos de grade por $X_i^{(k)}, ..., X_{rk}^{(k)}$, onde os pontos quantizam o intervalo I_k e $k = (1, p)$. A grade discretiza o espaço restrito pelo hipercubo. De acordo com Yager e Filev (1994):

> "Enfatizamos o significado desta discretização: os pontos de grade e os potenciais centróides dos grupos. Por um lado, quanto mais grosseira a discretização, menos pontos, menos cálculos são necessários, mas também são mais grosseiros os valores finais dos centróides dos grupo. Enfatizamos o fato de que o processo de obtenção do conjunto de potenciais centróides dos grupos, os pontos, não precisa ser baseado em uma grade

uniforme do espaço, podemos usar uma grade variável do espaço. Mas, geralmente, qualquer técnica que ofereça uma seleção representativa de pontos do espaço pode ser usada."

No método de montanha, os centróides dos grupos são restritos aos pontos $\underset{\sim}{y}_i$ da grade N em $\mathbb{R}^p$ e a função de montanha é calculada com

$$M(\underset{\sim}{y}_i) = \sum_{j=1}^{n} e^{-\alpha d(\underset{\sim}{x}_j, \underset{\sim}{y}_i)}, i = 1, 2, ... \tag{2.6}$$

em que α é uma constante positiva e $d(\underset{\sim}{x}_j, \underset{\sim}{y}_i)$ é a medida de distância entre o ponto de dados $\underset{\sim}{x}_j$ e os pontos $\underset{\sim}{y}_i$ da grade N. Segundo Yang e Wu (2005) o parâmetro α na equação (2.6) é muito importante, pois determina o raio da vizinhança em que os pontos de dados que estiverem fora deste raio terão uma pequena influência na função de montanha e também determina a forma de densidade aproximada do conjunto de dados.

Os valores das funções de montanha dos pontos podem ser considerados intimamente relacionados com a densidade dos pontos de dados na vizinhança. Eles também representam a capacidade potencial de um ponto de grade ser uma estimativa do centróide do grupo (YAGER; FILEV, 1994). Um ponto com muitos pontos de dados de vizinhança terá um grande valor de função de montanha. Seja M_1^* o valor máximo da função de montanha:

$$M_1^* = \max_{1 \leq i \leq n} [M(\underset{\sim}{y}_i)] \tag{2.7}$$

e $\underset{\sim}{y}_1^*$ é o primeiro centróide de grupo entre todos os pontos de grade. A função de montanha modificada usada para encontrar os centróides dos grupos subsequentes é definida como:

$$\widehat{M}^j(\underset{\sim}{y}_i) = \widehat{M}^{j-1}(\underset{\sim}{y}_i) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta d(\underset{\sim}{y}_{j-1}^*, \underset{\sim}{y}_i)} \tag{2.8}$$

onde β é uma constante positiva, $\widehat{M}^j$ é a nova função de montanha, $\widehat{M}^{j-1}$ a função de montanha antiga, M_{j-1}^* o valor máximo de $\widehat{M}^{j-1}$, e a localização desse valor máximo $\underset{\sim}{y}_{j-1}^*$ é o centróide de grupo recém-encontrado.

Na função de montanha modificada (2.8), os valores da função de montanha para os pontos que estão próximos dos centróides de grupo recém-encontrados, serão fortemente reduzidos e o parâmetro β determinará o raio da vizinhança que terá reduções mensuráveis na função de montanha.

Yager e Filev (1994) acharam mais efetivo demonstrar o funcionamento do método da montanha, assim:

$$\widehat{M}^j(\underset{\sim}{y}_i) = \max\{\widehat{M}^{j-1}(\underset{\sim}{y}_i)) - M^*_{j-1} \sum_{j=1}^{n} e^{-\beta d(\underset{\sim}{y}^*_{j-1}, \underset{\sim}{y}_i)}, 0\} \tag{2.9}$$

A aquisição de novos centróides de grupos e a redução da função de montanha repete-se até que o nível máximo atual M^*_{j-1} em comparação com o máximo original M^*_1, torne-se muito baixo. Isso significa que existem apenas alguns pontos em torno deste centróide de grupo e podem ser omitidos. Deve-se parar o processo de modificação da função da montanha quando a proporção:

$$\frac{M^*_{j-1}}{M^*_1} < \delta \tag{2.10}$$

onde δ é um parâmetro limite dado; p^* é denotado como o passo que satisfaz o critério de parada (2.10). Obviamente $(p^* - 1)$ é definido como centróides de grupo.

Este método é simples e eficaz para conjuntos de dados de baixa dimensão, no entanto a computação cresce exponencialmente quando aplicado a dados de alta dimensão (CHIU, 1997; KIM et al., 2005). Para reduzir a complexidade computacional deste método, Chiu (1994) sugeriu calcular a função de montanha nos pontos de dados em vez dos pontos de grade, uma abordagem conhecida como método de agrupamento subtrativo e no Capítulo 3 este método será apresentado em detalhes.

Capítulo 3

Métodos para Agrupamento de Dados

3.1 Método de Agrupamento K-médias

O método de agrupamento K-médias (rotulado nesta dissertação como KM) promove o particionamento de um conjunto de dados, descrito por $X = \{\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n\}$, $\Omega = \{1, ..., n\}$, em K grupos disjuntos, $P = \{P_1, ..., P_K\}$. O método busca formar grupos de modo que a função objetivo $J(P)$ seja minimizada:

$$J(P) = \sum_{k=1}^{K} \sum_{i \in P_k} ||\underset{\sim}{x}_i - \underset{\sim}{v}_k||^2 \tag{3.1}$$

em que $\underset{\sim}{v}_k$ é o centróide do k-ésimo grupo e $||\underset{\sim}{x}_i - \underset{\sim}{v}_k||^2$ é o quadrado da distância Euclidiana entre $\underset{\sim}{x}_i$ e $\underset{\sim}{v}_k$, denotada por:

$$\begin{aligned} ||\underset{\sim}{x}_i - \underset{\sim}{v}_k||^2 &= (\underset{\sim}{x}_i - \underset{\sim}{v}_k)^T (\underset{\sim}{x}_i - \underset{\sim}{v}_k) \\ &= \sum_{j=1}^{p} (x_{ij} - v_{kj})^2 \end{aligned} \tag{3.2}$$

A minimização da função objetivo $J(P)$ dada pela equação (3.1) com respeito ao centróide do k-ésimo grupo $\underset{\sim}{v}_k$ fornece a seguinte equação de atualização para os centróides dos grupos:

$$\underset{\sim}{v}_k = \frac{1}{|P_k|} \sum_{i \in P_k} \underset{\sim}{x}_i, k = 1, ..., K \tag{3.3}$$

Na definição da melhor partição do conjunto de padrões de entrada Ω, os centróides dos grupos $\underset{\sim}{v}_k$ $(k = 1, ..., K)$ estão fixos. Os grupos P_k $(k = 1, ..., K)$, que minimizam a função objetivo $J(P)$ dada são então atualizados de acordo com a seguinte regra de alocação:

$$P_k = \{i \in \Omega : ||\underset{\sim}{x}_i - \underset{\sim}{v}_k||^2 \leq ||\underset{\sim}{x}_i - \underset{\sim}{v}_r||^2, \forall r \neq k, r = 1, ..., K\} \tag{3.4}$$

3.2 Método de Agrupamento K-médias para Dados Simbólicos de Natureza Intervalar

Considere um conjunto de dados, descrito por $X = \{\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n\}$, $\Omega = \{1, ..., n\}$ e $\underset{\sim}{x}_i = ([a_i^1, b_i^1], ..., [a_i^p, b_i^p])$ $\forall i \in \Omega$, particionado em K grupos no espaço de entradas do tipo intervalo e $G = \{\underset{\sim}{g}_1, ..., \underset{\sim}{g}_K\}$ como os centróides dos grupos onde $\underset{\sim}{g}_k = ([\alpha_k^1, \beta_k^1], ..., [\alpha_k^p, \beta_k^p])$, $\forall k = 1, ..., K$. O método de agrupamento K-médias para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKM) consiste em encontrar uma partição em K grupos disjuntos $P^* = \{P_1, ..., P_K\}$ de modo que a seguinte função objetivo $J(P^*, G)$ seja minimizada:

$$J(P^*, G) = \sum_{k=1}^{K} \sum_{i \in P_k^*} ||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2 \tag{3.5}$$

em que $\underset{\sim}{g}_k \in \mathbb{R}^p$ x $\mathbb{R}^p$ é o centróide do k-ésimo grupo ($k = 1, ..., K$) e $||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2$ é o quadrado da distância Euclidiana entre $\underset{\sim}{x}_i$ e $\underset{\sim}{g}_k$, denotada por:

$$||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2 = (\underset{\sim}{x}_i - \underset{\sim}{g}_k)^T(\underset{\sim}{x}_i - \underset{\sim}{g}_k)$$

$$= (\underset{\sim}{x}_i)^T(\underset{\sim}{x}_i) - 2(\underset{\sim}{x}_i)^T(\underset{\sim}{g}_k) + (\underset{\sim}{g}_k)^T(\underset{\sim}{g}_k) \tag{3.6}$$

em que $||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2 = \sum_{k=1}^{K}(a_i - \alpha_k)^2 + \sum_{k=1}^{K}(b_i - \beta_k)^2$.

A minimização da função objetivo $J(P^*, G)$ dada pela equação (3.4) com respeito ao centróide do k-ésimo grupo $\underset{\sim}{g}_k$ fornece a seguinte equação de atualização para os centróides dos grupos:

$$\underset{\sim}{g}_k = \frac{1}{|P_k^*|} \sum_{i \in P_k^*} \underset{\sim}{x}_i, k = 1, ..., K \tag{3.7}$$

Na definição da melhor partição do conjunto de padrões de entrada, os centróides dos grupos $\underset{\sim}{g}_k$ ($k = 1, ..., K$) estão fixos. Os grupos P_k^* ($k = 1, ..., K$), que minimizam a função objetivo $J(P^*, G)$ dada são então atualizados de acordo com a seguinte regra de alocação:

$$P_k^* = \{i \in \Omega : ||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2 \leq ||\underset{\sim}{x}_i - \underset{\sim}{g}_s||^2, \forall s \neq k, s = 1, ..., K\} \tag{3.8}$$

3.3 Método de Agrupamento Subtrativo

Chiu (1994) propôs uma extensão do método de montanha [Yager e Filev (1994)], chamado de método de agrupamento subtrativo (rotulado nesta dissertação como SBC), no qual cada ponto de dados é considerado como um potencial centróide do grupo. Usando este método, o número de "pontos de grade" a serem avaliados é igual ao número de pontos de dados, independentemente da dimensão do problema.

O método de agrupamento subtrativo é semelhante ao método de montanha. O método de agrupamento subtrativo funciona da seguinte forma: considere uma coleção de n pontos de dados $X = \{\underset{\sim}{x}_1, \underset{\sim}{x}_2, ..., \underset{\sim}{x}_n\}$ com $\underset{\sim}{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$. Sem perda de generalidade, assumimos que os pontos de dados foram padronizados em cada dimensão. Consideramos cada ponto de dados como um potencial centróide do grupo e definimos a função potencial $M(\underset{\sim}{x}_i)$ da seguinte maneira:

$$M(\underset{\sim}{x}_i) = \sum_{j=1}^{n} e^{-\alpha||\underset{\sim}{x}_i - \underset{\sim}{x}_j||^2} \tag{3.9}$$

em que $\alpha = \dfrac{4}{r_a^2}$, sendo $r_a = 0.5$ e $||\underset{\sim}{x}_i - \underset{\sim}{x}_j||^2$ é o quadrado da distância Euclidiana entre $\underset{\sim}{x}_i$ e $\underset{\sim}{x}_j$, denotada por:

$$\begin{aligned} ||\underset{\sim}{x}_i - \underset{\sim}{x}_j||^2 &= (\underset{\sim}{x}_i - \underset{\sim}{x}_j)^T(\underset{\sim}{x}_i - \underset{\sim}{x}_j) \\ &= \sum_{k=1}^{p}(x_{ik} - x_{jk})^2 \end{aligned} \tag{3.10}$$

Usando a função potencial $M(\underset{\sim}{x}_i)$ da equação (3.7), os centróides dos grupos são selecionados de forma semelhante à usada no método de montanha original. Seja M_1^* o valor máximo da função potencial:

$$M_1^* = \max_{1 \le i \le n}[M(\underset{\sim}{x}_i)] \tag{3.11}$$

e $\underset{\sim}{x}_i^*$ é o ponto de dados cujo valor potencial é M_1^*, este ponto de dados é selecionado como o primeiro centróide do grupo.

A função potencial modificada usada para encontrar os centróides dos grupos subsequentes é definida como:

$$\widehat{M}^j(\underset{\sim}{x}_i) = \widehat{M}^{j-1}(\underset{\sim}{x}_i) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta||\underset{\sim}{x}_i - \underset{\sim}{x}_{j-1}^*||^2} \tag{3.12}$$

em que $\beta = \dfrac{4}{r_b^2}$, sendo r_b uma constante positiva e $\underset{\sim}{x}_{j-1}^*$ é o centróide recém-encontrado. Assim, subtraímos a função potencial da distância de cada ponto de dados a partir do primeiro centróide. Os pontos de dados próximos ao primeiro

centróide terão um potencial muito reduzido e, portanto, será improvável serem selecionados como o próximo centróide. O r_b constante é efetivamente o raio que define a vizinhança que terá reduções mensuráveis no potencial. Para evitar a obtenção de centróides bem espaçados, estabelecemos que r_b seja um pouco maior do que r_a , uma boa escolha é $r_b = 1.5r_a$ (CHIU, 1994).

3.4 Agrupamento Baseado em Kernel

Uma maneira de aumentar a precisão dos métodos de agrupamento é explorando uma função *kernel* no cálculo do valor potencial de cada ponto de dados, mapeando os pontos de dados do espaço de entrada para um espaço dimensional elevado em que a distância é medida usando uma função *kernel*. Os métodos de agrupamento baseados em *kernel*, calculados em um espaço dimensional elevado, são muito mais informativos que os métodos de agrupamento convencionais calculados no espaço original, levando a uma seleção mais precisa dos centróides dos grupos.

Uma função *kernel* é uma generalização da distância métrica que mede a distância entre dois pontos de dados à medida que os pontos de dados são mapeados em um espaço dimensional elevado no qual os dados são mais claramente separáveis (MÜLLER et al., 2001; GIROLAMI, 2002).

Dado um conjunto de dados $X = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_n\}$ no espaço p-dimensional $\mathbb{R}^p$, seja Φ uma função de mapeamento não linear a partir desse espaço de entrada para um espaço de características de alta dimensão $\mathcal{F}$:

$$\Phi : \mathbb{R}^p \to \mathcal{F}, \underline{x} \to \Phi(\underline{x}) \tag{3.13}$$

Consideremos o produto ${\underline{x}_i}^T \underline{x}_j$, referido como produto interno, que é usado como medida de similaridade em uma variedade de métodos de aprendizado da máquina. Ao aplicar a função de mapeamento não-linear Φ, o produto ${\underline{x}_i}^T \underline{x}_j$ no espaço de entrada é mapeado para $\Phi(\underline{x}_i)^T \Phi(\underline{x}_j)$ no espaço de características, que é uma medida de similaridade mais geral (SCHÖLKOPF; SMOLA, 2001).

A idéia chave no aprendizado baseado em *kernel* é que a função de mapeamento Φ não precisa ser explicitamente especificada, o produto interno no espaço de características de alta dimensão pode ser calculado através da função *kernel* $\mathcal{K}(\underline{x}_i, \underline{x}_j)$ no espaço de entrada $\mathbb{R}^p$ (SCHÖLKOPF; SMOLA, 2001)

$$\mathcal{K}(\underline{x}_i, \underline{x}_j) = \Phi(\underline{x}_i)^T \Phi(\underline{x}_j) \tag{3.14}$$

Exemplos de funções *kernel* tipicamente utilizadas são:

- Linear: $\mathcal{K}^{(l)}(\underline{x}_i, \underline{x}_j) = {\underline{x}_i}^T \underline{x}_j$

- Polinomial de grau p: $\mathcal{K}^{(p)}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) = (1 + \underset{\sim}{x}_i{}^T \underset{\sim}{x}_j)^p, p \in \mathbb{N}$
- Gaussiana: $\mathcal{K}^{(g)}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) = e^{-\frac{||\underset{\sim}{x}_i - \underset{\sim}{x}_j||^2}{2\sigma^2}}$, $\sigma \in \mathbb{R}$

em que, σ é parâmetro do *kernel*.

3.5 Agrupamento Baseado em Kernel para Intervalos

Um conjunto de dados simbólicos de natureza intervalar X (DIDAY; NOIRHOMME-FRAITURE, 2008) é uma correspondência definida de Ω em $\mathbb{R}^p$ tal que $i \in \Omega$, $X(i) = [a, b] \in I$, onde I é um conjunto de intervalos definido no espaço p-dimensional $\mathbb{R}^p$x$\mathbb{R}^p$. Considere um conjunto de n objetos ou padrões $\Omega = 1, ..., n$. Cada objeto i descrito por p variáveis simbólicas do tipo intervalo, é representado como um vetor de intervalos $\underset{\sim}{x}_i = ([a_i^1, b_i^1], ..., [a_i^p, b_i^p])^T$.

Considere $\Phi : \mathbb{R}^p$x$\mathbb{R}^p \rightarrow \mathcal{F}$, $\underset{\sim}{x}_i \rightarrow \Phi(\underset{\sim}{x}_i)$ uma função não-linear que realiza um mapeamento do espaço de entrada para intervalos finito X para um espaço de características de alta dimensão $\mathcal{F}$.

A escolha adequada da função *kernel* é um fator importante para que o problema não-linear no espaço de entrada possa ser linearmente separável no espaço de características de alta dimensão e uma escolha comum é a utilização da função *kernel* Gaussiana. Costa, Pimentel e Souza (2010) definiram duas funções *kernel* Gaussiana para intervalo e que serão utilizadas neste trabalho:

1. Função de *kernel* Gaussiana definida por uma componente:

$$\mathcal{K}^{(g)}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) = exp(-\frac{||\underset{\sim}{x}_i - \underset{\sim}{x}_j||^2}{2\sigma^2}) \tag{3.15}$$

   em que $||\underset{\sim}{x}_i - \underset{\sim}{x}_j||^2 = \sum_{j-1}^{p}[(x_i^I - x_j^I)^2 + (x_i^S - x_j^S)^2]$. Considere $\underset{\sim}{x}_i = (x_{i_1}^I, x_{i_1}^S, ..., x_{i_p}^I, x_{i_p}^S)^T$ e $\underset{\sim}{x}_j = (x_{j_1}^I, x_{j_1}^S, ..., x_{j_p}^I, x_{j_p}^S)^T$ como vetores de $2p$-dimensões que descrevem, respectivamente, o i-ésimo e j-ésimo objetos de Ω.

2. Função de *kernel* Gaussiana definida por duas componente:

$$\mathcal{K}^{(g)}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) = exp(-\frac{||\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I||^2}{2\sigma^2}) + exp(-\frac{||\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S||^2}{2\sigma^2}) \tag{3.16}$$

   em que $||\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I||^2 = \sum_{j-1}^{p}(x_i^I - x_j^I)^2$ e $||\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S||^2 = \sum_{j-1}^{p}(x_i^S - x_j^S)^2$. Considere também $\underset{\sim}{x}_i^I = (x_{i_1}^I, ..., x_{i_p}^I)^T$ e $\underset{\sim}{x}_i^S = (x_{i_1}^S, ..., x_{i_p}^S)^T$ como vetores de p-dimensões associados, respectivamente, aos limites inferior e superior dos intervalos que

descrevem o i-ésimo objeto de Ω, e $\underset{\sim}{x}_j^I = (x_{j_1}^I, ..., x_{j_p}^I)^T$ e $\underset{\sim}{x}_j^S = (x_{j_1}^S, ..., x_{j_p}^S)^T$ como vetores de p-dimensões que descrevem o j-ésimo objeto de Ω.

3.6 Método de Agrupamento Kernel K-médias Baseado em Kernelização da Métrica

Seja $\Omega = \{1, ..., n\}$ um conjunto de n padrões indexado por i e descrito por p variáveis. Seja $P = \{P_1, ..., P_K\}$ uma partição de Ω em K grupos. O método de agrupamento *kernel* K-médias baseado em kernelização da métrica (rotulado nesta dissertação como KKM), uma extensão do método de agrupamento K-médias, também busca formar grupos de modo que a seguinte função objetivo $J(\Phi(P))$ seja minimizada:

$$J(\Phi(P)) = \sum_{k=1}^{K} \sum_{i \in P_k} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{v}_k)||^2 \tag{3.17}$$

em que $\underset{\sim}{v}_k \in \mathbb{R}^p$ é o centróide do k-ésimo grupo ($k = 1, ..., K$).

O quadrado da distância euclidiana no espaço de características pode ser obtido da seguinte maneira:

$$\begin{aligned} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{v}_k)||^2 &= (\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{v}_k))^T (\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{v}_k)) \\ &= \Phi(\underset{\sim}{x}_i)^T \Phi(\underset{\sim}{x}_i) - 2\Phi(\underset{\sim}{x}_i)^T \Phi(\underset{\sim}{v}_k) + \Phi(\underset{\sim}{v}_k)^T \Phi(\underset{\sim}{v}_k) \\ &= \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_i) - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{v}_k) + \mathcal{K}(\underset{\sim}{v}_k, \underset{\sim}{v}_k) \end{aligned} \tag{3.18}$$

A derivação dos centróides dos grupos depende da escolha da função *kernel*. Considerando o *kernel* gaussiano, a função objetivo $J(\Phi(P))$ dada pela equação (3.17), pode ser expressa por:

$$\begin{aligned} J(\Phi(P)) &= \sum_{k=1}^{K} \sum_{i \in P_k} (1 - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{v}_k) + 1) \\ &= 2 \sum_{k=1}^{K} \sum_{i \in P_k} (1 - \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{v}_k)) \end{aligned} \tag{3.19}$$

Igualando a primeira derivada da equação (3.19) com relação ao centróide do k-ésimo grupo $\underset{\sim}{v}_k$ ao vetor nulo, obtemos a equação de atualização dos centróides dos grupos expressa da seguinte forma:

$$v_k = \frac{\sum_{i \in P_k} \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{v}_k) \underset{\sim}{x}_i}{\sum_{i \in P_k} \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{v}_k)}, k = 1, ..., K \tag{3.20}$$

Na definição da melhor partição do conjunto de padrões de entrada Ω, os centróides dos grupos $\underset{\sim}{v}_k$ $(k = 1, ..., K)$ estão fixos. Os grupos P_k $(k = 1, ..., K)$, que minimizam a função objetivo $J(\Phi(P))$ dada são então atualizados de acordo com a seguinte regra de alocação:

$$P_k = \{i \in \Omega : ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{v}_k)||^2 \leq ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{v}_h)||^2, \forall h \neq k, h = 1, ..., K\} \quad (3.21)$$

3.7 Método de Agrupamento Kernel K-médias Baseado em Kernelização da Métrica para Dados Simbólicos de Natureza Intervalar

Considere um conjunto de dados, descrito por $X = \{\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n\}$, $\Omega = \{1, ..., n\}$ e $\underset{\sim}{x}_i = ([a_i^1, b_i^1], ..., [a_i^p, b_i^p])$ $\forall i \in \Omega$, e $\Phi : \underset{\sim}{x} \to \Phi(\underset{\sim}{x})$ uma função linear que realiza um mapeamento do espaço de entrada para um espaço de características de alta dimensão $\mathcal{F}$. Considere também K centros no espaço de entradas do tipo intervalo $G = \{\underset{\sim}{g}_1, ..., \underset{\sim}{g}_K\}$ como os centróides dos grupos onde $\underset{\sim}{g}_k = ([\alpha_k^1, \beta_k^1], ..., [\alpha_k^p, \beta_k^p])$, $\forall k = 1, ..., K$. O método de agrupamento *kernel* K-médias baseado em kernelização da métrica para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKKM) consiste em encontrar uma partição de Ω em K grupos disjuntos $P^* = \{P_1, ..., P_K\}$ de modo que a seguinte função objetivo $J(\Phi(P^*, G))$ seja minimizada:

$$J(\Phi(P^*, G)) = \sum_{k=1}^{K} \sum_{i \in P_k^*} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k)||^2 \quad (3.22)$$

em que $\underset{\sim}{g}_k \in \mathbb{R}^p$ x $\mathbb{R}^p$ é o centróide do k-ésimo grupo $(k = 1, ..., K)$.

O quadrado da distância euclidiana no espaço de características pode ser obtido da seguinte maneira:

$$\begin{aligned} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k)||^2 &= (\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k))^T (\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k)) \\ &= \Phi(\underset{\sim}{x}_i)^T \Phi(\underset{\sim}{x}_i) - 2\Phi(\underset{\sim}{x}_i)^T \Phi(\underset{\sim}{g}_k) + \Phi(\underset{\sim}{g}_k)^T \Phi(\underset{\sim}{g}_k) \\ &= \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_i) - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{g}_k) + \mathcal{K}(\underset{\sim}{g}_k, \underset{\sim}{g}_k) \end{aligned} \quad (3.23)$$

Considerando o *kernel* gaussiano, a função objetivo $J(\Phi(P^*, G))$ dada pela equação (3.22), pode ser expressa por:

$$J(\Phi(P^*, G)) = \sum_{k=1}^{K} \sum_{i \in P_k^*} (1 - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{g}_k) + 1)$$

$$= 2\sum_{k=1}^{K}\sum_{i\in P_k^*}(1-\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{g}_k)) \tag{3.24}$$

em que $\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{g}_k) = exp(-\frac{||\underset{\sim}{x}_i-\underset{\sim}{g}_k||^2}{2\sigma^2})$, sendo $||\underset{\sim}{x}_i-\underset{\sim}{g}_k||^2 = \sum_{k=1}^{K}(a_i-\alpha_k)^2+\sum_{k=1}^{K}(b_i-\beta_k)^2$.

Para este caso, igualando a primeira derivada da equação (3.24) com relação ao centróide do k-ésimo grupo $\underset{\sim}{v}_k$ ao vetor nulo, obtemos a equação de atualização dos centróides dos grupos expressa da seguinte forma:

$$\underset{\sim}{g}_k = \frac{\sum_{i\in P_k^*}\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{g}_k)x_i}{\sum_{i\in P_k^*}\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{g}_k)} \tag{3.25}$$

Na definição da melhor partição do conjunto de padrões de entrada, os centróides dos grupos $\underset{\sim}{g}_k$ $(k=1,...,K)$ estão fixos. Os grupos P_k^* $(k=1,...,K)$, que minimizam a função objetivo $J(\Phi(P^*,G))$ dada são então atualizados de acordo com a seguinte regra de alocação:

$$P_k^* = \{i\in\Omega : ||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{g}_k)||^2 \leq ||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{g}_m)||^2, \forall m\neq k, m=1,...,K\} \tag{3.26}$$

3.8 Método de Agrupamento Subtrativo Baseado em Kernel

Kim et al. (2005) propôs um método de agrupamento subtrativo baseado em *kernel* (rotulado nesta dissertação como KSBC). Dado um ponto de dados $\underset{\sim}{x}_i \in \mathbb{R}^p$ $(1\leq i\leq n)$ e um mapeamento não-linear $\Phi : \mathbb{R}^p \to \mathcal{F}$, a função potencial em um ponto de dados $\underset{\sim}{x}_i$ é definida como:

$$M(\Phi(\underset{\sim}{x}_i)) = \sum_{j=1}^{n} e^{-\alpha(||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_j)||^2)} \tag{3.27}$$

em que α é uma constante positiva e $||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_j)||^2$ é o quadrado da distância Euclidiana entre $\Phi(\underset{\sim}{x}_i)$ e $\Phi(\underset{\sim}{x}_j)$. A distância no espaço de características é calculada através do *kernel* no espaço de entrada da seguinte maneira:

$$\begin{aligned}||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_j)||^2 &= (\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_j))^T(\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_j))\\ &= \Phi(\underset{\sim}{x}_i)^T\Phi(\underset{\sim}{x}_i)-2\Phi(\underset{\sim}{x}_i)^T\Phi(\underset{\sim}{x}_j)+\Phi(\underset{\sim}{x}_j)^T\Phi(\underset{\sim}{x}_j)\\ &= \mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_i)-2\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_j)+\mathcal{K}(\underset{\sim}{x}_j,\underset{\sim}{x}_j)\end{aligned} \tag{3.28}$$

e considerando o *kernel* gaussiano:

$$||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{x}_j)||^2 = 1 - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) + 1$$

$$= 2 - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) \tag{3.29}$$

Portanto, a função potencial $M(\Phi(\underset{\sim}{x}_i))$ dada pela equação (3.27) pode ser reescrita como:

$$M(\Phi(\underset{\sim}{x}_i)) = \sum_{j=1}^{n} e^{-\alpha(\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_i)-2\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_j)+\mathcal{K}(\underset{\sim}{x}_j,\underset{\sim}{x}_j))}$$

$$= \sum_{j=1}^{n} e^{-\alpha(2-2\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_j))} \tag{3.30}$$

O procedimento de seleção do centróide do grupo é semelhante ao método subtrativo. Depois de calcular a função potencial, o ponto de dados $\underset{\sim}{x}_i^*$ cujo valor potencial é $M_1^* = \max_{1\leq i\leq n}[M(\Phi(\underset{\sim}{x}_i))]$, é selecionado como o primeiro centróide do grupo.

A função potencial para encontrar os centróides subseqüentes é modificada da seguinte maneira:

$$\widehat{M}^j(\Phi(\underset{\sim}{x}_i)) = \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i)) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_{j-1}^*)||^2}$$

$$= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i)) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta(\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_i)-2\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_{j-1}^*)+\mathcal{K}(\underset{\sim}{x}_{j-1}^*,\underset{\sim}{x}_{j-1}^*))} \tag{3.31}$$

em que β é uma constante positiva e $\underset{\sim}{x}_{j-1}^*$ é o centróide recém-encontrado.

Capítulo 4

Métodos Propostos

4.1 Método de Agrupamento Subtrativo para Dados Simbólicos de Natureza Intervalar

No método subtrativo para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iSBC) consideramos cada ponto de dados como um potencial centróide do grupo e definimos uma função potencial do ponto de dados $\underset{\sim}{x}_i = (\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S) \in \mathbb{R}^p \mathrm{x} \mathbb{R}^p$ $(1 \leq i \leq n)$, como:

$$M(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S) = \sum_{j=1}^{n} e^{-\alpha(||\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I||^2 + ||\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S||^2)} \tag{4.1}$$

em que $\alpha = \dfrac{4}{r_a^2}$, sendo $r_a = 0.5$ e $||\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I||^2 + ||\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S||^2$ é o quadrado da distância Euclidiana entre os limites inferiores e superiores de $\underset{\sim}{x}_i$ e $\underset{\sim}{x}_j$, denotada por:

$$\begin{aligned} ||\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I|| + ||\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S||^2 &= (\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I)^T(\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I) + (\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S)^T(\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S) \\ &= \sum_{k=1}^{p}(\underset{\sim}{x}_{ik}^I - \underset{\sim}{x}_{jk}^I)^2 + \sum_{k=1}^{p}(\underset{\sim}{x}_{ik}^S - \underset{\sim}{x}_{jk}^S)^2 \end{aligned} \tag{4.2}$$

Usando a função potencial da equação (4.1), os centróides dos grupos são selecionados de forma semelhante à usada no método subtrativo original. Seja M_1^* o valor máximo da função potencial

$$M_1^* = \max_{1 \leq i \leq n}[M(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S)] \tag{4.3}$$

e $(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S)$ é o ponto de dados cujo valor potencial é M_1^*, este ponto de dados é selecionado como o primeiro centróide do grupo.

A função potencial modificada usada para encontrar os centróides dos grupos subseqüentes é definida como:

$$\widehat{M}^j(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S) = \widehat{M}^{j-1}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta(||\underset{\sim}{x}_i^I - \underset{\sim}{x}_{j-1}^{*I}||^2 + ||\underset{\sim}{x}_i^S - \underset{\sim}{x}_{j-1}^{*S}||^2)} \tag{4.4}$$

em que $\beta = \dfrac{4}{r_b^2}$, sendo r_b uma constante positiva e $(\underset{\sim}{x}_{j-1}^{*I}, \underset{\sim}{x}_{j-1}^{*S})$ é o centróide recém-encontrado. Para evitar a obtenção de centróides bem espaçados, assim como Chiu (1994), estabelecemos $r_b = 1.5r_a$.

O esquema do algoritmo iSBC é ilustrado a seguir:

1. Dado o número de grupos, δ, inicializar os parâmetros α, β;
2. Calcular a função potencial

$$M(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S) = \sum_{j=1}^{n} e^{-\alpha(||\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I||^2 + ||\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S||^2)};$$

3. Escolher o ponto de dados $\underset{\sim}{x}_i = (\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S)$ cuja função potencial $M(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S)$ é mais alta como centróide do grupo;
4. Modificar e recalcular a função potencial

$$\widehat{M}^j(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S) = \widehat{M}^{j-1}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta(||\underset{\sim}{x}_i^I - \underset{\sim}{x}_{j-1}^{*I}||^2 + ||\underset{\sim}{x}_i^S - \underset{\sim}{x}_{j-1}^{*S}||^2)};$$

5. Se o número de centróides encontrado for igual a δ, então pare; caso contrário voltar para o passo 3.

4.2 Método de Agrupamento Subtrativo Baseado em Kernel Definido por Uma Componente para Dados Simbólicos de Natureza Intervalar

No método subtrativo baseado em kernel definido por uma componente para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKSBC1C), dado um ponto de dados $\underset{\sim}{x}_i = (\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S) \in \mathbb{R}^p \mathrm{x} \mathbb{R}^p$ $(1 \leq i \leq n)$ e um mapeamento não-linear $\Phi : \mathbb{R}^p \mathrm{x} \mathbb{R}^p \rightarrow \mathcal{F}$, a função potencial em um ponto de dados $\underset{\sim}{x}_i$ é definida como:

$$M(\Phi(\underset{\sim}{x}_i)) = \sum_{j=1}^{n} e^{-\alpha(||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{x}_j)||^2)} \tag{4.5}$$

em que $\alpha = \frac{4}{r_a^2}$, sendo $r_a = 0.5$ e $||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{x}_j)||^2$ o quadrado da distância Euclidiana entre $\underset{\sim}{x}_i$ e $\underset{\sim}{x}_j$ em $\mathcal{F}$, considerando $\underset{\sim}{x}_i = (x_{i_1}^I, x_{i_1}^S, ..., x_{i_p}^I, x_{i_p}^S)^T$ e $\underset{\sim}{x}_j = (x_{j_1}^I, x_{j_1}^S, ..., x_{j_p}^I, x_{j_p}^S)^T$ como vetores de $2p$-dimensões que descrevem, respectivamente, o i-ésimo e j-ésimo objetos de Ω.

A distância no espaço de características é calculada através do *kernel* no espaço de entrada da seguinte maneira:

$$\begin{aligned} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{x}_j)||^2 &= (\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{x}_j))^T(\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{x}_j)) \\ &= \Phi(\underset{\sim}{x}_i)^T\Phi(\underset{\sim}{x}_i) - 2\Phi(\underset{\sim}{x}_i)^T\Phi(\underset{\sim}{x}_j) + \Phi(\underset{\sim}{x}_j)^T\Phi(\underset{\sim}{x}_j) \\ &= \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_i) - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) + \mathcal{K}(\underset{\sim}{x}_j, \underset{\sim}{x}_j) \end{aligned} \tag{4.6}$$

e considerando o *kernel* gaussiano:

$$\begin{aligned} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{x}_j)||^2 &= 1 - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) + 1 \\ &= 2 - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) \end{aligned} \tag{4.7}$$

em que $\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) = exp(-\frac{||\underset{\sim}{x}_j - \underset{\sim}{x}_j||^2}{2\sigma^2})$, sendo $||\underset{\sim}{x}_i - \underset{\sim}{x}_j||^2 = \sum_{j=1}^n (\underset{\sim}{x}_i^I - \underset{\sim}{x}_j^I)^2 + \sum_{j=1}^n (\underset{\sim}{x}_i^S - \underset{\sim}{x}_j^S)^2$.

Portanto, a função potencial $M(\Phi(\underset{\sim}{x}_i))$ pode ser reescrita como:

$$\begin{aligned} M(\Phi(\underset{\sim}{x}_i)) &= \sum_{j=1}^{n} e^{-\alpha(\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_i) - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_j) + \mathcal{K}(\underset{\sim}{x}_j, \underset{\sim}{x}_j))} \\ &= \sum_{j=1}^{n} e^{-\alpha(2 - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_j))} \end{aligned} \tag{4.8}$$

O procedimento de seleção do centróide do grupo é semelhante ao método subtrativo. Depois de calcular a função potencial, o ponto de dados $\underset{\sim}{x}_i^*$ cujo valor potencial é $M_1^* = \max_{1 \leq i \leq n}[M(\Phi(\underset{\sim}{x}_i))]$, é selecionado como o primeiro centróide do grupo.

A função potencial para encontrar os centróides subseqüentes é modificada da seguinte maneira:

$$\begin{aligned} \widehat{M}^j(\Phi(\underset{\sim}{x}_i)) &= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i)) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{x}_{j-1}^*)||^2} \\ &= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i)) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta(\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_i) - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_{j-1}^*) + \mathcal{K}(\underset{\sim}{x}_{j-1}^*, \underset{\sim}{x}_{j-1}^*))} \end{aligned} \tag{4.9}$$

em que $\beta = \dfrac{4}{r_b^2}$, sendo r_b uma constante positiva e $\underset{\sim}{x}^*_{j-1}$ é o centróide recém-encontrado. Para evitar a obtenção de centróides bem espaçados, assim como Chiu (1994), estabelecemos $r_b = 1.5r_a$.

O esquema do algoritmo iKSBC1C é ilustrado a seguir:

1. Dado o número de grupos, δ, e os valores escolhidos de α, β, escolher uma função *kernel* $\mathcal{K}$;

2. Calcular a função potencial

$$M(\Phi(\underset{\sim}{x}_i)) = \sum_{j=1}^{n} e^{-\alpha(||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_j)||^2)}$$

   em que $||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_j)||^2$ é o quadrado da distância Euclidiana entre $\underset{\sim}{x}_i$ e $\underset{\sim}{x}_j$ em $\mathcal{F}$, considerando $\underset{\sim}{x}_i = (x^I_{i_1}, x^S_{i_1}, ..., x^I_{i_p}, x^S_{i_p})^T$ e $\underset{\sim}{x}_j = (x^I_{j_1}, x^S_{j_1}, ..., x^I_{j_p}, x^S_{j_p})^T$ como vetores de $2p$-dimensões que descrevem, respectivamente, o i-ésimo e j-ésimo objetos de Ω;

3. Escolher o ponto de dados $\underset{\sim}{x}_i = (\underset{\sim}{x}^I_i, \underset{\sim}{x}^S_i)$ cuja função potencial $M(\Phi(\underset{\sim}{x}_i))$ é mais alta como centróide do grupo;

4. Modificar e recalcular a função potencial

$$\widehat{M}^j(\Phi(\underset{\sim}{x}_i)) = \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i)) - M^*_{j-1} \sum_{j=1}^{n} e^{-\beta||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}^*_{j-1})||^2}$$

$$= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i)) - M^*_{j-1} \textstyle\sum_{j=1}^{n} e^{-\beta(\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_i)-2\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}^*_{j-1})+\mathcal{K}(\underset{\sim}{x}^*_{j-1},\underset{\sim}{x}^*_{j-1}))};$$

5. Se o número de centróides encontrado for igual a δ, então pare; caso contrário voltar para o passo 3.

4.3 Método de Agrupamento Subtrativo Baseado em Kernel Definido por Duas Componentes para Dados Simbólicos de Natureza Intervalar

No método subtrativo baseado em kernel definido por duas componentes para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKSBC2C), dado um ponto de dados $\underset{\sim}{x}_i = (\underset{\sim}{x}^I_i, \underset{\sim}{x}^S_i) \in \mathbb{R}^p \mathrm{x} \mathbb{R}^p$ $(1 \leq i \leq n)$ e um mapeamento não-linear $\Phi : \mathbb{R}^p \mathrm{x} \mathbb{R}^p \rightarrow \mathcal{F}$, a função potencial em um ponto de dados $(\Phi(\underset{\sim}{x}^I_i), \Phi(\underset{\sim}{x}^S_i))$ é definida como:

$$M(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) = \sum_{j=1}^{n} e^{-\alpha(||\Phi(\underset{\sim}{x}_i^I)-\Phi(\underset{\sim}{x}_j^I)||^2+||\Phi(\underset{\sim}{x}_i^S)-\Phi(\underset{\sim}{x}_j^S)||^2)} \tag{4.10}$$

em que $\alpha = \frac{4}{r_a^2}$, sendo $r_a = 0.5$ e $||\Phi(\underset{\sim}{x}_i^I) - \Phi(\underset{\sim}{x}_j^I)||^2 + ||\Phi(\underset{\sim}{x}_i^S) - \Phi(\underset{\sim}{x}_j^S)||^2$ o quadrado da distância Euclidiana entre os limites inferiores e superiores de $\Phi(\underset{\sim}{x}_i)$ e $\Phi(\underset{\sim}{x}_j)$. A distância no espaço de características é calculada através do *kernel* no espaço de entrada da seguinte maneira:

$$\begin{aligned}
||\Phi(\underset{\sim}{x}_i^I) - \Phi(\underset{\sim}{x}_j^I)||^2 + ||\Phi(\underset{\sim}{x}_i^S) - \Phi(\underset{\sim}{x}_j^S)||^2 &= (\Phi(\underset{\sim}{x}_i^I) - \Phi(\underset{\sim}{x}_j^I))^T(\Phi(\underset{\sim}{x}_i^I) - \Phi(\underset{\sim}{x}_j^I))+ \\
&\quad (\Phi(\underset{\sim}{x}_i^S) - \Phi(\underset{\sim}{x}_j^S))^T(\Phi(\underset{\sim}{x}_i^S) - \Phi(\underset{\sim}{x}_j^S)) \\
&= \Phi(\underset{\sim}{x}_i^I)^T\Phi(\underset{\sim}{x}_i^I) - 2\Phi(\underset{\sim}{x}_i^I)^T\Phi(\underset{\sim}{x}_j^I)+ \\
&\quad \Phi(\underset{\sim}{x}_j^I)^T\Phi(\underset{\sim}{x}_j^I) + \Phi(\underset{\sim}{x}_i^S)^T\Phi(\underset{\sim}{x}_i^S)- \\
&\quad 2\Phi(\underset{\sim}{x}_i^S)^T\Phi(\underset{\sim}{x}_j^S) + \Phi(\underset{\sim}{x}_j^S)^T\Phi(\underset{\sim}{x}_j^S) \\
&= \mathcal{K}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^I) - 2\mathcal{K}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_j^I) + \mathcal{K}(\underset{\sim}{x}_j^I, \underset{\sim}{x}_j^I)+ \\
&\quad \mathcal{K}(\underset{\sim}{x}_i^S, \underset{\sim}{x}_i^S) - 2\mathcal{K}(\underset{\sim}{x}_i^S, \underset{\sim}{x}_j^S) + \mathcal{K}(\underset{\sim}{x}_j^S, \underset{\sim}{x}_j^S)
\end{aligned} \tag{4.11}$$

e considerando o *kernel* gaussiano:

$$\begin{aligned}
||\Phi(\underset{\sim}{x}_i^I) - \Phi(\underset{\sim}{x}_j^I)||^2 + ||\Phi(\underset{\sim}{x}_i^S) - \Phi(\underset{\sim}{x}_j^S)||^2 &= 1 - 2\mathcal{K}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_j^I) + 1 + 1 - 2\mathcal{K}(\underset{\sim}{x}_i^S, \underset{\sim}{x}_j^S) + 1 \\
&= 2 - 2\mathcal{K}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_j^I) + 2 - 2\mathcal{K}(\underset{\sim}{x}_i^S, \underset{\sim}{x}_j^S) \\
&= 2[2 - \mathcal{K}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_j^I) - \mathcal{K}(\underset{\sim}{x}_i^S, \underset{\sim}{x}_j^S)]
\end{aligned} \tag{4.12}$$

Portanto, a função potencial $M(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S))$ pode ser reescrita como:

$$\begin{aligned}
M(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) &= \sum_{j=1}^{n} e^{-\alpha(\mathcal{K}(\underset{\sim}{x}_i^I,\underset{\sim}{x}_i^I)-2\mathcal{K}(\underset{\sim}{x}_i^I,\underset{\sim}{x}_j^I)+\mathcal{K}(\underset{\sim}{x}_j^I,\underset{\sim}{x}_j^I)+\mathcal{K}(\underset{\sim}{x}_i^S,\underset{\sim}{x}_i^S)-2\mathcal{K}(\underset{\sim}{x}_i^S,\underset{\sim}{x}_j^S)+\mathcal{K}(\underset{\sim}{x}_j^S,\underset{\sim}{x}_j^S))} \\
&= \sum_{j=1}^{n} e^{-\alpha(2[2-\mathcal{K}(\underset{\sim}{x}_i^I,\underset{\sim}{x}_j^I)-\mathcal{K}(\underset{\sim}{x}_i^S,\underset{\sim}{x}_j^S)])}
\end{aligned} \tag{4.13}$$

O procedimento de seleção de centróide do grupo é semelhante ao método subtrativo baseado em *kernel*. Depois de calcular a função potencial, o ponto de dados $(\underset{\sim}{x}_i^{*I}, \underset{\sim}{x}_i^{*S})$ cujo valor potencial é $M_1^* = \max_{1 \leq i \leq n}[M(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S))]$, é selecionado como o primeiro centróide do grupo.

A função potencial para encontrar os centróides subsequentes é modificada da seguinte maneira:

$$\begin{aligned}
\widehat{M}^j(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) &= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta(||\Phi(\underset{\sim}{x}_i^I) - \Phi(\underset{\sim}{x}_{j-1}^{*I})||^2 + ||\Phi(\underset{\sim}{x}_i^S) - \Phi(\underset{\sim}{x}_{j-1}^{*S})||^2)} \\
&= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) - \\
&\quad M_{j-1}^* \textstyle\sum_{j=1}^{n} e^{-\beta(\mathcal{K}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^I) - 2\mathcal{K}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_{j-1}^{*I}) + \mathcal{K}(\underset{\sim}{x}_{j-1}^{*I}, \underset{\sim}{x}_{j-1}^{*I}) + \mathcal{K}(\underset{\sim}{x}_i^S, \underset{\sim}{x}_i^S) - 2\mathcal{K}(\underset{\sim}{x}_i^S, \underset{\sim}{x}_{j-1}^{*S}) + \mathcal{K}(\underset{\sim}{x}_{j-1}^{*S}, \underset{\sim}{x}_{j-1}^{*S}))} \\
&= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta(2[2 - \mathcal{K}(\underset{\sim}{x}_i^I, \underset{\sim}{x}_{j-1}^{*I}) - \mathcal{K}(\underset{\sim}{x}_i^S, \underset{\sim}{x}_{j-1}^{*S})])}
\end{aligned} \tag{4.14}$$

em que $\beta = \dfrac{4}{r_b^2}$, sendo r_b uma constante positiva e $(\underset{\sim}{x}_{j-1}^{*I}, \underset{\sim}{x}_{j-1}^{*S})$ é o centróide recém-encontrado. Para evitar a obtenção de centróides bem espaçados, assim como Chiu (1994), estabelecemos $r_b = 1.5 r_a$.

O esquema do algoritmo iKSBC2C é ilustrado a seguir:

1. Dado o número de grupos, δ, e os valores escolhidos de α, β, escolher uma função *kernel* $\mathcal{K}$;

2. Calcular a função potencial

$$M(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) = \sum_{j=1}^{n} e^{-\alpha(||\Phi(\underset{\sim}{x}_i^I) - \Phi(\underset{\sim}{x}_j^I)||^2 + ||\Phi(\underset{\sim}{x}_i^S) - \Phi(\underset{\sim}{x}_j^S)||^2)};$$

3. Escolher o ponto de dados $\underset{\sim}{x}_i = (\underset{\sim}{x}_i^I, \underset{\sim}{x}_i^S)$ cuja função potencial $M(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S))$ é mais alta como centróide do grupo;

4. Modificar e recalcular a função potencial

$$\begin{aligned}
\widehat{M}^j(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) &= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta(||\Phi(\underset{\sim}{x}_i^I) - \Phi(\underset{\sim}{x}_{j-1}^{*I})||^2 + ||\Phi(\underset{\sim}{x}_i^S) - \Phi(\underset{\sim}{x}_{j-1}^{*S})||^2)} \\
&= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i^I), \Phi(\underset{\sim}{x}_i^S)) -
\end{aligned}$$

$$M^*_{j-1}\sum_{j=1}^{n} e^{-\beta(\mathcal{K}(\underset{\sim}{x}_i^I,\underset{\sim}{x}_i^I)-2\mathcal{K}(\underset{\sim}{x}_i^I,\underset{\sim}{x}^{*I}_{j-1})+\mathcal{K}(\underset{\sim}{x}^{*I}_{j-1},\underset{\sim}{x}^{*I}_{j-1})+\mathcal{K}(\underset{\sim}{x}_i^S,\underset{\sim}{x}_i^S)-2\mathcal{K}(\underset{\sim}{x}_i^S,\underset{\sim}{x}^{*S}_{j-1})+\mathcal{K}(\underset{\sim}{x}^{*S}_{j-1},\underset{\sim}{x}^{*S}_{j-1}))}$$

$$= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i^I),\Phi(\underset{\sim}{x}_i^S)) - M^*_{j-1}\sum_{j=1}^{n} e^{-\beta(2[2-\mathcal{K}(\underset{\sim}{x}_i^I,\underset{\sim}{x}^{*I}_{j-1})-\mathcal{K}(\underset{\sim}{x}_i^S,\underset{\sim}{x}^{*S}_{j-1})])}$$

5. Se o número de centróides encontrado for igual a δ, então pare; caso contrário voltar para o passo 3.

4.4 Métodos de Agrupamento K-médias para Dados Simbólicos de Natureza Intervalar (Utilizando os Centróides dos Métodos de Agrupamento iSBC, iKSBC1C e iKSBC2C como entradas)

Considere um conjunto de dados, descrito por $X = \{\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n\}$, $\Omega = \{1, ..., n\}$ e $\underset{\sim}{x}_i = ([a_i^1, b_i^1], ..., [a_i^p, b_i^p])$ $\forall i \in \Omega$, particionado em K grupos no espaço de entradas do tipo intervalo e $G = \{\underset{\sim}{g}_1, ..., \underset{\sim}{g}_K\}$ como os centróides dos grupos onde $\underset{\sim}{g}_k = ([\alpha_k^1, \beta_k^1], ..., [\alpha_k^p, \beta_k^p])$, $\forall k = 1, ..., K$. O método de agrupamento K-médias para dados simbólicos de natureza intervalar consiste em encontrar uma partição em K grupos disjuntos $P^* = \{P_1, ..., P_K\}$ de modo que a seguinte função objetivo $J(P^*, G)$ seja minimizada:

$$J(P^*, G) = \sum_{k=1}^{K}\sum_{i\in P_k^*} ||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2 \tag{4.15}$$

em que $\underset{\sim}{g}_k \in \mathbb{R}^p$ x $\mathbb{R}^p$ é o centróide do k-ésimo grupo ($k = 1, ..., K$).

Esta seção propõem três métodos de agrupamento K-médias para dados simbólicos de natureza intervalar utilizando como estratégia os centróides dos métodos de agrupamento subtrativo já propostos nesta dissertação como entradas:

1. Método de agrupamento K-médias para dados simbólicos de natureza intervalar utilizando os centróides do método de agrupamento subtrativo para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKM+iSBC);

2. Método de agrupamento K-médias para dados simbólicos de natureza intervalar utilizando os centróides do método de agrupamento subtrativo baseado em *kernel* definido por uma componente para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKM+iKSBC1C);

3. Método de agrupamento K-médias para dados simbólicos de natureza intervalar utilizando os centróides do método de agrupamento subtrativo baseado em *kernel* definido por duas componentes para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKM+iKSBC2C).

O quadrado da distância Euclidiana entre $\underset{\sim}{x}_i$ e $\underset{\sim}{g}_k$ pode ser obtido da seguinte maneira:

$$||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2 = (\underset{\sim}{x}_i - \underset{\sim}{g}_k)^T(\underset{\sim}{x}_i - \underset{\sim}{g}_k)$$

$$= (\underset{\sim}{x}_i)^T(\underset{\sim}{x}_i) - 2(\underset{\sim}{x}_i)^T(\underset{\sim}{g}_k) + (\underset{\sim}{g}_k)^T(\underset{\sim}{g}_k) \tag{4.16}$$

em que $||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2 = \sum_{k=1}^{K}(a_i - \alpha_k)^2 + \sum_{k=1}^{K}(b_i - \beta_k)^2$.

Na definição da melhor partição do conjunto de padrões de entrada, os centróides dos grupos $\underset{\sim}{g}_k$ $(k = 1, ..., K)$ estão fixos. Os grupos P_k^* $(k = 1, ..., K)$, que minimizam a função objetivo $J(P^*, G)$ dada são então atualizados de acordo com a seguinte regra de alocação:

$$P_k^* = \{i \in \Omega : ||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2 \leq ||\underset{\sim}{x}_i - \underset{\sim}{g}_s||^2, \forall s \neq k, s = 1, ..., K\} \tag{4.17}$$

4.5 Métodos de Agrupamento Kernel K-médias Baseados em Kernelização da Métrica para Dados Simbólicos de Natureza Intervalar (Utilizando os Centróides dos Métodos de Agrupamento iSBC, iKSBC1C e iKSBC2C como entradas)

Considere um conjunto de dados, descrito por $X = \{\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n\}$, $\Omega = \{1, ..., n\}$ e $\underset{\sim}{x}_i = ([a_i^1, b_i^1], ..., [a_i^p, b_i^p])$ $\forall i \in \Omega$, e $\Phi : \underset{\sim}{x} \to \Phi(\underset{\sim}{x})$ uma função linear que realiza um mapeamento do espaço de entrada para um espaço de características de alta dimensão $\mathcal{F}$. Considere também K centros no espaço de entradas do tipo intervalo $G = \{\underset{\sim}{g}_1, ..., \underset{\sim}{g}_K\}$ como os centróides dos grupos onde $\underset{\sim}{g}_k = ([\alpha_k^1, \beta_k^1], ..., [\alpha_k^p, \beta_k^p])$, $\forall k = 1, ..., K$. O método de agrupamento *kernel* K-médias baseado em kernelização da métrica para dados simbólicos de natureza intervalar consiste em encontrar uma partição de Ω em K grupos disjuntos $P^* = \{P_1, ..., P_K\}$ de modo que a seguinte função objetivo $J(\Phi(P^*, G))$ seja minimizada:

$$J(\Phi(P^*, G)) = \sum_{k=1}^{K} \sum_{i \in P_k^*} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k)||^2 \tag{4.18}$$

em que $\underset{\sim}{g}_k \in \mathbb{R}^p$ x $\mathbb{R}^p$ é o centróide do k-ésimo grupo ($k = 1, ..., K$).

Esta seção propõem três métodos de agrupamento *kernel* K-médias baseado em kernelização da métrica para dados simbólicos de natureza intervalar utilizando como estratégia os centróides dos métodos de agrupamento subtrativo já propostos nesta dissertação como entradas:

1. Método de agrupamento *kernel* K-médias baseado em kernelização da métrica para dados simbólicos de natureza intervalar utilizando os centróides do método de agrupamento subtrativo para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKKM+iSBC);

2. Método de agrupamento *kernel* K-médias baseado em kernelização da métrica para dados simbólicos de natureza intervalar utilizando os centróides do método de agrupamento subtrativo baseado em *kernel* definido por uma componente para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKKM+iKSBC1C);

3. Método de agrupamento *kernel* K-médias baseado em kernelização da métrica para dados simbólicos de natureza intervalar utilizando os centróides do método de agrupamento subtrativo baseado em *kernel* definido por duas componentes para dados simbólicos de natureza intervalar (rotulado nesta dissertação como iKKM+iKSBC2C).

O quadrado da distância euclidiana no espaço de características pode ser obtido da seguinte maneira:

$$\begin{aligned} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k)||^2 &= (\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k))^T (\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k)) \\ &= \Phi(\underset{\sim}{x}_i)^T \Phi(\underset{\sim}{x}_i)) - 2\Phi(\underset{\sim}{x}_i)^T \Phi(\underset{\sim}{g}_k) + \Phi(\underset{\sim}{g}_k)^T \Phi(\underset{\sim}{g}_k) \\ &= \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{x}_i) - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{g}_k) + \mathcal{K}(\underset{\sim}{g}_k, \underset{\sim}{g}_k) \end{aligned} \tag{4.19}$$

Considerando o *kernel* gaussiano, a função objetivo $J(\Phi(P^*, G))$ dada pela equação (4.19), pode ser expressa por:

$$J(\Phi(P^*, G)) = \sum_{k=1}^{K} \sum_{i \in P_k^*} (1 - 2\mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{g}_k) + 1)$$

$$= 2\sum_{k=1}^{K}\sum_{i\in P_k^*}(1-\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{g}_k)) \tag{4.20}$$

em que $\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{g}_k) = exp(-\frac{||\underset{\sim}{x}_i-\underset{\sim}{g}_k||^2}{2\sigma^2})$, sendo $||\underset{\sim}{x}_i-\underset{\sim}{g}_k||^2 = \sum_{k=1}^{K}(a_i-\alpha_k)^2+\sum_{k=1}^{K}(b_i-\beta_k)^2$.

Na definição da melhor partição do conjunto de padrões de entrada, os centróides dos grupos $\underset{\sim}{g}_k$ $(k=1,...,K)$ estão fixos. Os grupos P_k^* $(k=1,...,K)$, que minimizam a função objetivo $J(\Phi(P^*,G))$ dada são então atualizados de acordo com a seguinte regra de alocação:

$$P_k^* = \{i\in\Omega : ||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{g}_k)||^2 \leq ||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{g}_m)||^2, \forall m\neq k, m=1,...,K\} \tag{4.21}$$

Capítulo 5

Avaliação dos Métodos Propostos

Os métodos de agrupamento subtrativo para dados simbólicos de natureza intervalar (iSBC), subtrativo baseado em *kernel* definido por uma componente para dados simbólicos de natureza intervalar (iSBC1C), subtrativo baseado em *kernel* definido por duas componentes para dados simbólicos de natureza intervalar (iSBC2C), K-médias para dados simbólicos de natureza intervalar utilizando os centróides dos métodos subtrativos propostos (iKM+iSBC, iKM+iSBC1C e iKM+iSBC2C) e *kernel* K-médias para dados simbólicos de natureza intervalar utilizando os centróides dos métodos subtrativos propostos (iKKM+iSBC, iKKM+iSBC1C e iKKM+iSBC2C) serão implementados em linguagem *R*. Os métodos e estratégias propostas serão avaliados através de experimentos utilizando dados reais e serão comparados aos métodos K-médias para dados do tipo intervalo (iKM) e *kernel* K-médias para dados simbólicos do tipo intervalo (iKKM). Para comparar os métodos de agrupamento considerados nesta dissertação, utilizaremos o Índice de Rand Ajustado (IRA) Hubert e Arabie (1985) e a taxa total de erro de alocação (TEA) Breiman et al. (1984).

Tabela 5.1: Matriz de confusão.

	Grupos					
Classes	P_1	$\cdots$	P_k	$\cdots$	P_K	$\sum$
$\mathcal{P}_1$	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1K}	$n_{1\bullet} = \sum_{k=1}^{K} n_{1k}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\mathcal{P}_i$	n_{i1}	$\cdots$	n_{ik}	$\cdots$	n_{iK}	$n_{i\bullet} = \sum_{k=1}^{K} n_{ik}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\mathcal{P}_c$	n_{c1}	$\cdots$	n_{ck}	$\cdots$	n_{cK}	$n_{c\bullet} = \sum_{k=1}^{K} n_{ck}$
$\sum$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$\cdots$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$\cdots$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$n = \sum_{i=1}^{c} \sum_{k=1}^{K} n_{ik}$

Seja $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_i, \ldots, \mathcal{P}_c\}$ a partição *a priori* de $\Omega = \{1, \ldots, n\}$ em c classes e seja $P = \{P_1, \ldots, P_k, \ldots, P_K\}$ uma partição rígida de $\Omega = \{1, \ldots, n\}$ em K grupos

fornecidos por um algoritmo de agrupamento. As quantidades n_{ik}, $i = 1, \ldots, c$, $k = 1, \ldots, K$, representam o número de observações que estão na classe $\mathcal{P}_i$ e no grupo P_k e podem ser representadas na forma da Tabela 5.1, denominada matriz de confusão.

O Índice de Rand Ajustado (IRA) é obtido como

$$IRA = \frac{\sum_{i=1}^{c}\sum_{k=1}^{K}\binom{n_{ik}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{c}\binom{n_{i\bullet}}{2}\sum_{k=1}^{K}\binom{n_{\bullet k}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{c}\binom{n_{i\bullet}}{2} + \sum_{k=1}^{K}\binom{n_{\bullet k}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{c}\binom{n_{i\bullet}}{2}\sum_{k=1}^{K}\binom{n_{\bullet k}}{2}}, \tag{5.1}$$

onde $\binom{n}{2} = \frac{n(n-1)}{2}$, n_{ik} representa o número de observações que estão na classe $\mathcal{P}_i$ e no grupo P_k, $n_{i\bullet}$ representa o número de observações na classe $\mathcal{P}_i$, $n_{\bullet k}$ representa o número de observações no grupo P_k, e n é o número total de observações no conjunto de dados.

O IRA avalia o grau de concordância (similaridade) entre uma partição *a priori* e uma partição fornecida por um método de agrupamento. Além disso, o IRA não é sensível ao número de classes nas partições ou à distribuição das observações nos grupos. Finalmente, o IRA assume valores no intervalo $[-1, 1]$, no qual o valor 1 indica concordância perfeita entre as partições, enquanto que valores próximos de zero ou negativos correspondem a concordância entre partições encontrada ao acaso (MILLIGAN; COOPER, 1986).

Em problemas de classificação, cada grupo P_k é associado a uma classe *a priori* $\mathcal{P}_i$ e esta associação deve ser interpretada como se a verdadeira classe *a priori* fosse $\mathcal{P}_i$. Dessa forma, para uma observação pertencente a um dado grupo P_k a decisão está correta se a classe *a priori* dessa observação é $\mathcal{P}_i$. Para obter uma taxa de erro de classificação mínima, precisamos encontrar uma regra de decisão que minimize a probabilidade de erro.

Seja $\ell(\mathcal{P}_i, P_k)$ a probabilidade equiprovável *a posteriori* de que uma observação pertença à classe $\mathcal{P}_i$ quando associado ao grupo P_k. Seja $\ell(P_k)$ a probabilidade de que a observação pertença ao grupo P_k. A função ℓ é conhecida como função de verossimilhança.

A estimativa da máxima probabilidade *a posteriori* é a moda da probabilidade *a posteriori* $\ell(\mathcal{P}_i, P_k)$ e o índice da classe *a priori* associada a esta moda é dada por

$$MAP(P_k) = \arg\max_{1 \leq i \leq c} \ell(\mathcal{P}_i, P_k). \tag{5.2}$$

A regra de decisão de Bayes que minimiza a probabilidade média de erro é selecionar a classe *a priori* que maximiza a probabilidade *a posteriori*. A taxa de erro de alocação do grupo P_k é igual a $1 - \ell(\mathcal{P}_{MAP(P_k)}, P_k)$ e a taxa total de erro de

alocação (TEA) é igual a

$$TEA = \sum_{k=1}^{K} \ell(P_k)(1 - \ell(\mathcal{P}_{MAP(P_k)}, P_k)). \tag{5.3}$$

Para uma amostra,

$$\ell(\mathcal{P}_{MAP(P_k)}, P_k) = \max_{1 \leq i \leq c} \frac{n_{ik}}{n_{\bullet k}}. \tag{5.4}$$

A taxa total de erro de alocação (TEA) foi concebida de modo a medir a habilidade de um algoritmo de agrupamento encontrar as classes *a priori* presentes em um conjunto de dados e é calculada da forma:

$$TEA = \sum_{k=1}^{K} \frac{n_{\bullet k}}{n} \left(1 - \max_{1 \leq i \leq c} n_{ik}/n_{\bullet k}\right) = 1 - \frac{\sum_{k=1}^{K} \max_{1 \leq i \leq c} n_{ik}}{n}. \tag{5.5}$$

O índice TEA assume valores no intervalo $[0, 1]$, no qual valores próximos de zero indicam maior habilidade de um algoritmo na detecção de classes *a priori*.

Capítulo 6

Resultados

Nos estudos de métodos de agrupamento são realizados experimentos para tratar dados conhecidos como dados de aplicação ou dados reais (COSTA, 2011). Neste capítulo, apresentamos uma avaliação experimental dos métodos de agrupamento propostos no Capítulo 4. Para avaliar o desempenho dos métodos propostos foram realizados experimentos no software *R*, que classifica os conjuntos de dados reais do tipo intervalo e avalia os resultados da classificação usando o Índice de Rand Ajustado (IRA) e a taxa total de erro de alocação (TEA). A seguir serão descritos alguns dos conjuntos de dados reais utilizados e apresentados os resultados da execução dos métodos de agrupamento propostos.

6.1 Conjunto de dados: Carros

O conjunto de dados simbólicos de natureza intervalar *Carros* consiste de 33 modelos de carros descritos por 8 variáveis intervalares: Preço, Cilindrada do motor, Velocidade máxima, Aceleração, Passo, Comprimento, Largura e Altura. Este conjunto de dados está agrupado em quatro classes a priori de tamanhos diferentes: 8 modelos de carros são da categoria Ammiraglia, 8 modelos da categoria Berlina, 7 da categoria Sportiva e 10 são da categoria Utilitaria. Como pode ser visto no Apêndice H, que descreve os modelos de carros de acordo com suas variáveis e categorias.

A Tabela 6.1 ilustra os valores do cálculo do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) obtidos na comparação do conjunto de partições a priori e as partições obtidas como resultado da aplicação dos métodos de agrupamento propostos. Para este conjunto de dados o método iKSBC1C não teve desempenho melhor que o iSBC, de acordo com o IRA teve um diminuição no valor e a TEA teve um aumento, no entanto o iKSBC2C teve um resultado satisfatório se comparado aos resultados dos métodos iSBC e ao IKSBC1C. Em relação ao iKM, se comparado aos métodos IKM+SBC, IKM+iSBC1C e IKM+iSBC2C, é visto que os três métodos utilizando os centróides dos métodos subtrativos tiveram resultados

satisfatórios. Já para os métodos iKKM+iSBC e IKM+iSBC1C o IRA aumentou em relação ao IKKM e a TEA diminuiu, no entanto o IKM+iSBC2C obteve os mesmo resultados que o método IKKM.

Tabela 6.1: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Carros*

Métodos de agrupamento	IRA	TEA
iSBC	**0,3817**	**0,3636**
iKSBC1C	0,2819	0,3939
iKSBC2C	**0,4204**	**0,2727**
iKM	0,2329	0,4848
iKM+iSBC	**0,3541**	**0,3636**
iKM+iKSBC1C	**0,3541**	**0,3636**
iKM+iKSBC2C	**0,3541**	**0,3636**
iKKM	-0,0141	0,6667
iKKM+iSBC	-0,0029	0,6364
iKKM+iKSBC1C	-0,0029	0,6364
iKKM+iKSBC2C	-0,0141	0,6667

6.2 Conjunto de dados: Desenvolvimento dos países

O conjunto de dados simbólicos de natureza intervalar *Desenvolvimento dos países* consiste de 10 países descritos por 4 variáveis intervalares: população, espectativa de vida, PIB (Produto Interno Bruto) e IDH (Indíce de Desenvolvimento Humano). Este conjunto de dados está agrupado em duas classes a priori de tamanhos iguais: 5 países estão na categoria de país desenvolvido e 5 países estão na categoria de país em desenvolvimento. O Apêndice I descreve os países de acordo com suas variáveis e desenvolvimento.

A Tabela 6.2 mostra que, os métodos iKM+SBC, IKM+iSBC1C e IKM+iSBC2C propostos tiveram o mesmo desempenho que o método K-médias para dados simbólicos de natureza intervalar (IKM), apresentando um índice de Rand ajustado negativo e uma taxa total de erro de alocação relativamente alta. Os métodos iKKM+SBC, IKKM+iSBC1C e IKKM+iSBC2C também tiveram o mesmo desempenho que o método iKKM, mostrando que para este conjunto de dados os métodos propostos não tiveram desempenho satisfatório se comparados aos métodos os quais eles são extensão. No entanto os métodos iSBC, iKSBC1C e iKSBC2C obtiveram resultados satisfatórios se comparados ao método iKM e suas extensões, e ao método IKKM e suas extensões.

Tabela 6.2: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Desenvolvimento dos países*

Métodos de agrupamento	IRA	TEA
iSBC	**0,2800**	**0,2000**
iKSBC1C	**0,2800**	**0,2000**
iKSBC2C	**0,2800**	**0,2000**
iKM	-0,0766	0,5000
iKM+iSBC	-0,0766	0,5000
iKM+iKSBC1C	-0,0766	0,5000
iKM+iKSBC2C	-0,0766	0,5000
iKKM	0,0000	0,4000
iKKM+iSBC	0,0000	0,4000
iKKM+iKSBC1C	0,0000	0,4000
iKKM+iKSBC2C	0,0000	0,4000

6.3 Conjunto de dados: Facedata

O conjunto de dados simbólicos de natureza intervalar *facedata* (http://127.0.0.1:19650/library/RSDA/html/facedata.html) é um conjunto de dados intervalares usado como exemplo no pacote **RSDA** do software *R* e consiste de 6 variáveis intervalares agrupadas em oito classes, sendo cada classe de tamanho 3. O Apêndice J descreve os dados de acordo com suas variáveis e categoria.

Tabela 6.3: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Facedata*

Métodos de agrupamento	IRA	TEA
iSBC	**1,0000**	**0,0000**
iKSBC1C	**1,0000**	**0,0000**
iKSBC2C	**1,0000**	**0,0000**
iKM	0,7783	0,1250
iKM+iSBC	**1,0000**	**0,0000**
iKM+iKSBC1C	**1,0000**	**0,0000**
iKM+iKSBC2C	**1,0000**	**0,0000**
iKKM	0,7783	0,2500
iKKM+iSBC	**1,0000**	**0,0000**
iKKM+iKSBC1C	**1,0000**	**0,0000**
iKKM+iKSBC2C	**1,0000**	**0,0000**

Os resultados do IRA e da TEA após a aplicação dos algoritmos dos métodos de agrupamento para dados simbólicos de natureza intervalar propostos podem ser vizualizados na Tabela 6.3. Observando os resultados percebe-se que todos os métodos de agrupamento subtrativo propostos (iSBC, iSBC1C e iSBC2C)

apresentaram resultados totalmente satisfatórios. E os métodos K-médias e *kernel* K-médias, ambos utilizando os centróides dos métodos subtrativos propostos (iKM+iSBC, iKM+iKSBC1C, iKM+iKSBC2C, iKKM+iSBC, iKKM+iKSBC1C e iKKM+iKSBC2C) obtiveram resultados superiores aos métodos iKM e iKKM, dos quais eles são extensões. Mostrando assim que para classificação de padrões deste conjunto de dados os métodos subtrativos, K-médias e *kernel* K-médias, para dados do tipo intervalo, propostos são eficazes.

6.4 Conjunto de dados: Fluxos de água

Este conjunto de dados simbólicos de natureza intervalar *Fluxos de água* (*https://lhedjazi.jimdo.com/useful-links/*) consiste de registros de fluxos de água de 30 minutos coletados durante 1 ano na rede de distribuição de água da cidade de Barcelona na Espanha. Contém dados referente a 316 dias do ano, considerando que, em alguns dias, valores de medição foram descartados em função de problemas de funcionamento nos sensores que coletam as informações referente aos fluxos.

Os registros de fluxos de água são descritos por 48 variáveis intervalares que descrevem a variação mínima e máxima observada com base em 3 medições consecutivas de 10 minutos do fluxo de água. E os dados são distribuídos entre duas classes: Final de semana (sábados, domingos e feriados) e Dia útil, como pode ser visto no Apêndice K que descreve os dias de acordo com suas variáveis e classes. A caracterização de intervalos de tempo correspondentes a variação do consumo de água dos usuários, possibilita prever o fluxo a ser fornecido de acordo com cada tipo de dia da semana, possibilitando a empresa fornecedora de água uma certa flexibilidade para regular o fluxo de água de acordo com o período associado ao tipo de dia.

Como mencionado, o conjunto de dados originalmente possui 48 variáveis do tipo intervalo, no entanto neste trabalho foi retirada a variável IF17, por se tratar de uma concatenação das duas variáveis anteriores (IF15 e IF16).

A Tabela 6.4 apresenta os resultados obtidos após a aplicação dos algoritmos referente aos métodos de agrupamento para dados simbólicos de natureza intervalar propostos. É possível ver que os métodos baseados em *kernel* iKSBC1C e iKSBC2C obtiveram resultados superiores ao método subtrativo para dados simbólicos de natureza intervalar (iSBC). Tanto o método iKM quanto suas extensões apresentaram o mesmo desempenho. No entanto em relação as extensões do método IKKM é possível notar uma melhora no desempenho do IRA e apesar de apresentar valor baixo no iKKM+iSBC o resultado já é considerado bom e pode representar algum nível de reconhecimento nos padrões avaliados, pois segundo Costa (2011) este conjunto tem um tendência a apresentar uma taxa total de erro de alocação relativamente

Tabela 6.4: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Fluxos de água*

Métodos de agrupamento	IRA	TEA
iSBC	0,0205	0,2943
iKSBC1C	**0,0349**	**0,2943**
iKSBC2C	**0,0410**	**0,2943**
iKM	**0,0995**	**0,3323**
iKM+iSBC	**0,0995**	**0,3323**
iKM+iKSBC1C	**0,0995**	**0,3323**
iKM+iKSBC2C	**0,0995**	**0,3323**
iKKM	-0,0037	0,2943
iKKM+iSBC	-0,0113	0,2943
iKKM+iKSBC1C	**0,0857**	**0,2911**
iKKM+iKSBC2C	**0,1363**	**0,2690**

alta e no caso dos métodos propostos apresentou TEA relativamente baixa.

6.5 Conjunto de dados: Fórmula 1

O conjunto de dados simbólicos de natureza intervalar *Fórmula 1* consiste de 10 equipes da fórmula 1 (F1) descritos por 9 variáveis intervalares: Posição, Número de pontos, Ano de nascimento, Altura, Peso, Ano de ínicio na F1, Número de grandes prêmios (GP), Poles, Número de vitórias em 2005. Uma variável deste conjunto de dados, em relação aos títulos dos pilotos, está agrupada em duas classes a priori de tamanhos diferentes: 3 equipes conquistaram títulos e 7 não conquistaram títulos. E outra variável deste conjunto de dados, em relação aos pódios no ano de 2005, está agrupada em três classes a priori de tamanhos diferentes: 3 equipes subiram ao pódio em 1º lugar, 4 equipes não subiram ao pódio e 3 equipes subiram ao pódio, mas não em 1º lugar. O Apêndice L descreve as equipes de acordo com suas variáveis e conquistas de títulos e pódios dos pilotos.

Os resultados obtidos para este conjunto de dados em relação a variável título dos pilotos estão apresentados na Tabela 6.5. É possível notar que os métodos iSBC e iSBC2C obtiveram os mesmos resultados, já o método iSBC2C obteve um resultado negativo no IRA mas apresentou uma TEA relativamente baixa. Dentre os métodos de agrupamento para dados do tipo intervalo propostos, o método iKM+iKSBC2C, extensão do método iKM, foi o que apresentou melhor desempenho. Os métodos iKKM+iKSBC1C e iKKM+iKSBC2C apresentaram resultados iguais ao iKM, já o iKKM+iSBC obteve desempenho inferior ao método iKM e suas extensões mas assim como o iSBC2C obteve uma TEA relativamente baixa.

Tabela 6.5: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Fórmula 1* (Variável de agrupamento: Títulos dos pilotos)

Métodos de agrupamento	IRA	TEA
iSBC	0,0483	0,3000
iKSBC1C	-0,0714	0,3000
iKSBC2C	0,0483	0,3000
iKM	0,0708	0,3000
iKM+iSBC	0,0483	0,3000
iKM+iKSBC1C	0,0708	0,3000
iKM+iKSBC2C	**0,2857**	**0,2000**
iKKM	0,2593	0,2000
iKKM+iSBC	-0,1111	0,3000
iKKM+iKSBC1C	0,2593	0,2000
iKKM+iKSBC2C	0,2593	0,2000

A Tabela 6.6 apresenta os resultados obtidos para este conjunto de dados em relação a variável Pódios no ano de 2005. É possível observar que o desempenho dos métodos subtrativos baseados em *kernel*, iSBC1C e iSBC2C, foram superiores ao desempenho do método iSBC. As extensões do método iKM obtiveram o mesmo desempenho que ele, com excessão do método iKM+iSBC. Já os métodos iKKM+iSBC, iKKM+iSBC1C e iKKM+iSBC2C apresentaram desempenho superior ao método iKKM.

Tabela 6.6: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Fórmula 1* (Variável de agrupamento: Pódios no ano de 2005)

Métodos de agrupamento	IRA	TEA
iSBC	0,1223	0,4000
iKSBC1C	**0,3478**	**0,3000**
iKSBC2C	**0,3478**	**0,3000**
iKM	0,3478	0,3000
iKM+iSBC	0,1223	0,4000
iKM+iKSBC1C	0,3478	0,3000
iKM+iKSBC2C	0,3478	0,3000
iKKM	-0,0372	0,5000
iKKM+iSBC	**0,1223**	**0,4000**
iKKM+iKSBC1C	**0,1223**	**0,4000**
iKKM+iKSBC2C	**0,1223**	**0,4000**

6.6 Conjunto de dados: Fungos

As informações do conjunto de dados simbólicos de natureza intervalar *Fungos* (*http://www.mykoweb.com/CAF/genera/*) foram extraídas a partir de três gêneros de espécies de fungos presentes no estado da Califórnia, Estados Unidos: Agaricus, Amanita e Boletus. Agaricus é um gênero de cogumelos espinhosos, de médio a grande porte. São membros bem conhecidos do gênero Agaricus o A. *bisporus*, o "cogumelo botão"comum nos supermercados e o A. *campestris*, o "cogumelo do prado", um cogumelo popular comestível que é comum em pastagens e prados em muitas áreas temperadas do mundo. Amanita também é um gênero de cogumelos espinhosos, de médio a grande porte. O membro mais famoso do gênero Amanita é o A. *muscaria*, o "fly agaric", que possui um gorro vermelho adornado por verrugas brancas em sua parte superior. Amanita contém os cogumelos mais mortais do mundo (A. *phalloides*, A. *virosa*) e alguns excelentes cogumelos comestíveis. Já a principal característica do gênero Boletus são os esporos tipicamente castanhos a castanho-oliva.

Esse conjunto possui 55 espécies distribuídas em: 16 espécies do gênero *Amanita*, 24 espécies do gênero *Agaricus* e 15 espécies do gênero *Boletus*, onde cada espécie é descrita por cinco variáveis do tipo intervalo: largura do píleo, largura do estipe, espessura do estipe, altura dos esporos e largura dos esporos. O Apêndice M descreve as espécies de acordo com suas variáveis e espécies.

Tabela 6.7: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Fungos*

Métodos de agrupamento	IRA	TEA
iSBC	0,4887	0,2545
iKSBC1C	**0,7761**	**0,0727**
iKSBC2C	0,4564	0,2727
iKM	0,2417	0,3636
iKM+iSBC	0,0526	0,5091
iKM+iKSBC1C	**0,3921**	**0,3091**
iKM+iKSBC2C	**0,4265**	**0,2909**
iKKM	0,1623	0,4182
iKKM+iSBC	0,0169	0,5273
iKKM+iKSBC1C	**0,2268**	**0,3273**
iKKM+iKSBC2C	0,1497	0,4364

Os resultados do IRA e da TEA para este conjunto de dados podem ser vistos na Tabela 6.7. Observando os resultados é possível notar que o método iKSBC1C obteve desempenho superior aos métodos iSBC e iKSBC2C, assim como em relação ao método baseado em *kernel* iKKM e suas extensões, o método utilizando os centróides

do iKSBC1C (iKKM+iKSBC1C) também obteve desempenho superior aos métodos iKKM, iKKM+iSBC e iKKM+iKSBC2C. No entanto em relação ao método iKM e suas extensões, o método que obteve o melhor resultado foi o iKM+iKSBC2C.

6.7 Conjunto de dados: Iris

O conjunto de dados simbólicos de natureza intervalar Iris (Billard; Diday, 2006) exibe observações com valores de intervalo para 30 categorias de íris. Essas categorias foram geradas pela agregação de grupos consecutivos de cinco das 150 observações individuais publicadas em Fisher (1936). Tal agregação pode ocorrer se as cinco observações em cada categoria corresponderem às medições de cinco flores diferentes no mesmo local, ou cinco flores diferentes na mesma planta, etc. Existem quatro variáveis aleatórias: Comprimento da sépala, Largura da sépala, Comprimento da pétala e Largura da pétala. Esse conjunto está agrupado em 3 classes a priori de tamanhos iguais de acordo as espécies: setosa (S), versicolor (Ve) e virginica (Vi). Os dados, portanto, listam 30 observações, 10 para cada uma das três espécies, como pode ser visto no Apêndice N que descreve as observações de acordo com suas variáveis e espécies.

Tabela 6.8: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Iris*

Métodos de agrupamento	IRA	TEA
iSBC	**1,0000**	**0,0000**
iKSBC1C	**1,0000**	**0,0000**
iKSBC2C	**1,0000**	**0,0000**
iKM	1,0000	0,0000
iKM+iSBC	1,0000	0,0000
iKM+iKSBC1C	1,0000	0,0000
iKM+iKSBC2C	1,0000	0,0000
iKKM	0,5797	0,2000
iKKM+iSBC	**1,0000**	**0,0000**
iKKM+iKSBC1C	**1,0000**	**0,0000**
iKKM+iKSBC2C	**1,0000**	**0,0000**

Após a aplicação dos algoritmos referente aos métodos de agrupamento para dados simbólicos de natureza intervalar propostos é possível notar que todos os métodos obtiveram desempenho totalmente satisfatório em relação aos valores do cálcula do IRA e da TEA, exceto o método já existente *kernel* K-médias para dados simbólicos do tipo intervalo (iKKM), desenvolvido por Costa (2011), como pode ser visto na Tabela 6.8.

6.8 Conjunto de dados: Peixes

As informações do conjunto de dados simbólicos de natureza intervalar *Peixes* (*https://lhedjazi.jimdo.com/useful-links/*) foram coletadas por pesquisadores do laboratório LEESA - Laboratoire d'Ecophysiologie et d'Ecotoxicologie des Systemes Aquatiques, com o objetivo de obter um melhor conhecimento a respeito dos níveis anormais de contaminação de mercúrio em algumas regiões da Guyana francesa (COSTA, 2011).

Este conjunto de dados simbólicos de natureza intervalar consiste em 12 espécies de peixes, descritas por 13 variáveis do tipo intervalo: Comprimento, Peso, Músculo, Intestino, Estômago, Brânquias, Fígado, Rins, Fígado/músculo, Rins/músculo, Brânquias/músculo, Instestino/músculo, Estômago/músculo. Estas espécies estão agrupadas em 4 classes a priori de tamanhos diferentes de acordo com a dieta: Carnívoros, Onívoros, Detritívoros e Herbívoros. O Apêndice O descreve as espécies de acordo com suas variáveis e dieta.

Tabela 6.9: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Peixes*

Métodos de agrupamento	IRA	TEA
iSBC	0,1172	0,3333
iKSBC1C	0,1951	0,4167
iKSBC2C	**0,4535**	**0,2500**
iKM	0,0331	0,5000
iKM+iSBC	0,0331	0,5000
iKM+iKSBC1C	**0,0522**	**0,4167**
iKM+iKSBC2C	0,0331	0,5000
iKKM	0,0209	0,4167
iKKM+iSBC	-0,0942	0,5000
iKKM+iKSBC1C	-0,0366	0,5000
iKKM+iKSBC2C	-0,0366	0,5000

A Tabela 6.9 ilustra os resultados do IRA e da TEA para o conjunto de dados simbólicos de natureza intervalar *Peixes*. O desempenho do método iKSBC2C foi superior ao dos métodos iSBC e iKSBC1C. Já em relação ao método iKM e suas extensões o único que obteve resultado diferente e superior ao método já existente foi o iKM+iKSBC1C, as demais extensões obtiveram desempenho igual ao método iKM, $0,0331$ como resultado do cálculo do IRA e $0,5000$ como resultado do cálculo da TEA. O método IKKM obteve o melhor resultado comparando com os métodos que são extensões dele.

6.9 Conjunto de dados: Temperatura das cidades

O conjunto de dados simbólicos de natureza intervalar *Temperatura das cidades* foi inicialmente apresentado em Guru, Kiranagi e Nagabhushan (2004), relativo a 37 cidades. Estas cidades são descritas por 12 variáveis intervalares que indicam as temperaturas mínima e máxima dos 12 meses do ano em graus centígrados. E está agrupado em quatro classes a priori de tamanhos diferentes: quinze cidades pertencem a classe 1, vinte pertencem a classe 2, uma cidade pertence a classe 3 e uma a classe 4. Como pode ser visto no Apêndice P, que descreve as cidades de acordo com suas variáveis e classes.

Tabela 6.10: Resultado do Índice de Rand Ajustado (IRA) e da taxa total de erro de alocação (TEA) para o conjunto de dados *Temperatura das cidades*

Métodos de agrupamento	IRA	TEA
iSBC	0,3372	0,4324
iKSBC1C	**0,4804**	**0,2973**
iKSBC2C	0,3935	0,3243
iKM	0,4834	0,2703
iKM+iSBC	0,3387	0,4324
iKM+iKSBC1C	**0,5655**	**0,2703**
iKM+iKSBC2C	0,3459	0,3514
iKKM	-0,0356	0,4324
iKKM+iSBC	-0,0488	0,4595
iKKM+iKSBC1C	**0,0207**	**0,3784**
iKKM+iKSBC2C	-0,0468	0,4595

A Tabela 6.10 mostra que dentre os métodos subtrativos propostos o que obteve melhor desempenho foi o método iKSBC1C. Bem como em relação aos métodos iKM e suas extensões e iKKM e suas extensões, os que obtiveram melhores resultados foram os que utilizam os centróides do método iKSBC1C, iKM+iKSBC1C e iKKM+iKSBC1C, respectivamente. Em relação ao método iKKM e suas extensões também é possível notar que o único método que não obteve resultado muito baixo para o cálculo do IRA foi o iKKM+iKSBC1C. E apesar deste conjunto de dados ter muitas variáveis, o intuito da avaliação dos métodos para dados simbólicos de natureza intervalar propostos neste tipo de cenário é reforçar a validação deles, observando seu comportamento diante de cenários reais.

Capítulo 7

Conclusões

Neste trabalho foram propostos três conjuntos de métodos de agrupamento para dados simbólicos de natureza intervalar. O primeiro conjunto de métodos são extensões do método de agrupamento subtrativo, onde em uma extensão a principal diferença ocorre na capacidade de interpretar dados de natureza intervalar, enquanto que na outras duas extensões a principal diferença também ocorre na capacidade de interpretar dados de natureza intervalar, mas também na utilização de funções *kernel* para auxiliar no processo de identificação de grupos não-lineares. Os outros dois conjuntos de métodos são estratégias para melhorar o desempenho dos métodos de agrupamento K-médias para dados simbólicos do tipo intervalo baseado em distância L_2 e *kernel* K-médias para dados simbólicos do tipo intervalo, onde a diferença ocorre na utilização dos centróides dos métodos subtrativos propostos nesta dissertação como entradas destes métodos.

As variações do método subtrativo, para dados intervalares utilizando funções *kernel* foram tratadas de duas formas: na primeira o intervalo foi considerado como uma informação única (definido por uma componente) e na segunda existe um tratamento separado para os limites inferiores e superiores de cada intervalo (definido por duas componentes). Em ambos os casos o objetivo é identificar a não-linearidade dos dados representados por meio de conjuntos de intervalos, caracterizando a variabilidade e incerteza das informações. As variações do método *kernel* K-médias foram aplicadas sob o agrupamento no espaço de entrada, onde os centróides dos grupos podem ser encontrados explicitamente no espaço de dados.

Foram realizados experimentos com dados simbólicos de natureza intervalar reais, com nove conjuntos: Carros, Desenvolvimento dos países, Facedata, Fluxos de água, Fórmula 1, Fungos, Iris, Peixes e Temperatura das cidades. Para isto, adotou-se o Índice de Rand Ajustado (IRA) e a taxa total de erro de alocação (TEA) como forma de avaliar os métodos.

Como conclusão principal, os métodos de agrupamento subtrativo baseados em *kernel* para dados simbólicos de natureza intervalar se mostraram mais eficientes do

que o método de agrupamento subtrativo para dados simbólicos de natureza intervalar, quando avaliados através do IRA e da TEA. Desta forma, esta dissertação contribuiu para o desenvolvimento de métodos para o reconhecimento/classificação de dados não-linearmente distribuídos, no âmbito dos algoritmos de agrupamento para dados simbólicos intervalares. E como contribuição deste trabalho podemos citar a aplicação e extensão de métodos baseados em *kernel* para literatura de Análise de Dados Simbólicos (ADS).

Especificamente sobre as variações dos métodos K-médias e *kernel* K-médias podemos afirmar que, a utilização dos centróides dos métodos subtrativos propostos como entradas destes métodos minimizaram a sensibilidade que os métodos originais tinham em relação a escolha do centróide para a definição da partição inicial. E na maioria dos resultados apresentados as variações utilizando os centróides dos métodos subtrativos baseados em *kernel* para dados intervalares obtiveram desempenho superior aos métodos para dados intervalares originais e aos métodos utilizando como centróides fixos os centróides dos método subtrativo para dados simbólicos de natureza intervalar.

Como trabalhos futuros podemos citar a realização de experimentos com dados simbólicos do tipo intervalo simulados, com o objetivo de que os experimentos possam fornecer resultados referentes a eficiência dos métodos propostos em encontrar partições de dados não-linearmente separáveis para o espaço de entrada.

Os conjuntos de dados reais utilizados nos experimentos podem ser vistos nos Apêndices deste trabalho e também podem ser solicitados por e-mail (camila.ravena@hotmail.com).

Referências Bibliográficas

[1] Arthur, D.; Vassilvitskii, S. (2007). K-means^{++}: The advantages of careful seeding. Proceeding SODA'07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, New Orleans, Louisiana. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, pp. 1027-1035.

[2] Backer, E. (1978). Cluster analysis by optimal decomposition of induced fuzzy sets. Delft University Press.

[3] Ball, G.; Hall, D. (1965). ISODATA, a novel method of data analysis and pattern classification. Tech. rept. NTIS AD 699616. Stanford Research Institute, Menlo Park, CA. Architectural design, vl. 699, edt. 616.

[4] Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. (2005). Clustering with Bregman Divergences. The Journal of Machine Learning Research, 6: 1705–1749.

[5] Bertrand, P.; Goupil, F. (2000). 'Descriptive Statistics for Symbolic Data'. In: Bock, H.-H., Diday, E. (Eds.), Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information From Complex Data, Berlin: Springer-Verlag. New York. pp. 106-124 [416,418].

[6] Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

[7] Billard, L.; Diday, E. (2000). Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.-P., Groenen, P.J.F., Schader, M. (Eds.), Data Analysis, Classification, and Related Methods, Springer-Verlag, Berlin. pp. 369-374.

[8] Billard, L.; Diday, E. (2003). Billard, L.; Diday, E. From the statistics of data to the statistics of knowledge: symbolic data analysis. Journal of the American Statistical Association, 98(462):470-487.

[9] Billard, L.; Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons Ltd., England.

[10] Bock, H.-H.; Diday, E. (2000). Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag Berlin Heidelberg.

[11] Bock, H.-H. (2002). Clustering algorithms and Kohonen maps for symbolic data. Journal of the Japanese Society of Computational Statistics. 15: 1-13.

[12] Bock, H. (2008). Visualizing symbolic data by Kohonen maps. In: Diday, E., Noirhomme-Fraiture, M. (Eds.), Symbolic Data Analysis and the SODAS Software, Wiley, Chichester, pp. 205-234.

[13] Bradlay, P. S.; Fayyad, U.; Reina, C. (1998). Scaling clustering algorithms to large databases. In: KDD'98 - Fourth International Conference on Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, New York, NY, pp. 9-15.

[14] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C.J. (1984). Classification and regression trees. Califórnia City: Wadsworth International Group, 358 p.

[15] Brito, P. (2002). Hierarchical and pyramidal clustering for symbolic data. Journal of the Japanese Society of Computational Statistics, 15(2): 231-244.

[16] Camastra, F.; Verri, A. (2005). A novel kernel method for clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(5): 801-805.

[17] Chavent, M.; de Carvalho, F.A.T.; Lechevallier, Y.; Verde, R. (2006). New clustering methods for interval data. Computational Statistics, 21: 211-230.

[18] Cheung, Y. M. (2003) k*-Means: A new generalized k-means clustering algorithm. Pattern Recognition Letters 24(15): 2883–2893.

[19] Chinrungrueng, C.; Séquin, C.H. (1995). Optimal Adaptative K-means Algorithm with Dynamic Adjustment of Learning Rate. IEEE Transactions Neural Network 6(1): 157-169.

[20] Chiu, S.L. (1994). Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent and Fuzzy Systems. John Wiley & Sons, 2: 267-278.

[21] Chiu, S.L. (1997). Extracting Fuzzy Rules from Data for Function Approximation and Pattern Classification. Chapter 9 in Fuzzy Information Engineering: A Guided Tour of Applications, ed. D. Dubois, H. Prade, and R. Yager, John Wiley & Sons.

[22] Chouakria, A.; Diday, E.; Cazes, P. (1998). An improved factorial representation of symbolic objects. In Studies and Research, Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98), pp. 276-289. Luxembourg: Office for Official Publications of the European Communities.

[23] Costa, A.F.B.F.; Pimentel, B.A.; Souza, R.M.C.R. (2010). A kernel k-means clustering method for symbolic interval data. In: International Joint Conference on Neural Networks, IJCNN 2010, Barcelona, Spain. [S.l.:s.n.]. p.1-6. ISBN 978-1-4244-6916-1.

[24] de Carvalho, F.A.T; Brito, P.; Bock, H.-H. (2006). Dynamic Clustering for Interval Data Based on L_2 Distance. Computational Statistics, 21(2): 231-250.

[25] Diday, E; Lemaire, J.; Pouget, J.; Testu, F. (1982). Éléments d'analyse de données. Dunod, Paris.

[26] Diday, E. (1986). Orders and overlapping cluster hi pyramids. Multidimensional Data Analysis, DSWO Press, Leiden, pp. 201-234.

[27] Diday, E. (1987). The symbolic approach in clustering and related methods of data analysis. In: Bock, H.-H. (Ed.), Classification and Related Methods of Data Analysis, North-Holland, Amsterdam. pp. 673-684.

[28] Diday, E.; Noirhomme-Fraiture, M. (2008). Symbolic Data Analysis and the SODAS Software. Wiley-Interscience. New York, NY.

[29] Duarte Silva, A.P.; Brito, P. (2006). Linear discriminant analysis for interval data. Computational Statistics, 21(2): 289-308.

[30] Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3(3): 32-57.

[31] El Golli, A.; Conan-Guez, B.; Rossi, F. (2004). A self-organizing map for dissimilarity data. In: D. Banks; L. House; F.R. McMorris; P. Arabie; W. Gaul (Eds.): Classification, clustering, and data mining applications. Studies

in Classification, Data Analysis, and Knowledge Organization. Springer, Heidelberg, pp. 61-68.

[32] Everitt, B.S.; Landau, S.; Leese, M.; Stahl, D. (2011). Cluster Analysis. 5th Edition. Wiley Series in Probability and Statistics. Wiley.

[33] Filippone, M.; Camastra, F.; Masulli, F.; Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. Pattern Recognition, 41:176-190.

[34] Fisher, R.A. (1936). The Use of Multiple Measurement in Taxonomic Problems. Annals of Eugenics 7, 179-188.

[35] Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics, 21(3): 768–769.

[36] Gioia, F.; Lauro, C.N. (2006). Principal component analysis on interval data. In: Computational Statistics, Physica-Verlag, v.21 n.2, p.343-363.

[37] Girolami, M. (2002). Mercer kernel-based clustering in feature space. IEEE Trans. Neural Networks, 13(3):780-784.

[38] Guru, D. S.; Kiranagi, B. B.; Nagabhushan, P. Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. Pattern Recognition Letters, Elsevier Science Inc., New York, NY, USA, v. 25, p. 1203-1213, July 2004. ISSN 0167-8655.

[39] Hamada, A.; Minami, H.; and Mizuta, M. (2008). Principal component analysis for modal interval-valued data. Proceedings of IASC2008, the Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, 512-519.

[40] Hand, D. J.; Mannila, H.; Smyth, P. (2001). Principles of Data Mining. The MIT Press, Massachusetts.

[41] Hansen, P.; Mladenović, N. (2001) J-Means: a new local search heuristic for minimum sum of squares clustering. Pattern Recognition 34(2): 405–413.

[42] Har-Peled, S.; Mazumdar, S. (2004). Coresets for k-means and k-median clustering. Proceeding STOC '04 Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, Chicago, IL, USA. ACM New York, NY, USA, 16: 291-300.

[43] Hastie, T.; Tibshirani, R.; Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York.

[44] Hayes-Roth, F.; McDermott, J. (1978). An interference matching technique for inducing abstractions. Communications of the ACM, New York, NY, USA 21(5): 401-411.

[45] Höppner, F.; Klawonn, F.; Kruse, R.; Runkler, T. (1999). Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. John Wiley & Sons, New York.

[46] Hubert, L.; Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1): 193-218.

[47] Jain, A. K.; Murty, M.N.; Flynn, P.J. (1999). Data Clustering: A Review. Journal ACM Computing Surveys, 31(3): 264-323.

[48] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. Pattern Recogn. Lett., 31(8):651-666.

[49] Kao, C.-H.; Nakano, J.; Shieh, S.-H.; Tien, Y.-J.; Wu, H.-M.; Yang, C.-K.; Chen, C.-H. (2014). Exploratory data analysis of interval-valued symbolic data with matrix visualization. Computational Statistics & Data Analysis, Elsevier Science Publishers B.V. Amsterdam, Netherlands, 79: 14-29.

[50] Kashima, H.; Hu, J.; Ray, B.; Singh, M. (2008). K-means Clustering of Proportional Data Using L_1 Distance. Pattern Recognition, 2008. ICPR 2008 - 19th International Conference on Pattern Recognition, Tampa, FL, USA.

[51] Kauffman, L.; Rousseeuw, P.J. (1990). Finding groups in data: An itroduction to cluster analysis. Volume 603 de Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey.

[52] Kim, D. W.; Lee, K. Y.; Lee, D.; Lee, K. H. (2005). A kernel-based subtractive clustering method. Pattern Recognition Letters, 26:879-891.

[53] Kohonen, T. (1989). Self-organization and associative memory. Springer-Verlag Berlin Heidelberg, vl. 8.

[54] Kohonen, T. (2013). Essentials of the self-organizing map. Neural Netw., 37:52-65.

[55] Krishna, K.; Murtyg, M.N. (1999). Genetic K-means algorithm. Journal IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Press Piscataway, NJ, USA, 29(3): 433–439.

[56] Lauro, N.C.; Verde, R.; Palumbo, F. (2000). Factorial discriminant analysis on symbolic objects. In: Bock, H.-H., Diday, E. (Eds.), Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information From Complex Data, Springer-Verlag Berlin Heidelberg, pp. 212-233.

[57] Lebart, L.; Morineau, A.; Warwick, K.M. (1984) Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley, New York.

[58] Likas, A.; Vlassis, N.; Verbeek, J.J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2): 451-461.

[59] Linde, Y.; Buzo, A.; Gray, R. (1980). An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications, 28(1): 84–94.

[60] Mao, J.; Jain, A.K. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). IEEE Transactions on Neural Networks, 7(1): 16–29.

[61] Martinetz, T. M.; Berkovich, S. G.; Schulten, K. J. (1993). 'neural gas' network for vector quantization and its application to time-series prediction. IEEE Trans. Neural Networks, 4(4):558-569.

[62] Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integrals equations. In Proc. R. Soc. London, volume 209, pages 415-446.

[63] Michalski, R. S. (1973). "AQVAL/1–Computer Implementation of a Variable-Valued Logic System VL1 and Examples of its Application to Pattern Recognition,". Proceedings of the First International Joint Conference on Pattern Recognition, Washington, DC, pp. 3-17.

[64] Milligan, G.W.; Cooper, M.C. (1986). A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. Multivariate Behavioral Research, 21(4): 441-458.

[65] Mota Filho, F. O. M. (2005). Aplicação de modelos de estimação de fitness em algoritmos geneticos. Dissertação de mestrado. Universidade Estadual de Campinas, Faculdade de Engenharia Eletrica e de Computação, Campinas, SP.

[66] Müller, K.-R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. (2001). An Introduction to Kernel-based Learning Algorithms. IEEE Trans. Neural Networks, 12(2):181-201.

[67] Neto, E.A.L; de Carvalho, F.A.T. (2008). Centre and Range method for fitting a linear regression model to symbolic interval data, Computational Statistics & Data Analysis, 52(3): 1500-1515.

[68] Neto, E.A.L; de Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables, Computational Statistics & Data Analysis, 54(2): 333-347.

[69] Palumbo, F.; Lauro, C.N. (2003). A PCA for interval valued data based on midpoints and radii. In: Yanai, H., Okada, A., Shigemasu, K., Kano, Y., Meulman, J.J. (Eds.), New Developments in Psychometrics, Springer-Verlag, Tokyo. pp. 641-648.

[70] Pelleg, D.; Moore, A. (1999). Accelerating Exact k-means Algorithms with Geometric Reasoning. In: KDD'99 - Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA. ACM New York, NY, USA, pp. 277-281.

[71] Russell, S. J.; Norvig, P. (2010). Artificial Inteligence: A modern Approach. 3rd edition, Pearson Educationn, Inc., New Jersey.

[72] Schölkopf, B., Smola, A.; Müller, K. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Massachusetts Institute of Technology, MIT Press Cambridge, MA, USA, Journal Neural Computation, 10(5): 1299–1319.

[73] Schölkopf, B.; Smola, A.J. (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge.

[74] Sneath, P.H.A.; Sokal, R.R. (1973). Numerical Taxonomy: The Principles and Pratice of Numerical Classification. W.H. Freeman and Company, San Francisco, USA.

[75] Steinbach, M.; Karypis, G.; Kumar, V. (2000). A Comparison of Document Clustering Techniques. In: KDD Workshop on Text Mining.

[76] Steinhaus, H. (1956). Sur la division des corp materiels en parties. Bulletin de L'Académie Polonaise des Sciences, (Cl. III), IV(12): 801–804.

[77] Souza, R.M.C.R.; de Carvalho, F.A.T. (2004). Clustering of interval data based on city-block distances. Pattern Recognition Letters, 25(3): 353-365.

[78] Tukey, J.W. (1977). Exploratory Data Analysis. Addison-Wesley, Mass.

[79] Yager, R. R.; Filev, D. P. (1994). Approximate clustering via the mountain method. IEEE Trans. Systems, Man, Cybernet., 24(8):1279-1284.

[80] Yang, M.-S; Wu, K.-L. (2005). A modified mountain clustering algorithm. Pattern Anal Applic. 8: 125–138. Springer-Verlag London Limited.

Apêndice A

Algoritmo de Montanha

1. Inicializar os parâmetros α, β e o intervalo I_k, $1 \leq k \leq p$;

2. Quantizar os intervalos e formar a grade;

3. Calcular os valores da função de montanha

$$M(\underset{\sim}{y}_i)) = \sum_{j=1}^{n} e^{-\alpha d(\underset{\sim}{x}_j, \underset{\sim}{y}_i)}, i = 1, 2, ...;$$

4. Escolher o ponto de grade $\underset{\sim}{y}_i$ para o qual o valor da função de montanha $M(\underset{\sim}{y}_i)$ é a mais alta como centróide de grupo;

5. Modificar e recalcular a função de montanha

$$\widehat{M}^j(\underset{\sim}{y}_i)) = \widehat{M}^{j-1}(\underset{\sim}{y}_i) - M^*_{j-1} \sum_{j=1}^{n} e^{-\beta d(\underset{\sim}{y}^*_{j-1}, \underset{\sim}{y}_i))};$$

6. Se o número de centróides encontrados for igual ao número pré-especificado de grupos δ, de acordo com $\dfrac{M^*_{j-1}}{M^*_1} < \delta$, então pare; Caso contrário, vá para o passo 4.

Apêndice B

Algoritmo KM

1. Inicialização
   Fixe K (o número de grupos), $2 \leq K < n$; escolha aleatoriamente uma partição inicial P em K grupos $P_1, ..., P_K$ ou, alternativamente, escolha K padrões distintos $\underset{\sim}{v}_1, ..., \underset{\sim}{v}_K$ como centróides iniciais e aloque cada padrão i de acordo com o centróide mais próximo v_r ($r = \arg\min_{1 \leq k \leq K} ||\underset{\sim}{x}_i - \underset{\sim}{v}_k||^2$) para obter a partição inicial $P = \{P_1, ..., P_K\}$.

2. Etapa 1: Definição dos melhores centróides dos grupos
   Atualize os centróides dos grupos v_k ($k = 1, ..., K$) de acordo com

$$\underset{\sim}{v}_k = \frac{1}{|P_k|} \sum_{i \in P_k} \underset{\sim}{x}_i.$$

3. Etapa 2: Definição da melhor partição
   $test \leftarrow 0$
   para $i = 1$ até n faça
   defina o grupo vencedor P_r tal que
   $r = \arg\min_{1 \leq k \leq K} ||\underset{\sim}{x}_i - \underset{\sim}{v}_k||^2$
   se $i \in P_k$ e $r \neq k$
   $test \leftarrow 1$
   $P_r \leftarrow P_r \cup \{i\}$
   $P_k \leftarrow P_k \setminus \{i\}$

4. Critério de parada
   Se $test = 0$, então, PARE, caso contrário, volte ao passo 2.

Apêndice C

Algoritmo iKM

1. Inicialização
   Fixe K (o número de grupos), $2 \leq K < n$; escolha aleatoriamente uma partição inicial P^* em K grupos $P_1, ..., P_K$ ou, alternativamente, escolha K padrões distintos $\underset{\sim}{g}_1, ..., \underset{\sim}{g}_K$, onde $\underset{\sim}{g}_k = ([\alpha_k^1, \beta_k^1], ..., [\alpha_k^p, \beta_k^p])$, $\forall k = 1, ..., K$, como centróides iniciais e aloque cada padrão i de acordo com o centróide mais próximo g_s ($s = \arg\min_{1 \leq k \leq K} ||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2$) para obter a partição inicial $P^* = \{P_1, ..., P_K\}$.

2. Etapa 1: Definição dos melhores centróides dos grupos
   Atualize os centróides dos grupos $\underset{\sim}{g}_k$ $(k = 1, ..., K)$ de acordo com

$$\underset{\sim}{g}_k = \frac{1}{|P_k^*|} \sum_{i \in P_k^*} \underset{\sim}{x}_i.$$

3. Etapa 2: Definição da melhor partição
   $test \leftarrow 0$
   para $i = 1$ até n faça
   defina o grupo vencedor P_s^* tal que
   $s = \arg\min_{1 \leq k \leq K} ||\underset{\sim}{x}_i - \underset{\sim}{g}_k||^2$
   se $i \in P_k^*$ e $s \neq k$
   $test \leftarrow 1$
   $P_s^* \leftarrow P_s^* \cup \{i\}$
   $P_k^* \leftarrow P_k^* \setminus \{i\}$

4. Critério de parada
   Se $test = 0$, então, PARE, caso contrário, volte ao passo 2.

Apêndice D

Algoritmo SBC

1. Dado o número de grupos, δ, inicializar os parâmetros α, β;
2. Calcular a função potencial

$$M(\underset{\sim}{x}_i) = \sum_{j=1}^{n} e^{-\alpha||\underset{\sim}{x}_i - \underset{\sim}{x}_j||^2};$$

3. Escolher o ponto de dados $\underset{\sim}{x}_i$ cuja função potencial $M(\underset{\sim}{x}_i)$ é mais alta como centróide do grupo;
4. Modificar e recalcular a função potencial

$$\widehat{M}^j(\underset{\sim}{x}_i) = \widehat{M}^{j-1}(\underset{\sim}{x}_i) - M_{j-1}^* \sum_{j=1}^{n} e^{-\beta||\underset{\sim}{x}_i - \underset{\sim}{x}_{j-1}^*||^2};$$

5. Se o número de centróides encontrado for igual a δ, então pare; caso contrário voltar para o passo 3.

Apêndice E

Algoritmo KKM

1. Inicialização
   Fixe K (o número de grupos), $2 \leq K < n$; escolha aleatoriamente uma partição inicial P de Ω em K grupos $P_1, ..., P_K$ ou, alternativamente, escolha K padrões distintos $\underset{\sim}{v}_1, ..., \underset{\sim}{v}_K$ pertencendo a Ω como centróides iniciais e aloque cada padrão i de acordo com o centróide mais próximo v_h ($h = \arg\min_{1 \leq k \leq K} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{v}_k)||^2$) para obter a partição inicial $P = \{P_1, ..., P_K\}$.

2. Etapa 1: Definição dos melhores centróides dos grupos
   Atualize os centróides dos grupos v_k $(k = 1, ..., K)$ de acordo com

$$v_k = \frac{\sum_{i \in P_k} \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{v}_k)\underset{\sim}{x}_i}{\sum_{i \in P_k} \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{v}_k)}.$$

3. Etapa 2: Definição da melhor partição
   $test \leftarrow 0$
   para $i = 1$ até n faça
   defina o grupo vencedor P_h tal que
   $h = \arg\min_{1 \leq k \leq K} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{v}_k)||^2$
   se $i \in P_k$ e $h \neq k$
   $test \leftarrow 1$
   $P_h \leftarrow P_h \cup \{i\}$
   $P_k \leftarrow P_k \setminus \{i\}$

4. Critério de parada
   Se $test = 0$, então, PARE, caso contrário, volte ao passo 2.

Apêndice F

Algoritmo iKKM

1. Inicialização
   Fixe K (o número de grupos), $2 \leq K < n$; escolha aleatoriamente uma partição inicial P^* de Ω em K grupos $P_1, ..., P_K$ ou, alternativamente, escolha K padrões distintos $\underset{\sim}{g}_1, ..., \underset{\sim}{g}_K$ pertencendo a Ω, onde $\underset{\sim}{g}_k = ([\alpha_k^1, \beta_k^1], ..., [\alpha_k^p, \beta_k^p])$, $\forall k = 1, ..., K$, como centróides iniciais e aloque cada padrão i de acordo com o centróide mais próximo g_m ($m = \arg\min_{1 \leq k \leq K} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k)||^2$) para obter a partição inicial $P^* = \{P_1, ..., P_K\}$.

2. Etapa 1: Definição dos melhores centróides dos grupos
   Atualize os centróides dos grupos $\underset{\sim}{g}_k$ ($k = 1, ..., K$) de acordo com

$$\underset{\sim}{g}_k = \frac{\sum_{i \in P_k^*} \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{g}_k) x_i}{\sum_{i \in P_k^*} \mathcal{K}(\underset{\sim}{x}_i, \underset{\sim}{g}_k)}.$$

3. Etapa 2: Definição da melhor partição
   $test \leftarrow 0$
   para $i = 1$ até n faça
      defina o grupo vencedor P_m^* tal que
         $m = \arg\min_{1 \leq k \leq K} ||\Phi(\underset{\sim}{x}_i) - \Phi(\underset{\sim}{g}_k)||^2$
      se $i \in P_k^*$ e $m \neq k$
         $test \leftarrow 1$
         $P_m^* \leftarrow P_m^* \cup \{i\}$
         $P_k^* \leftarrow P_k^* \setminus \{i\}$

4. Critério de parada
   Se $test = 0$, então, PARE, caso contrário, volte ao passo 2.

Apêndice G

Algoritmo KSBC

1. Dado o número de grupos, δ, e os valores escolhidos de α, β, escolher uma função *kernel* $\mathcal{K}$;

2. Calcular a função potencial

$$M(\Phi(\underset{\sim}{x}_i)) = \sum_{j=1}^{n} e^{-\alpha(||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}_j)||^2)};$$

3. Escolher o ponto de dados $\underset{\sim}{x}_i$ cuja função potencial $M(\Phi(\underset{\sim}{x}_i))$ é mais alta como centróide do grupo;

4. Modificar e recalcular a função potencial

$$\widehat{M}^j(\Phi(\underset{\sim}{x}_i)) = \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i)) - M^*_{j-1} \sum_{j=1}^{n} e^{-\beta||\Phi(\underset{\sim}{x}_i)-\Phi(\underset{\sim}{x}^*_{j-1})||^2}$$

$$= \widehat{M}^{j-1}(\Phi(\underset{\sim}{x}_i)) - M^*_{j-1} \textstyle\sum_{j=1}^{n} e^{-\beta(\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}_i)-2\mathcal{K}(\underset{\sim}{x}_i,\underset{\sim}{x}^*_{j-1})+\mathcal{K}(\underset{\sim}{x}^*_{j-1},\underset{\sim}{x}^*_{j-1}))};$$

5. Se o número de centróides encontrado for igual a δ, então pare; caso contrário voltar para o passo 3.

Apêndice H

Conjunto de dados simbólicos de natureza intervalar Carros

Tabela H.1: Conjunto de dados simbólicos de natureza intervalar Carros

Modelos	Categoria	Preço	Cilindrada	Velocidade_Max	Aceleração	Passo	Comprimento	Largura	Altura
Alfa 145_U	Utilitaria	[27806,33596]	[1370,1910]	[185,211]	[8.3,11.2]	[254,254]	[406,406]	[171,171]	[143,143]
Alfa 156_B	Berlina	[41593,62291]	[1598,2492]	[200,227]	[8.5,10.5]	[260,260]	[443,443]	[175,175]	[142,142]
Alfa 166_L	Ammiraglia	[64499,88760]	[1970,2959]	[204,211]	[9.8,9.9]	[270,270]	[472,472]	[182,182]	[142,142]
Aston Martin_S	Sportiva	[260500,460000]	[5935,5935]	[298,306]	[4.7,5]	[259,269]	[465,467]	[183,192]	[124,132]
Audi A3_U	Utilitaria	[40230,68838]	[1595,1781]	[189,238]	[6.8,10.9]	[250,251]	[415,415]	[174,174]	[142,142]
Audi A6_B	Berlina	[68216,140265]	[1781,4172]	[216,250]	[6.7,9.7]	[276,276]	[480,480]	[181,181]	[145,145]
Audi A8_L	Ammiraglia	[123849,171417]	[2771,4172]	[232,250]	[5.4,10.1]	[289,289]	[503,503]	[188,188]	[144,144]
Bmw serie 3_B	Berlina	[45407,76392]	[1796,2979]	[201,247]	[6.6,10.9]	[273,273]	[447,447]	[174,174]	[142,142]
Bmw serie 5_L	Ammiraglia	[70292,198792]	[2171,4398]	[226,250]	[6.7,9.1]	[283,283]	[478,478]	[180,180]	[144,144]
Bmw serie 7_L	Ammiraglia	[104892,276792]	[2793,5397]	[228,240]	[7,8.6]	[293,307]	[498,512]	[186,186]	[143,143]
Ferrari_S	Sportiva	[240292,391692]	[3586,5474]	[295,298]	[4.5,5.2]	[260,260]	[476,476]	[192,192]	[130,130]
Punto_U	Utilitaria	[19229,30885]	[1242,1910]	[155,170]	[12.2,14.3]	[246,246]	[380,384]	[166,166]	[148,148]
Fiesta_U	Utilitaria	[19242,24742]	[1242,1753]	[167,167]	[13.1,13.9]	[245,245]	[383,383]	[163,163]	[132,132]
Focus_B	Berlina	[27492,34092]	[1596,1753]	[185,193]	[10.8,11]	[262,262]	[415,415]	[170,170]	[143,143]
Honda NSK_S	Sportiva	[205242,215242]	[2977,3179]	[260,270]	[5.7,6.5]	[253,253]	[414,414]	[175,175]	[129,129]
Lamborghini_S	Sportiva	[413000,423000]	[5992,5992]	[335,335]	[3.9,3.9]	[265,265]	[447,447]	[204,204]	[111,111]
Lancia Y_U	Utilitaria	[19837,29034]	[1242,1242]	[158,174]	[11.2,14.1]	[238,238]	[372,372]	[169,169]	[144,144]
Lancia K_L	Ammiraglia	[58806,81306]	[1998,2959]	[212,220]	[8.9,9.2]	[270,270]	[469,469]	[183,183]	[146,146]
Maserati GT_S	Sportiva	[155000,159500]	[3217,3217]	[280,290]	[5.1,5.7]	[266,266]	[451,451]	[182,182]	[131,131]
Mercedes SL_S	Sportiva	[132800,262500]	[2799,5987]	[232,250]	[6.1,9.7]	[252,252]	[447,447]	[181,181]	[129,129]
Mercedes Classe C_B	Berlina	[55902,115248]	[1998,3199]	[210,250]	[5.2,11]	[272,272]	[453,453]	[173,173]	[143,143]
Mercedes Classe E_L	Ammiraglia	[69243,389405]	[1998,5439]	[222,250]	[5.7,9.7]	[283,283]	[482,482]	[180,180]	[144,144]
Mercedes Classe S_L	Ammiraglia	[128202,394342]	[3199,5786]	[210,240]	[7.2,8.4]	[297,309]	[504,516]	[186,186]	[144,144]
Nissan Micra_U	Utilitaria	[18492,24192]	[998,1348]	[150,164]	[12.5,15.5]	[236,236]	[375,375]	[160,160]	[144,144]
Corsa_U	Utilitaria	[19212,30612]	[973,1796]	[155,202]	[9,17]	[249,249]	[382,382]	[165,165]	[144,144]
Vectra_B	Berlina	[36492,49092]	[1598,2171]	[193,207]	[10.5,12.5]	[264,264]	[450,450]	[171,171]	[143,143]
Porsche_S	Sportiva	[147704,246412]	[3387,3600]	[280,305]	[4.2,5.2]	[235,235]	[443,444]	[177,183]	[130,131]
Twingo_U	Utilitaria	[16992,23492]	[1149,1149]	[151,168]	[11.7,13.4]	[235,235]	[343,343]	[163,163]	[142,142]
Rover 25_U	Utilitaria	[21492,33042]	[1119,1994]	[160,185]	[10.7,15]	[251,251]	[399,399]	[169,169]	[142,142]
Rover 75_B	Berlina	[50490,65399]	[1796,2497]	[195,210]	[10.2,11.6]	[275,275]	[475,475]	[178,178]	[143,143]
Skoda Fabia_U	Utilitaria	[19519,32686]	[1397,1896]	[157,183]	[11.5,16.5]	[246,246]	[396,396]	[165,165]	[145,145]
Skoda Octavia_B	Berlina	[27419,48679]	[1585,1896]	[190,191]	[11.1,11.8]	[251,251]	[452,452]	[173,173]	[143,143]
Passat_L	Ammiraglia	[39676,63455]	[1595,2496]	[192,220]	[9.6,12.7]	[270,270]	[470,470]	[175,175]	[146,146]

Apêndice I

Conjunto de dados simbólicos de natureza intervalar Desenvolvimento dos países

Tabela I.1: Conjunto de dados simbólicos de natureza intervalar Desenvolvimento dos países

Países	População	Esp_vida	PIB_hab	IDH	Categoria
1	[1161,67884]	[56.7,70.7]	[3419,9107]	[0.6,0.76]	país desenvolvido
2	[4469,82017]	[77.2,79.3]	[22093,28433]	[0.92,0.94]	país desenvolvido
3	[15211,283230]	[67.2,78.5]	[5749,31872]	[0.71,0.94]	país desenvolvido
4	[814,19138]	[68.4,78.7]	[4799,24574]	[0.76,0.94]	país desenvolvido
5	[1988,145491]	[66.1,78.2]	[6041,25443]	[0.77,0.94]	país em desenvolvimento
6	[4913,1275130]	[65.1,80.5]	[2857,24898]	[0.68,0.93]	país desenvolvido
7	[2576,25662]	[64,76]	[3561,8860]	[0.63,0.82]	país em desenvolvimento
8	[16189,1008940]	[62.3,71.9]	[2248,8209]	[0.57,0.77]	país em desenvolvimento
9	[1757,35119]	[40.6,52.3]	[501,5468]	[0.31,0.6]	país em desenvolvimento
10	[159,4809]	[55.6,70.5]	[2367,4047]	[0.53,0.7]	país em desenvolvimento

Apêndice J

Conjunto de dados simbólicos de natureza intervalar Facedata

Tabela J.1: Conjunto de dados simbólicos de natureza intervalar Facedata

Categoria	AD	BC	AH	DH	EH	GH
HUS1	[168.86,172.84]	[58.55,63.39]	[102.83,106.53]	[122.38,124.52]	[56.73,61.07]	[60.44,64.54]
HUS2	[169.85,175.03]	[60.21,64.38]	[102.94,108.71]	[120.24,124.52]	[56.73,62.37]	[60.44,66.84]
HUS3	[168.76,175.15]	[61.40,63.51]	[104.35,107.45]	[120.93,125.18]	[57.20,61.72]	[58.14,67.08]
INC1	[155.26,160.45]	[53.15,60.21]	[95.88,98.49]	[91.68,94.37]	[62.48,66.22]	[58.90,63.13]
INC2	[156.26,161.31]	[51.09,60.07]	[95.77,99.36]	[91.21,96.83]	[54.92,64.20]	[54.41,61.55]
INC3	[154.47,160.31]	[55.08,59.03]	[93.54,98.98]	[90.43,96.43]	[59.03,65.86]	[55.97,65.80]
ISA1	[164.00,168.00]	[55.01,60.03]	[120.28,123.04]	[117.52,121.02]	[54.38,57.45]	[50.80,53.25]
ISA2	[163.00,170.00]	[54.04,59.00]	[118.80,123.04]	[116.67,120.24]	[55.47,58.67]	[52.43,55.23]
ISA3	[164.01,169.01]	[55.00,59.01]	[117.38,123.11]	[116.67,122.43]	[52.80,58.31]	[52.20,55.47]
JPL1	[167.11,171.19]	[61.03,65.01]	[118.23,121.82]	[108.30,111.20]	[63.89,67.88]	[57.28,60.83]
JPL2	[169.14,173.18]	[60.07,65.07]	[118.85,120.88]	[108.98,113.17]	[62.63,69.07]	[57.38,61.62]
JPL3	[169.03,170.11]	[59.01,65.01]	[115.88,121.38]	[110.34,112.49]	[61.72,68.25]	[59.46,62.94]
KHA1	[149.34,155.54]	[54.15,59.14]	[111.95,115.75]	[105.36,111.07]	[54.20,58.14]	[48.27,50.61]
KHA2	[149.34,155.32]	[52.04,58.22]	[111.20,113.22]	[105.36,111.07]	[53.71,58.14]	[49.41,52.80]
KHA3	[150.33,157.26]	[52.09,60.21]	[109.04,112.70]	[104.74,111.07]	[55.47,60.03]	[49.20,53.41]
LOT1	[152.64,157.62]	[51.35,56.22]	[116.73,119.67]	[114.62,117.41]	[55.44,59.55]	[53.01,56.60]
LOT2	[154.64,157.62]	[52.24,56.32]	[117.52,119.67]	[114.28,117.41]	[57.63,60.61]	[54.41,57.98]
LOT3	[154.83,157.81]	[50.36,55.23]	[117.59,119.75]	[114.04,116.83]	[56.64,61.07]	[55.23,57.80]
PHI1	[163.08,167.07]	[66.03,68.07]	[115.26,119.60]	[116.10,121.02]	[60.96,65.30]	[57.01,59.82]
PHI2	[164.00,168.03]	[65.03,68.12]	[114.55,119.60]	[115.26,120.97]	[60.96,67.27]	[55.32,61.52]
PHI3	[161.01,167.00]	[64.07,69.01]	[116.67,118.79]	[114.59,118.83]	[61.52,68.68]	[56.57,60.11]
ROM1	[167.15,171.24]	[64.07,68.07]	[123.75,126.59]	[122.92,126.37]	[51.22,54.64]	[49.65,53.71]
ROM2	[168.15,172.14]	[63.13,68.07]	[122.33,127.29]	[124.08,127.14]	[50.22,57.14]	[49.93,56.94]
ROM3	[167.11,171.19]	[63.13,68.03]	[121.62,126.57]	[122.58,127.78]	[49.41,57.28]	[50.99,60.46]

Apêndice K

Conjunto de dados simbólicos de natureza intervalar Fluxos de água

Tabela K.1: Conjunto de dados simbólicos de natureza intervalar Fluxos de água

Dias do ano	IF1	IF2	IF3	...	IF46	IF47	IF48	Categoria
1	[1.67,2.17]	[1.33,2.17]	[1.67,2]	...	[1,1.17]	[1,1.33]	[1.83,2]	Final de semana
2	[1.83,2.17]	[1.5,2.17]	[1.5,1.67]	...	[1.67,2.17]	[1.67,1.83]	[2.17,2.17]	Dia útil
3	[1.5,1.5]	[1.17,1.5]	[0.5,1.17]	...	[1.5,1.67]	[1.17,1.33]	[0.83,1.83]	Dia útil
4	[1.83,2]	[1.33,1.67]	[0.67,1.33]	...	[0.83,1.17]	[0.83,1.17]	[0.67,1.5]	Final de semana
5	[0.5,0.67]	[0.33,1]	[0.17,0.5]	...	[6.67,6.83]	[6.83,6.83]	[6.5,6.67]	Dia útil
6	[6.67,7]	[6.33,7.17]	[6.33,6.67]	...	[1.33,1.83]	[1.33,1.83]	[1,1.5]	Dia útil
7	[1.83,2.5]	[1,1.17]	[0.83,1.17]	...	[1.83,6.67]	[1,2]	[1.67,2]	Dia útil
8	[2.17,2.33]	[1.67,2.5]	[0.83,1.5]	...	[0.67,1.33]	[1.17,1.67]	[0.83,1.67]	Final de semana
9	[1.67,2.33]	[1.5,2.17]	[0.33,1]	...	[0.67,1.17]	[0.67,0.83]	[0.5,1.5]	Final de semana
10	[1.67,2]	[1.17,2]	[0.5,1.33]	...	[0.5,1]	[0.83,1]	[0.33,1.5]	Dia útil
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
312	[0.67,1]	[0.67,1.17]	[0.33,0.83]	...	[0.5,0.83]	[0.67,1.33]	[0.5,0.67]	Dia útil
313	[0.5,1]	[0.33,1.17]	[0.33,0.5]	...	[0.83,1]	[0.67,0.83]	[0.5,0.83]	Dia útil
314	[0.33,1]	[0.67,1.33]	[0.33,0.33]	...	[0.67,0.83]	[0.5,1.17]	[0.5,1.33]	Final de semana
315	[0.5,1]	[0.33,1]	[0.5,0.5]	...	[0.5,0.83]	[0.33,0.5]	[0.33,0.67]	Final de semana
316	[0.33,1]	[0.83,1.33]	[0.67,0.83]	...	[1.17,1.5]	[1.17,1.33]	[1.17,1.33]	Dia útil

Apêndice L

Conjunto de dados simbólicos de natureza intervalar Fórmula 1

Tabela L.1: Conjunto de dados simbólicos de natureza intervalar Fórmula 1

Equipes	Posição	Num_pontos	Ano_Nascim	Altura	Peso	Ano_inicio_F1
Renault	[1,5]	[58,133]	[1973,1981]	[171,172]	[68,70]	[1996,2001]
Mclaren	[2,20]	[4,112]	[1971,1979]	[168,186]	[70,74]	[1997,2001]
Ferrari	[3,8]	[38,62]	[1969,1972]	[172,174]	[74,79]	[1991,1993]
Toyota	[6,7]	[43,45]	[1974,1975]	[173,178]	[60,73]	[1997,1997]
Williams	[10,22]	[2,36]	[1976,1980]	[174,185]	[68,75]	[2000,2003]
Red bull	[12,24]	[1,24]	[1971,1983]	[169,182]	[68,75]	[1994,2005]
Sauber	[13,14]	[9,11]	[1971,1981]	[161,168]	[58,63]	[1996,2002]
BAR	[9,23]	[1,37]	[1977,1980]	[163,173]	[60,68]	[2000,2002]
Minardi	[21,21]	[3,3]	[1980,1980]	[178,178]	[68,68]	[2005,2005]
Jordan	[16,18]	[5,7]	[1976,1977]	[172,174]	[64,66]	[2001,2005]

Num_GP	Poles	Num_vitórias_2005	Títulos_pilotos	Pódios_2005
[19,19]	[1,5]	[1,7]	Não conquistou	Subiu - 1º lugar
[1,19]	[0,5]	[0,7]	Conquistou	Subiu - 1º lugar
[19,19]	[0,1]	[0,1]	Conquistou	Subiu - 1º lugar
[18,19]	[1,1]	[0,0]	Não conquistou	Subiu - não 1º lugar
[14,19]	[0,1]	[0,0]	Conquistou	Subiu - não 1º lugar
[1,19]	[0,0]	[0,0]	Não conquistou	Não subiu
[19,19]	[0,0]	[0,0]	Não conquistou	Não subiu
[16,18]	[0,1]	[0,0]	Não conquistou	Subiu - não 1º lugar
[11,11]	[0,0]	[0,0]	Não conquistou	Não subiu
[19,19]	[0,0]	[0,0]	Não conquistou	Não subiu

Apêndice M

Conjunto de dados simbólicos de natureza intervalar Fungos

Tabela M.1: Conjunto de dados simbólicos de natureza intervalar Fungos

Espécies	Gênero	Largura do píleo	Largura do estipe	Espessura do estipe	Altura dos esporos	Largura dos esporos
Amanita aprica	Amanita	[5,15]	[3.3,9.1]	[1.4,3.5]	[9.5,13]	[6.5,8.5]
Amanita constricta	Amanita	[6,12]	[9,17]	[1,2]	[9.5,11.5]	[8.5,10]
Amanita Gemmata	Amanita	[3,11]	[4,15]	[0.5,2]	[8,11.5]	[6,9]
Amanita Lanei	Amanita	[8,25]	[10,20]	[1.5,4]	[8,11]	[5,6]
Amanita Muscaria	Amanita	[6,39]	[7,16]	[2,3]	[9,13]	[6.5,9.5]
Amanita Novinupta	Amanita	[5,14]	[6,12]	[1.5,3.5]	[7,8.5]	[5.5,6]
Amanita Ocreata	Amanita	[5,13]	[10,22]	[1.5,3]	[9,12.5]	[7,9]
Amanita Pachycolea	Amanita	[8,18]	[10,25]	[1,3]	[11.5,14]	[10,12]
Amanita Pantherina	Amanita	[4,15]	[7,11]	[1,2.5]	[9.5,13]	[7,9.5]
Amanita Phalloides	Amanita	[3.5,15]	[4,18]	[1,3]	[7,12]	[6,10]
Amanita Protecta	Amanita	[4,14]	[5,15]	[1,3]	[10,12]	[8.5,10]
Amanita Smithiana	Amanita	[5,17]	[6,18]	[1,3.5]	[8.5,12]	[6,8]
Amanita Vaginata	Amanita	[5.5,10]	[6,13]	[1.2,2]	[8,11.5]	[7.5,10]
Amanita Velosa	Amanita	[5,11]	[4,11]	[1,2.5]	[8.5,12]	[7,11]
Agaricus albolutescens	Agaricus	[6,12]	[2,7]	[1.5,3]	[6,7.5]	[4,5]
Agaricus fissuratus	Agaricus	[6,21]	[4,14]	[1,3.5]	[6.5,9]	[4.5,6]
Agaricus fissuratus	Agaricus	[6,32]	[10,37]	[6,6]	[7.5,10.5]	[5,6.5]
Agaricus benesii	Agaricus	[4,8]	[5,1]	[1,2]	[5,6]	[3,4]
Agaricus bernardii	Agaricus	[7,16]	[4,7]	[3,4.5]	[5.5,7]	[5.5,6.5]
Agaricus bisporus	Agaricus	[5,12]	[2.5,2.5]	[1.5,2.5]	[6,8.5]	[4.5,6]
Agaricus bitorquis	Agaricus	[5,15]	[4,10]	[2,4]	[5,6.5]	[4.5,5]
Agaricus californicus	Agaricus	[4,11]	[3,7]	[0.4,1]	[5,7.5]	[4.5,5]
Agaricus campestris	Agaricus	[5,10]	[3,6]	[1,2]	[5.5,8]	[3.5,5]
Agaricus comtulus	Agaricus	[2.5,4]	[3,5]	[0.4,0.7]	[4,5]	[3.5,5]
Agaricus cupreo-brunneus	Agaricus	[2.5,6]	[1.5,3.5]	[1,1.5]	[7,8]	[5,6]
Agaricus fuscofibrillosus	Agaricus	[4,15]	[4,15]	[1.5,2.5]	[5,6.5]	[3.5,4]
Agaricus fuscovelatus	Agaricus	[3.5,8]	[4,10]	[1,2]	[7,8]	[5,6]
Agaricus hondensis	Agaricus	[7,14]	[8,14]	[1.5,2.5]	[4,6]	[3,4.5]
Amanita augusta	Amanita	[4,12]	[5,15]	[1,2]	[8,12]	[6,8]
Amanita calyptroderma	Amanita	[8,25]	[10,20]	[1.5,4]	[8,11]	[5,6]
Agaricus lilaceps	Agaricus	[8,20]	[9,19]	[3,5]	[5,6.5]	[4,5]
Agaricus micromegathus	Agaricus	[2.5,4]	[2.5,4.5]	[4,7]	[4.5,5]	[3,3.5]
Agaricus pattersonae	Agaricus	[5,15]	[6,15]	[2.5,3.5]	[7,9]	[6,6.5]
Agaricus perobscurus	Agaricus	[8,12]	[6,12]	[1.5,2]	[6.5,8]	[4.5,5]
Agaricus praeclaresquamosus	Agaricus	[7,19]	[8,15]	[2,3.5]	[4,6]	[3.5,4.5]
Agaricus semotus	Agaricus	[2,6]	[3,7]	[0.4,0.8]	[4.5,5.5]	[3,3.5]
Agaricus sylvicola	Agaricus	[6,12]	[6,12]	[1.5,2]	[5.5,6.5]	[3.5,4]
Agaricus smithii	Agaricus	[7,12]	[5,12]	[2,3]	[7,9]	[5,5.5]
Agaricus subrutilescens	Agaricus	[6,14]	[6,16]	[1,2]	[4,6]	[3.5,4.5]
Agaricus xanthodermus	Agaricus	[5,17]	[4,14]	[1,3.5]	[5,6]	[4,5.5]
Boletus amygdalinus	Boletus	[4,10]	[4,7]	[1.5,3]	[11,14]	[5,6.5]
Boletus appendiculatus	Boletus	[7,14]	[5,9]	[3,6]	[11.5,13.5]	[3.5,4.5]
Boletus chrysenteron	Boletus	[4,7]	[5,10]	[1,1.5]	[11.5,14]	[4,6]
Boletus citriniporus	Boletus	[4,8]	[4,7]	[1,3]	[12,13.5]	[3.75,4.5]
Boletus dryophilus	Boletus	[4,12]	[4,8]	[1,2.5]	[11.5,16]	[5,6.5]
Boletus edulis	Boletus	[7,25]	[7,20]	[3,8]	[12,17]	[4,6]
Boletus flaviporus	Boletus	[6,11]	[6,12]	[1,2]	[12,15]	[5,6]
Boletus orovillus	Boletus	[8,15]	[5,9]	[2.5,4.5]	[5.5,6.5]	[3.5,4]
Boletus regineus	Boletus	[7,14]	[7,13]	[3,4]	[11.5,13.5]	[3.5,4.5]
Boletus rex-veris	Boletus	[9,18]	[5,10]	[2,6]	[12.5,18]	[4,5]
Boletus rubripes	Boletus	[6,16]	[6,15]	[3,5]	[12,16.5]	[4.5,5.5]
Boletus smithii	Boletus	[7,15]	[7,15]	[3.5,7]	[13.5,16]	[4.5,5.5]
Boletus subtomentosus	Boletus	[4,12]	[4,8]	[1,2]	[10,15]	[4,5]
Boletus truncatus	Boletus	[5,10]	[5,10]	[1.5,2.5]	[12,15]	[4.5,6]
Boletus zelleri	Boletus	[4,11]	[5,10]	[1,3]	[11,15]	[4,6]

Apêndice N

Conjunto de dados simbólicos de natureza intervalar Iris

Tabela N.1: Conjunto de dados simbólicos de natureza intervalar Iris

Observações	Espécies	Comprimento da sépala	Largura da sépala	Comprimento da pétala	Largura da pétala
w1	S	[4.6,5.1]	[3.0,3.6]	[1.3,1.5]	[0.2,0.2]
w2	S	[4.4,5.4]	[2.9,3.9]	[1.4,1.7]	[0.1,0.4
w3	S	[4.3,5.8]	[3.0,4.0]	[1.1,1.6]	[0.1,0.2]
w4	S	[5.1,5.7]	[3.5,4.4]	[1.3,1.7]	[0.3,0.4]
w5	S	[4.6,5.4]	[3.3,3.7]	[1.0,1.9]	[0.2,0.5]
w6	S	[4.7,5.2]	[3.0,3.5]	[1.4,1.6]	[0.2,0.4]
w7	S	[4.8,5.5]	[3.1,4.2]	[1.4,1.6]	[0.1,0.4]
w8	S	[4.4,5.5]	[3.0,3.5]	[1.3,1.5]	[0.1,0.2]
w9	S	[4.4,5.1]	[2.3,3.8]	[1.3,1.9]	[0.2,0.6]
w10	S	[4.6,5.3]	[3.0,3.8]	[1.4,1.6]	[0.2,0.3]
w11	Ve	[5.5,7.0]	[2.3,3.2]	[4.0,4.9]	[1.3,1.5]
w12	Ve	[4.9,6.6]	[2.4,3.3]	[3.3,4.7]	[1.0,1.6]
w13	Ve	[5.0,6.1]	[2.0,3.0]	[3.5,4.7]	[1.0,1.5]
w14	Ve	[5.6,6.7]	[2.2,3.1]	[3.9,4.5]	[1.0,1.5]
w15	Ve	[5.9,6.4]	[2.5,3.2]	[4.0,4.9]	[1.2,1.8]
w16	Ve	[5.7,6.8]	[2.6,3.0]	[3.5,5.0]	[1.0,1.7]
w17	Ve	[5.4,6.0]	[2.4,3.0]	[3.7,5.1]	[1.0,1.6]
w18	Ve	[5.5,6.7]	[2.3,3.4]	[4.0,4.7]	[1.3,1.6]
w19	Ve	[4.0,6.1]	[2.3,3.0]	[3.3,4.6]	[1.0,1.4]
w20	Ve	[4.1,6.2]	[2.5,3.0]	[3.0,4.3]	[1.1,1.3]
w21	Vi	[5.8,7.1]	[2.7,3.3]	[5.1,6.0]	[1.8,2.5]
w22	Vi	[4.9,7.6]	[2.5,3.6]	[4.5,6.6]	[1.7,2.5]
w23	Vi	[5.7,6.8]	[2.5,3.2]	[5.0,5.5]	[1.9,2.4]
w24	Vi	[6.0,7.7]	[2.2,3.8]	[5.0,6.9]	[1.5,2.3]
w25	Vi	[5.6,7.7]	[2.7,3.3]	[4.9,6.7]	[1.8,2.3]
w26	Vi	[6.1,7.2]	[2.8,3.2]	[4.8,6.0]	[1.6,2.1]
w27	Vi	[6.1,7.9]	[2.6,3.8]	[5.1,6.4]	[1.4,2.2]
w28	Vi	[6.0,7.7]	[3.0,3.4]	[4.8,6.1]	[1.8,2.4]
w29	Vi	[5.8,6.9]	[2.7,3.3]	[5.1,5.9]	[1.9,2.5]
w30	Vi	[5.0,6.7]	[2.5,3.4]	[5.0,5.4]	[1.8,2.3]

Apêndice O

Conjunto de dados simbólicos de natureza intervalar Peixes

Tabela O.1: Conjunto de dados simbólicos de natureza intervalar Peixes

Espécie	Comprimento	Peso	Músculo	Intestino	Estômago	Brânquias	Fígado
Ageneiosusbrevifili	[22.50, 35.5]	[170,625]	[1425,5043]	[333, 2980.06]	[0, 1761.1]	[393.71 , 853.1]	[642, 7105.77]
Cynodongibbus	[19.00,32]	[77,359]	[2393,8737]	[0,2653]	[478.34 , 10860.7]	[354.22 , 1976.38]	[2684.83 ,43014]
Hopliasaïmara	[25.50 ,63]	[340,5500]	[1269.65 ,5937]	[454.5 ,2181]	[443,1122]	[270, 823.23]	[822, 6271.15]
Potamotrygonhystrix	[20.50 ,45]	[400,6250]	[664, 936.46]	[0, 921.44]	[0, 464.63]	[0,0]	[1331.45 , 7014.86]
Leporinusfasciatus	[18.80 ,25]	[125,273]	[1231,2013]	[0,0]	[181.02 , 243.06]	[0, 145.19]	[463.18 , 1124.99]
Leporinusfrederici	[23.00, 24.5]	[290,350]	[255,1085]	[45, 261.2]	[43.48 , 625.2]	[20, 41.3]	[242.45 , 617.07]
Dorasmicropoeus	[19.20 ,31]	[128,505]	[813,1947]	[0, 1430.82]	[0, 1294.93]	[85.57 , 199.83]	[2157.69 , 10491.1]
Platydorascostatus	[13.70 ,25]	[60,413]	[317,923]	[206.62 , 649.67]	[0,288]	[38.16 , 186.22]	[1216,6218]
Pseudoancistrusbarbatus	[13.00, 20.5]	[55,210]	[93,114]	[0, 216.95]	[55.87 , 127.67]	[0,0]	[332, 1727.65]
Semaprochilodusvari	[22.00,28]	[330,700]	[293.52 , 556.61]	[143.19 ,933]	[0, 694.98]	[55.12 , 98.65]	[5283.84 , 22362.9]
Acnodonoligacanthus	[10.00, 16.2]	[34.9 , 154.7]	[26,115]	[0,95]	[26.09 , 178.34]	[2.8 , 147.71]	[101.26 ,807]
Myleusrubripinis	[12.30 ,18]	[80,275]	[8,35]	[0,0]	[10.76 , 41.93]	[0, 9.45]	[190.12 , 394.52]

Rins	Fígado/músculo	Rins/músulo	Brânquias/músculo	Intestino/músculo	Estômago/músculo	Categoria
[0, 3969.05]	[0.45 , 1.41]	[0, 2.02]	[0.15 , 0.3]	[0.23 , 0.63]	[0, 0.55]	Carnívoros
[1437.82 , 27514.6]	[1.12 , 4.92]	[0.6 , 3.24]	[0.15 , 0.24]	[0, 0.5]	[0.2 , 1.24]	Carnívoros
[1071,22015]	[0.65 , 2.45]	[0.84 , 4.97]	[0.14 , 0.22]	[0.11 , 0.49]	[0.09 , 0.4]	Carnívoros
[1435.93 , 8053.03]	[1.93 , 7.49]	[2.16 , 8.6]	[0,0]	[0, 1.25]	[0, 0.5]	Carnívoros
[963.17 , 2845.22]	[0.38 , 0.56]	[0.73 , 1.41]	[0, 0.07]	[0,0]	[0.12 , 0.17]	Onívoros
[113, 896.26]	[0.44 , 0.96]	[0.39 , 0.83]	[0.04 , 0.08]	[0.18 , 0.24]	[0.13 , 0.58]	Onívoros
[0, 19600.9]	[2.24 , 8.36]	[0, 14.68]	[0.09 , 0.17]	[0, 1.48]	[0, 0.79]	Detritívoros
[709,7366]	[1.51 , 11.28]	[0.98 , 15.67]	[0.05 , 0.28]	[0.3 , 1.45]	[0, 0.61]	Detritívoros
[0, 42.34]	[3.53 , 17.11]	[0, 0.37]	[0,0]	[0, 2.31]	[0.49 , 1.36]	Detritívoros
[935.49 , 2805.21]	[14.82 , 52.6]	[3.05 , 6.77]	[0.17 , 0.25]	[0.4 , 1.68]	[0, 1.25]	Detritívoros
[153, 566.69]	[1.33 , 18.34]	[1.71 , 21.8]	[0.06 , 3.36]	[0, 2.16]	[0.23 , 5.97]	Herbívoros
[72.3 , 112.54]	[7.12 , 30.35]	[2.42 , 10.23]	[0, 0.85]	[0,0]	[0.31 , 4.33]	Herbívoros

Apêndice P

Conjunto de dados simbólicos de natureza intervalar Temperatura das cidades

Tabela P.1: Conjunto de dados simbólicos de natureza intervalar Temperatura das cidades

Cidades	Categoria	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Amssterdam	2	[-4,4]	[-5,3]	[2,12]	[5,15]	[7,17]	[10,20]	[10,20]	[12,23]	[10,20]	[5,15]	[1,10]	[-1,4]
Athens	2	[6,12]	[6,12]	[8,16]	[11,19]	[16,25]	[19,29]	[22,32]	[22,32]	[19,28]	[16,23]	[11,18]	[8,14]
Bahrain	1	[13,19]	[14,19]	[17,23]	[21,27]	[25,32]	[28,34]	[29,36]	[30,36]	[28,34]	[24,31]	[20,26]	[15,21]
Bombay	1	[19,28]	[19,28]	[22,30]	[24,32]	[27,33]	[26,32]	[25,30]	[25,30]	[24,30]	[24,32]	[23,32]	[20,30]
Cairo	1	[8,20]	[9,22]	[11,25]	[14,29]	[17,33]	[20,35]	[22,36]	[22,35]	[20,33]	[18,31]	[14,26]	[10,20]
Calcutta	1	[13,27]	[16,29]	[21,34]	[24,36]	[26,36]	[26,33]	[26,32]	[26,32]	[26,32]	[24,32]	[18,29]	[13,26]
Colombo	1	[22,30]	[22,30]	[23,31]	[24,31]	[25,31]	[25,30]	[25,29]	[25,29]	[25,30]	[24,29]	[23,29]	[22,30]
Copenhagen	2	[-2,2]	[-3,2]	[-1,5]	[3,10]	[8,16]	[11,20]	[14,22]	[14,21]	[11,18]	[7,12]	[3,7]	[1,4]
Dubal	1	[13,23]	[14,24]	[17,28]	[19,31]	[22,34]	[25,36]	[28,39]	[28,39]	[25,37]	[21,34]	[17,30]	[14,26]
Frankfurt	2	[-10,9]	[-8,16]	[-4,17]	[0,24]	[3,27]	[7,30]	[8,32]	[8,31]	[5,27]	[0,22]	[-3,14]	[-8,10]
Geneva	2	[-3,5]	[-6,6]	[3,9]	[7,13]	[10,17]	[15,17]	[16,24]	[16,23]	[11,19]	[6,13]	[3,8]	[-2,6]
HongKong	1	[13,17]	[12,16]	[15,19]	[19,23]	[22,27]	[25,29]	[25,30]	[25,30]	[25,29]	[22,27]	[18,23]	[14,19]
KulaLumpur	1	[22,31]	[23,32]	[23,33]	[23,33]	[23,32]	[23,32]	[23,31]	[23,32]	[23,32]	[23,31]	[23,31]	[23,31]
Lisbon	2	[8,13]	[8,14]	[8,16]	[11,18]	[13,21]	[16,24]	[17,26]	[18,27]	[17,24]	[14,21]	[11,17]	[8,14]
London	2	[2,6]	[2,7]	[3,10]	[5,13]	[8,17]	[11,20]	[13,22]	[13,21]	[11,19]	[8,14]	[5,10]	[3,7]
Madras	1	[20,30]	[20,31]	[22,33]	[26,35]	[28,39]	[27,38]	[26,36]	[26,35]	[25,34]	[24,32]	[22,30]	[21,29]
Madrid	2	[1,9]	[1,12]	[3,16]	[6,19]	[9,24]	[13,29]	[16,34]	[16,33]	[13,28]	[8,20]	[4,14]	[1,9]
Manila	1	[21,27]	[22,27]	[24,29]	[24,31]	[25,31]	[25,31]	[23,29]	[24,28]	[25,28]	[24,29]	[22,28]	[22,27]
Mauritius	3	[22,28]	[22,29]	[22,29]	[21,28]	[19,25]	[18,24]	[17,23]	[17,23]	[17,24]	[18,25]	[19,27]	[21,28]
MexicoCity	1	[6,22]	[15,23]	[17,25]	[18,27]	[18,27]	[18,27]	[18,27]	[18,26]	[18,26]	[16,25]	[14,25]	[8,23]
Moscow	2	[-13,-6]	[-15,-12]	[-8,0]	[0,8]	[7,18]	[11,23]	[13,24]	[11,22]	[6,16]	[1,8]	[-5,0]	[-11,-5]
Munich	2	[-6,1]	[-5,3]	[-2,9]	[3,14]	[7,18]	[10,21]	[12,23]	[11,23]	[8,20]	[4,13]	[0,7]	[-4,2]
Nairobi	1	[12,25]	[13,26]	[14,25]	[14,24]	[13,22]	[12,21]	[11,21]	[11,21]	[11,24]	[13,24]	[12,23]	[13,23]
NewDelhi	1	[6,21]	[10,24]	[14,29]	[20,36]	[26,40]	[28,39]	[27,35]	[26,34]	[24,34]	[18,34]	[11,28]	[7,23]
NewYork	2	[-2,4]	[-3,4]	[1,9]	[6,15]	[12,22]	[17,27]	[21,29]	[18,20]	[16,24]	[11,19]	[5,12]	[-2,6]
Paris	2	[1,7]	[1,7]	[2,12]	[5,16]	[8,19]	[12,22]	[14,24]	[13,24]	[11,21]	[7,16]	[4,10]	[1,6]
Rome	2	[4,11]	[5,13]	[7,16]	[10,19]	[13,23]	[17,28]	[20,31]	[20,31]	[17,27]	[13,21]	[9,16]	[5,12]
SanFrancisco	2	[6,13]	[6,14]	[7,17]	[8,18]	[10,19]	[11,21]	[12,22]	[12,22]	[12,23]	[11,22]	[8,18]	[6,14]
Seoul	2	[0,7]	[1,6]	[1,8]	[6,16]	[12,22]	[16,25]	[18,31]	[16,30]	[9,28]	[3,24]	[7,19]	[1,8]
Singapore	1	[23,30]	[23,30]	[24,31]	[24,31]	[24,30]	[25,30]	[25,30]	[25,30]	[24,30]	[24,30]	[24,30]	[23,30]
Stockholm	2	[-9,-5]	[-9,-6]	[-4,-2]	[1,8]	[6,15]	[11,19]	[14,22]	[13,20]	[9,15]	[5,9]	[1,4]	[-2,2]
Sydney	1	[20,30]	[20,30]	[18,26]	[16,23]	[12,20]	[5,17]	[8,16]	[9,17]	[11,20]	[13,22]	[16,26]	[20,30]
Tehran	4	[0,5]	[5,8]	[10,15]	[15,18]	[20,25]	[28,30]	[36,38]	[38,40]	[29,30]	[18,20]	[9,12]	[-5,0]
Tokyo	2	[0,9]	[0,10]	[3,13]	[9,18]	[14,23]	[18,25]	[22,29]	[23,31]	[20,27]	[13,21]	[8,16]	[2,12]
Toronto	2	[-8,-1]	[-8,-1]	[-4,4]	[-2,11]	[-8,18]	[13,24]	[16,27]	[16,26]	[12,22]	[6,14]	[-1,17]	[-5,1]
Vienna	2	[-2,1]	[-1,3]	[1,8]	[5,14]	[10,19]	[13,22]	[15,24]	[14,23]	[11,19]	[7,13]	[2,7]	[1,3]
Zurich	2	[-11,9]	[-8,15]	[-7,18]	[-1,21]	[2,27]	[6,30]	[10,31]	[8,25]	[5,23]	[3,22]	[0,19]	[-11,8]

Apêndice Q

Resultados Complementares

Tabela Q.1: Desempenho dos métodos propostos nos conjuntos de dados simbólicos de natureza intervalar: Índice de Rand Ajustado (IRA) e taxa total de erro de alocação (TEA) das melhores soluções

Métodos de agrupamento	Carros				Desenvolvimento dos países			
	IRA		TEA		IRA		TEA	
iSBC	**0,3817**	**(2)**	**0,3636**	**(2)**	**0,2800**	**(1)**	**0,2000**	**(1)**
iKSBC1C	0,2819	(6)	0,3939	(6)	**0,2800**	**(1)**	**0,2000**	**(1)**
iKSBC2C	**0,4204**	**(1)**	**0,2727**	**(1)**	**0,2800**	**(1)**	**0,2000**	**(1)**
iKM	0,2329	(7)	0,4848	(7)	-0,0766	(8)	0,5000	(8)
iKM + iSBC	0,3541	(3)	**0,3636**	**(2)**	-0,0766	(8)	0,5000	(8)
iKM + iKSBC1C	0,3541	(3)	**0,3636**	**(2)**	-0,0766	(8)	0,5000	(8)
iKM + iKSBC2C	0,3541	(3)	**0,3636**	**(2)**	-0,0766	(8)	0,5000	(8)
iKKM	-0,0141	(10)	0,6667	(10)	0,0000	(4)	0,4000	(4)
iKKM + iSBC	-0,0029	(8)	0,6364	(8)	0,0000	(4)	0,4000	(4)
iKKM + iKSBC1C	-0,0029	(8)	0,6364	(8)	0,0000	(4)	0,4000	(4)
iKKM + iKSBC2C	-0,0141	(10)	0,6667	(10)	0,0000	(4)	0,4000	(4)
Métodos de agrupamento	Facedata				Fluxos de água			
	IRA		TEA		IRA		TEA	
iSBC	**1,0000**	**(1)**	**0,0000**	**(1)**	0,0205	(9)	0,2943	(3)
iKSBC1C	**1,0000**	**(1)**	**0,0000**	**(1)**	0,0349	(8)	0,2943	(3)
iKSBC2C	**1,0000**	**(1)**	**0,0000**	**(1)**	0,0410	(7)	0,2943	(3)
iKM	0,7783	(10)	0,1250	(10)	**0,0995**	**(2)**	0,3323	(8)
iKM + iSBC	**1,0000**	**(1)**	**0,0000**	**(1)**	**0,0995**	**(2)**	0,3323	(8)
iKM + iKSBC1C	**1,0000**	**(1)**	**0,0000**	**(1)**	**0,0995**	**(2)**	0,3323	(8)
iKM + iKSBC2C	**1,0000**	**(1)**	**0,0000**	**(1)**	**0,0995**	**(2)**	0,3323	(8)
iKKM	0,7783	(10)	0,2500	(10)	-0,0037	(10)	0,2943	(3)
iKKM + iSBC	**1,0000**	**(1)**	**0,0000**	**(1)**	-0,0113	(11)	0,2943	(3)
iKKM + iKSBC1C	**1,0000**	**(1)**	**0,0000**	**(1)**	0,0857	(6)	**0,2911**	**(2)**
iKKM + iKSBC2C	**1,0000**	**(1)**	**0,0000**	**(1)**	**0,1363**	**(1)**	**0,2690**	**(1)**

... continuação da Tabela Q.1

Métodos de agrupamento	Fórmula 1 - Títulos dos pilotos				Fórmula 1 - Pódios no ano de 2005			
	IRA		TEA		IRA		TEA	
iSBC	0,0483	(7)	0,3000	(5)	0,1223	(6)	0,4000	(6)
iKSBC1C	-0,0714	(10)	0,3000	(5)	**0,3478**	**(1)**	**0,3000**	**(1)**
iKSBC2C	0,0483	(7)	0,3000	(5)	**0,3478**	**(1)**	**0,3000**	**(1)**
iKM	0,0708	(5)	0,3000	(5)	**0,3478**	**(1)**	**0,3000**	**(1)**
iKM + iSBC	0,0483	(7)	0,3000	(5)	0,1223	(6)	0,4000	(6)
iKM + iKSBC1C	0,0708	(5)	0,3000	(5)	**0,3478**	**(1)**	**0,3000**	**(1)**
iKM + iKSBC2C	**0,2857**	**(1)**	**0,2000**	**(1)**	**0,3478**	**(1)**	**0,3000**	**(1)**
iKKM	**0,2593**	**(2)**	**0,2000**	**(1)**	-0,0372	(11)	0,5000	(11)
iKKM + iSBC	-0,1111	(11)	0,3000	(5)	0,1223	(6)	0,4000	(6)
iKKM + iKSBC1C	**0,2593**	**(2)**	**0,2000**	**(1)**	0,1223	(6)	0,4000	(6)
iKKM + iKSBC2C	**0,2593**	**(2)**	**0,2000**	**(1)**	0,1223	(6)	0,4000	(6)
Métodos de agrupamento	Fungos				Iris			
	IRA		TEA		IRA		TEA	
iSBC	**0,4887**	**(2)**	**0,2545**	**(2)**	**1,0000**	**(1)**	**0,0000**	**(1)**
iKSBC1C	**0,7761**	**(1)**	**0,0727**	**(1)**	**1,0000**	**(1)**	**0,0000**	**(1)**
iKSBC2C	0,4564	(3)	0,2727	(3)	**1,0000**	**(1)**	**0,0000**	**(1)**
iKM	0,2417	(7)	0,3636	(7)	**1,0000**	**(1)**	**0,0000**	**(1)**
iKM + iSBC	0,0526	(10)	0,5091	(10)	**1,0000**	**(1)**	**0,0000**	**(1)**
iKM + iKSBC1C	0,3921	(5)	0,3091	(5)	**1,0000**	**(1)**	**0,0000**	**(1)**
iKM + iKSBC2C	0,4265	(4)	0,2909	(4)	**1,0000**	**(1)**	**0,0000**	**(1)**
iKKM	0,1623	(8)	0,4182	(8)	0,5797	(11)	0,2000	(11)
iKKM + iSBC	0,0169	(11)	0,5273	(11)	**1,0000**	**(1)**	**0,0000**	**(1)**
iKKM + iKSBC1C	0,2268	(6)	0,3273	(6)	**1,0000**	**(1)**	**0,0000**	**(1)**
iKKM + iKSBC2C	0,1497	(9)	0,4364	(9)	**1,0000**	**(1)**	**0,0000**	**(1)**
Métodos de agrupamento	Peixes				Temperatura das cidades			
	IRA		TEA		IRA		TEA	
iSBC	0,1172	(3)	**0,3333**	**(2)**	0,3372	(7)	0,4324	(7)
iKSBC1C	**0,1951**	**(2)**	0,4167	(3)	0,4804	(3)	0,2973	(3)
iKSBC2C	**0,4535**	**(1)**	**0,2500**	**(1)**	0,3935	(4)	0,3243	(4)
iKM	0,0331	(5)	0,5000	(6)	**0,4834**	**(2)**	**0,2703**	**(1)**
iKM + iSBC	0,0331	(5)	0,5000	(6)	0,3387	(6)	0,4324	(7)
iKM + iKSBC1C	0,0522	(4)	0,4167	(3)	**0,5655**	**(1)**	**0,2703**	**(1)**
iKM + iKSBC2C	0,0331	(5)	0,5000	(6)	0,3459	(5)	0,3514	(5)
iKKM	0,0209	(8)	0,4167	(3)	-0,0356	(9)	0,4324	(7)
iKKM + iSBC	-0,0942	(11)	0,5000	(6)	-0,0488	(11)	0,4595	(10)
iKKM + iKSBC1C	-0,0366	(9)	0,5000	(6)	0,0207	(8)	0,3784	(6)
iKKM + iKSBC2C	-0,0366	(9)	0,5000	(6)	-0,0468	(10)	0,4595	(10)