Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Rotulação de Grupos em Algoritmos de Agrupamento Baseados em Distância Utilizando Grau de Pertinência

Francisco das Chagas Imperes Filho

Teresina-PI, Abril de 2018

Francisco das Chagas Imperes Filho

Rotulação de Grupos em Algoritmos de Agrupamento Baseados em Distância Utilizando Grau de Pertinência

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Piauí (área de concentração: Computação Aplicada), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Vinicius Ponte Machado

Teresina-PI

Abril de 2018

Francisco das Chagas Imperes Filho

Rotulação de Grupos em Algoritmos de Agrupamento Baseados em Distância Utilizando Grau de Pertinência/ Francisco das Chagas Imperes Filho. – Teresina-PI, Abril de 2018-

62 p. : il.

Orientador: Vinicius Ponte Machado

Qualificação (Mestrado) – Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação, Abril de 2018.

1. Rotulação de grupos de dados. 2. Agrupamento de dados. 3. Aprendizagem de Máquina I. Vinícius Ponte Machado. II. Universidade Federal do Piauí. III. Programa de Pós-Graduação em Ciência da Computação. IV. Rotulação de Grupos em Algoritmos de Agrupamento Baseados em Distância Utilizando Grau de Pertinência

CDU 02:141:005.7

??

Francisco das Chagas Imperes Filho

Rotulação de Grupos em Algoritmos de Agrupamento Baseados em Distância Utilizando Grau de Pertinência

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Piauí (área de concentração: Computação Aplicada), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho Aprovado. Teresina-PI. 13 de abril de 2018.

Vinicius Ponte Machado
Orientador

Kelson Romulo Teixeira Aires
Interno

Rodrigo de Melo Souza Veras
Interno

Ricardo Augusto Souza Fernandes
Externo

Teresina-PI
Abril de 2018

Aos meus pais, esposa, filhos, irmãos e amigos.

Agradecimentos

Agradeço a Deus.

Agradeço ao meu pai, Francisco Imperes (in memory) e, especialmente a minha mãe Gonçala de Abreu, pelo amor, carinho e por sempre está ao meu lado nos momentos em que precisei.

Aos meus irmãos, Rosemere, Renato e Roberto, por sempre me apoiarem em todas as fases de minha vida.

A minha esposa, Izabel de Macêdo, pelo amor, atenção, zelo, apoio e, principalmente, pela paciência durante a execução desse projeto.

Aos meus filhos, Monalisa e Isaac Imperes, por me darem alegrias, paz e descontração em todos os momentos de minha vida.

Aos amigos e colegas de curso e de trabalho, pela paciência, apoio e colaboração para o desenvolvimento deste projeto.

Agradeço ao meu orientador, Vinícius Ponte Machado, por todos os conselhos, pela paciência e ajuda antes e durante o período de pós-graduação.

"A humildade é uma
das maiores virtudes
do ser humano."
(Autor Desconhecido)

Resumo

O agrupamento de dados vem sendo considerado um item relevante na subárea de Aprendizagem de Máquina (AM), mais especificamente Aprendizagem de Máquina Não Supervisionada. Por esse motivo, nos últimos anos este tópico vem ganhando destaque no campo da Inteligência Artificial (IA). O problema relacionado ao agrupamento (*clustering*) é abordado com frequência em muitos trabalhos, e a compreensão dos grupos (*clusters*) é tão importante quanto a sua formação. Definir grupos pode auxiliar na interpretação e, consequentemente, direcionar esforços para tomada de decisão levando em consideração as peculiaridades de cada grupo formado. As interpretações dos grupos podem ser bastante úteis quando é necessário saber o que torna um elemento pertencente a um grupo, quais as principais características de um grupo, quais as diferenças e similaridades entre os grupos, entre outras situações. Devido a problemática relacionada a encontrar definições, ou rótulos, capazes de identificar facilmente os grupos formados, este trabalho descreve um modelo que elabora rótulos para encontrar características relevantes nos elementos de cada grupo e identificá-los de forma única. A proposta está dividida em duas partes. Na primeira o modelo transforma a saída padrão de um algoritmo de agrupamento não supervisionado baseado em distância em Grau de Pertinência (GP). Nessa etapa cada elemento da base de dados analisada recebe um GP em relação a cada grupo formado. Na segunda, os elementos com seus respectivos GPs são utilizados para formular faixas de valores para os rótulos. Estes, por sua vez, são capazes de identificar grupos de forma única em bases de dados bem difundidas na literatura. O método foi submetido a uma análise comparativa com outro modelo de rotulação que tem por objetivo identificar características únicas em grupos de dados, facilitando sua compreensão. Os rótulos produzidos pela proposta deste trabalho conseguiram representar um grande número de elementos de cada grupo, favorecendo seu entendimento. Na análise comparativa, o modelo conseguiu produzir rótulos atingindo média de percentual de acertos de $94,66\%$ nas bases de dados analisadas, permitindo uma fácil interpretação das definições geradas. Por fim, a proposta foi analisada utilizando outras bases de dados atingindo média de percentual de acertos de $92,01\%$. Os experimentos realizados demonstram que o modelo proposto é capaz de construir rótulos para a identificação de grupos, melhorando a sua compreensão.

Palavras-chaves: Rotulação de Dados. Definição de Dados. Agrupamento de Dados. Aprendizagem de Máquina.

Abstract

Data grouping has been considered a relevant item in the Machine Learning (ML) subarea, more specifically Unsupervised Machine Learning. For this reason, in recent years this topic has been gaining prominence in the field of Artificial Intelligence (AI). Data grouping is frequently discussed in many papers, and to understand clusters is as important as to form them. Defining groups can assist in their interpretation and, consequently, in directing efforts to decision making, taking into account the peculiarities of each group. The interpretation of clusters can be very useful when it is necessary to know what makes an element belonging to a group, what are the main characteristics of a group, what are the differences and similarities between the groups, among other situations. Due to problems related to the finding of definitions, or labels, able to easily identify the groups, this work describes a model to find relevant characteristics in the elements of each group and to identify them uniquely. The proposal is divided in two steps. On the first step, the model transforms the standard output of an unsupervised clustering algorithm based in distance into a pertinence degree. In this step each element of the analyzed database receives a pertinence degree in relation to each formed group. In the second, the elements and their respective pertinence degrees are used to formulate ranges of values for the labels. These, in turn, are able to uniquely identify groups in databases well diffused in the literature. The method was submitted to a comparative analysis with another labeling model that aims to identify unique characteristics in data groups, facilitating their comprehension. The labels produced by the proposed model managed to represent a large number of elements in each group, favoring their understanding. In the comparative analysis, the model was able to produce labels in the analyzed databases at a success rate of 94.66%, allowing an easy interpretation of the generated definitions. Finally, the proposal was analyzed using other databases and reached a success rate of 92.01%. The experiments demonstrated that the proposed model is able to build labels for group identification, improving their understanding.

Keywords: Data Labeling. Data Definition. Data Grouping. Machine Learning.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AM	Aprendizagem de Máquina
EFD	*Equal Frequence Discretization*
EWD	*Equal Widths Discretization*
GP	Grau de Pertinência
GS	Grau de Seleção
IA	Inteligência Artificial
IGS	Incremento do Grau de Seleção
LDA	*Latent Dirichlet Allocation*
MLP	*Multilayer Perceptron*
RBD	Rotulação Baseada em Distância
RNA	Rede Neural Artificial
TEDA	*Typicality and Eccentricity Data Analytics*
TFxICF	*Term Frequency-Inverse Cluster Frequency*

Lista de símbolos

$\sum$	Letra grega Sigma maiúscula, representando somatório
χ	Letra grega Qui, representando um conjunto de k centros
k	Número de grupos ou conjunto de dados
SO_2	Dióxido de Enxofre ou Anidrido Sulfuroso

Sumário

1 Introdução

A capacidade de produção de dados gerados através dos mais variados meios, tem dificultado o processo de interpretação para muitos especialistas que tem na análise das informações seu maior recurso para tomada de decisão. Nesse contexto, prover mecanismos que possibilitem a correta interpretação e uso racional dos dados tem sido motivo de estudos para muitos pesquisadores. Uma possível forma de tratar os dados produzidos em demasia é através do agrupamento de dados, uma sub-área da Aprendizagem de Máquina Não Supervisionada. Segundo Coppin (2010), métodos que se enquadram nessa categoria aprendem sem qualquer intervenção humana. Uma tarefa comum da Aprendizagem de Máquina Não Supervisionada é o agrupamento. Esta técnica tem como objetivo selecionar um conjunto de dados e agrupá-los de acordo com alguma similaridade. Os algoritmos de agrupamento de dados foram desenvolvidos como uma ferramenta para relacionar grande quantidade de dados gerados por diferentes sistemas (RASIM et al., 2016). Diante desta problemática, esse trabalho tem como objetivo utilizar um algoritmo de agrupamento de dados baseado em distância e apresentar um modelo capaz de analisar grupos de dados e produzir rótulos para ajudar na compreensão e auxiliar especialistas no processo de tomada de decisão.

1.1 Contexto e Justificativa

O agrupamento de dados tem sido considerado um dos tópicos mais relevantes dentre aqueles existentes na área de aprendizagem de máquina e mineração de dados (AGGARWAL; REDDY, 2013). O desenvolvimento e aperfeiçoamento de algoritmos que tratam dessa temática tem sido o centro de muitas pesquisas, porém poucos trabalhos visam especificamente o estudo das definições e compreensão dos grupos.

As interpretações dos grupos podem ser bastante úteis quando é necessário saber o que torna um elemento pertencente a um grupo, quais as principais características de um grupo, quais as diferenças e similaridades entre os grupos, entre outras situações. A solução dessas questões pode ajudar na otimização de soluções ou em simples análises para saber como os dados estão distribuídos nos grupos (RIBEIRO, 2016). Como exemplo, pode-se citar uma base de dados relacionada ao perfil de conhecimento dos colaboradores sobre a área de atuação e produtos comercializados por uma determinada organização. Nesse cenário, a análise para saber quais as características e necessidades determinantes em cada grupo de funcionários pode servir para tomada de decisão dos gestores, direcionando-os para capacitação nas áreas onde forem mais deficitários.

1.2 Proposta

Este trabalho descreve um modelo capaz de analisar grupos e produzir definições que também serão chamadas de rótulos. O modelo utiliza o algoritmo de agrupamento não supervisionado baseado em distância *K-Means* como base para o desenvolvimento da proposta. Como ponto de partida, a saída padrão do algoritmo (distância) é transformada para que cada elemento da base de dados estudada receba um Grau de Pertinência (GP) em relação a cada grupo formado.

O modelo utiliza o GP para selecionar os elementos relevantes em cada grupo. Os elementos com seus respectivos GPs são utilizados para formular as faixas de valores dos rótulos. Cada faixa de valor é formada por meio da seleção dos valores máximo e mínimo de cada atributo, utilizando para isso os elementos selecionados como relevantes. Devido ao objetivo de formular rótulos únicos para cada grupo, eles devem conter pelo menos uma faixa de valor única em cada grupo. Com isso, a cada iteração o modelo seleciona elementos com GP maiores, com o objetivo de selecionar elementos mais relevantes e eliminar as interseções entre as faixas de valores, identificando desta foram cada grupo de forma única.

1.3 Objetivos

Geral

O objetivo desse trabalho é utilizar um algoritmo de agrupamento de dados baseado em distância e apresentar um modelo capaz de analisar grupos de dados e produzir rótulos para ajudar na compreensão e auxiliar especialistas no processo de tomada de decisão.

Específicos

- Transformar a saída padrão de um algoritmo de agrupamento de dados baseado em distância para grau de pertinência.
- Formular faixas de valores para cada atributo em cada grupo formado, verificar a existência de interseções entre faixas de valores, montar e exibir rótulos de cada grupo com faixas de valores que não possuem interseção.
- Avaliar o modelo por meio da quantidade de elementos que cada rótulo é capaz de representar.
- Realizar uma análise comparativa com o trabalho proposto por Ribeiro (2016) e Machado, Ribeiro e Rabelo (2015).

1.4 Estrutura do Trabalho

O trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica e as pesquisas relacionadas para melhor compreensão da proposta. O Capítulo 3 demonstra o modelo implementado. O Capítulo 4 discute os detalhes da implementação, os resultados obtidos e os rótulos gerados pelo modelo. No Capítulo 5 é apresentada uma análise comparativa com outro modelo de rotulação, tomando como referência bases de dados bastantes difundidas na literatura. O Capítulo 6 exibe as conclusões e as sugestões de trabalhos futuros que podem contribuir para a expansão do modelo proposto. Por fim, o Apêndice A apresenta resultados de análises de outras bases de dados submetidas ao processo de formulação de grupos de dados e produção de rótulos.

2 Referencial Teórico

Este capítulo apresenta o referencial teórico que serve de sustentação para o entendimento do conteúdo abordado neste trabalho. Inicialmente são apontados conceitos relacionados a Aprendizagem de Máquina, agrupamento de dados e na sequência, uma descrição do funcionamento do algoritmo de agrupamento *K-Means*. Por fim, serão apresentados os trabalhos relacionados com a pesquisa, destacando-se os que têm em suas propostas o mesmo objetivo desta preposição.

2.1 Aprendizagem de Máquina

Aprendizagem de Máquina (AM), do inglês *Machine Learning*, é uma subárea da Inteligência Artificial (IA), de acordo com Coppin (2010) IA é o estudo dos sistemas que agem de um modo que a um observador qualquer pareceria ser inteligente. Desta forma, algoritmos que se enquadram na categoria de AM tentam construir um modelo a partir de entradas específicas e usam essas entradas para fazer previsões, ao invés de seguir o conjunto fixo de instruções definidas pelo usuário (PARTH et al., 2015).

Segundo Russel e Norvig (2013), a aprendizagem de máquina deve ter a capacidade de se adaptar a novas circunstâncias, detectar e extrapolar padrões. Portanto, a AM surgiu da percepção de criar programas computacionais que aprendem um determinado comportamento ou padrão automaticamente a partir de exemplos ou observações.

A AM apresenta alguma relação com o aprendizado humano, onde seres humanos são capazes de generalizar (aprender) a partir de exemplos ou observações. Algoritmos de AM podem ser vistos como mecanismos que extraem um padrão de comportamento a partir de dados (exemplos). Tais algoritmos têm sido utilizados em várias áreas do conhecimento, como: imagens médicas (SILVA; FILHO; SILVA, 2015), segurança e detecção de intrusão (ANNA; ERHAN, 2016), bioinformática (KUN et al., 2016), redes de sensores sem fio (HANEN; RIDHA, 2017), análise de vulnerabilidade de software (GONG; KUANG; LIU, 2016), soluções para eficiência energética (MEHMET, 2015) e análise de sentimentos (EBRU; AKCAYOL, 2016).

Basicamente, aprendizagem de máquina subdivide-se em duas grandes técnicas: supervisionada e não supervisionada. A primeira utiliza dados rotulados e busca mapear entradas em saídas previamente definidas. A segunda, por sua vez, utiliza dados não rotulados e o objetivo é encontrar características semelhantes que possam agrupá-los de acordo com um padrão de similaridade. O presente trabalho foca no uso da segunda técnica.

2.2 Agrupamento (*Clustering*)

Segundo Chang, Pen e Chen (2017), agrupamento é uma técnica de análise de dados estatísticos que agrupa dados que possuem atributos semelhantes em grupos. Ela é usada em vários contextos, incluindo aprendizado de máquinas, bioinformática, processamento de imagem, descoberta de conhecimento e reconhecimento de padrões.

O agrupamento de dados é um mecanismo muito utilizado para a análise de dados. Basicamente, os algoritmos de agrupamento dividem os dados em grupos chamados de *clusters*, de modo que a similaridade *intra-cluster* é maximizada e a similaridade *inter-cluster* é minimizada. Um método de agrupamento é caracterizado principalmente por sua escolha de medida de similaridade (ATILGAN; NASIBOV, 2016).

O objetivo do agrupamento é encontrar uma atribuição de *cluster* de consenso em cada grupo, combinando padrões de similaridade. As informações de cada grupo devem conter características que os representam dentro do universo dos elementos pesquisados. Esse aspecto torna os algoritmos que se enquadram nessa categoria como excelentes métodos para agrupamento de dados.

2.3 *K-Means*

MacQueen (1967) foi o primeiro pesquisador a descrever o algoritmo *K-Means* como um processo de particionamento de uma população N-dimensional em k conjuntos que são razoavelmente eficientes no sentido da variância dentro da classe.

Linder (2009), descreve o algoritmo *K-Means* como uma heurística de agrupamento não hierárquico que busca minimizar a distância dos elementos a um conjunto de k centros dado por $\chi = \{x_1, x_2, ..., x_k\}$ de forma iterativa. A distância entre um ponto p_i e um conjunto de grupos, dada por $d(p_i, \chi)$, é definida como sendo a distância do ponto ao centro mais próximo dele. A função a ser minimizada é dada pela Equação 2.1,

$$(P, \chi) = \frac{1}{n} \sum_{i=1}^{n} d(p_i, \chi)^2 \tag{2.1}$$

onde:

P = distância do ponto ao centro mais próximo;

χ = conjunto de k centros;

n = número de pontos;

d = distância entre um ponto e um conjunto de grupos.

Por depender de um parâmetro k (número de pontos/elementos) definido de forma explícita, esta característica costuma ser um problema tendo em vista que nem sempre se

sabe quantos grupos existem a priori. Mesmo existindo outras formas/possibilidades para descrever este método de particionamento de um conjunto de dados, de forma genérica o algoritmo *K-Means* pode ser descrito conforme Pseudocódigo[1] 1.

Algoritmo 1: K-Means

```
Entrada: K
Selecione os k objetos que serão os centroides dos grupos
início
    para todos os objetos restantes faça
        Calcule a distância entre o elemento e os centroides
        Adicione o elemento ao grupo que possuir a menor distância até ele
        Recalcule o centroide do grupo
    fim
    para todos os k grupos faça
        Calcule a Soma de Quadrados Residual
    fim
    repita
        para todos os n elementos faça
            Mova os elementos para os outros agrupamentos
            Recalcule a Soma de Quadrados Residual
            se soma dos Quadrados Residual diminuiu então
                O objeto passa a fazer parte do grupo que produzir maior ganho
                Recalcule a Soma de Quadrados Residual dos grupos alterados
            fim
        fim
    até Número de iterações = i ou Não ocorra mudança de objetos;
fim
```

O *K-Means* é facilmente programado e é computacionalmente econômico, tendo sua complexidade expressa pela Fórmula 2.2,

$$O(n * K * i * d) \tag{2.2}$$

onde:

n = número de pontos;

K = número de grupos;

i = número de iterações;

d = número de atributos/características.

[1] Fonte: Fernando Sarturi Prass. Disponível em: <http://fp2.com.br/blog/index.php/2013/algoritmo-de-k-means/>. Acesso em: 23 abr. 2018.

Devido a sua complexidade, com o algoritmo *K-Means* é viável processar amostras muito grandes de dados. Algumas possíveis aplicações que podem utilizar a eficiência do algoritmo não supervisionado *K-Means* incluem métodos para agrupamento de similaridades, previsão não-linear, aproximação de distribuições multivariadas, testes não paramétricos para independência entre várias variáveis e classificação de árvores baseadas em distância (MACQUEEN, 1967).

O *K-Means* tende a convergir para uma configuração estável, na qual nenhum elemento está designado para um *cluster* cujo centro não lhe seja o mais próximo. Um exemplo da execução do algoritmo é apresentado na Figura 1. Na Figura 1(a), é colocada a situação inicial de um conjunto de dados em um espaço bidimensional. Na Figura 1(b), são atribuídos centroides aos grupos de forma aleatória ($k = 3$). Na Figura 1(c), cada elemento foi designado para um dos três centroides iniciais. Na Figura 1(d), os centroides de cada grupo foram calculados e os elementos foram designados para os grupos cujos centroides lhe estão mais próximos. Finalizando na Figura 1(e), os centroides foram recalculados e apresenta os grupos já em sua forma final. Caso não estivessem, os passos demonstrados na Figura 1(d) e 1(e) seriam repetidos até que houvesse uma convergência total.

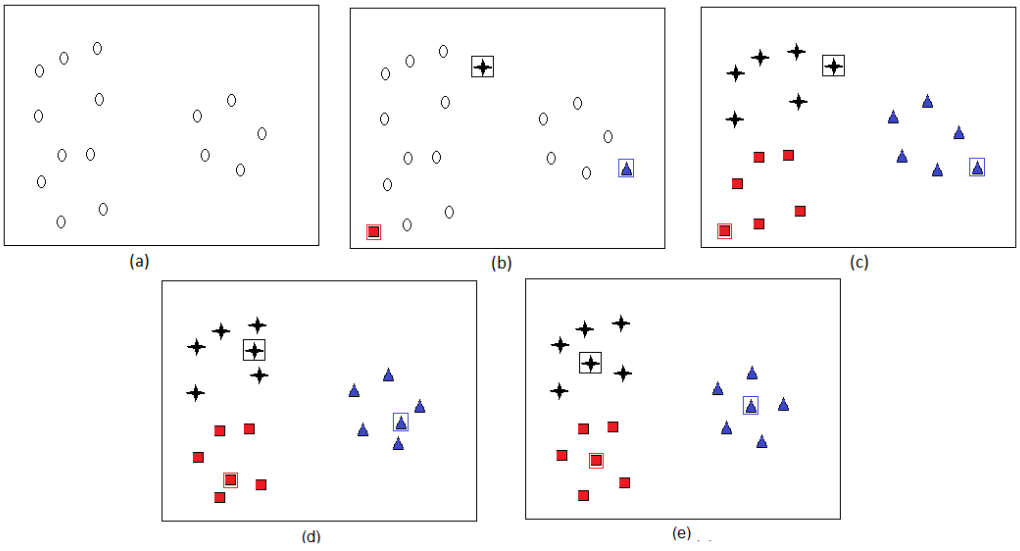

Figura 1 – Exemplo de execução do algoritmo *K-Means*.

A convergência estável eventualmente pode gerar um problema que enfatiza a questão da homogeneidade e ignora a importante questão da boa separação dos grupos. Isto pode causar uma má separação dos grupos no caso de uma má inicialização dos centroides, inicialização esta que é feita de forma arbitrária (aleatória) no início da execução. A Figura 2 demonstra o efeito de uma má inicialização na execução do algoritmo *K-Means*. Na Figura 2(a) existem três grupos naturais no conjunto de dados, um deles bem afastado dos demais. Na Figura 2(b), dois grupos foram atribuídos ao centroide representado por um quadrado. O problema é que como o terceiro grupo está bem separado dos outros dois, se dois centroides forem inicializados daquele lado do plano os elementos ficarão vinculados a

estes centroides, separando um grupo natural do conjunto de dados (LINDER, 2009).

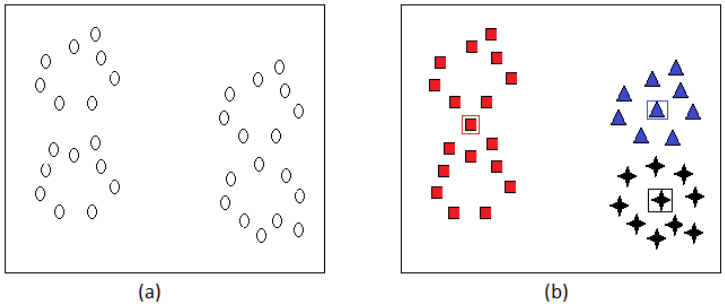

Figura 2 – Exemplo de má inicialização na execução do algoritmo *K-Means*.

Outro ponto que pode afetar a qualidade dos resultados é a escolha do número de grupos feita pelo usuário. Um número pequeno demais pode causar a junção de dois grupos naturais, enquanto que um número grande demais pode fazer com que um grupo natural seja quebrado equivocadamente em dois.

2.4 Trabalhos Relacionados

O trabalho de Lopes, Machado e Rabelo (2014) discorre sobre a mesma problemática dessa investigação: formula rótulos para um conjunto de grupos fornecidos. Os rótulos têm o objetivo de representar os elementos, facilitando a compreensão dos grupos. Os grupos geralmente são formados por algoritmos de aprendizagem não supervisionada.

O modelo presente no trabalho de Lopes, Machado e Rabelo (2014) utiliza o algoritmo *K-Means*, porém, como os autores afirmam, outros algoritmos de agrupamento podem ser utilizados. A metodologia da proposta é descrita da seguinte forma: de posse de uma base de dados como entrada, um algoritmo com aprendizagem não-supervisionada é aplicado com o objetivo de formar grupos a partir dos elementos inicialmente fornecidos. Para cada grupo formado um segundo algoritmo, desta vez com aprendizagem supervisionada, é utilizado para a identificação de possíveis características importantes. Adicionalmente, faz-se uso de um método de discretização e de algumas estratégias de decisões necessárias para a concretização da abordagem subdividida em 4 etapas. A Figura 3 representa o fluxograma de execução do modelo proposto por Lopes, Machado e Rabelo (2014).

A discretização dos dados é realizada caso a base de dados possua valores contínuos, caso contrário, os valores não são alterados. O principal propósito de utilizar um método de discretização consiste em permitir a inferência de um conjunto de valores para uma determinada característica de um rótulo. Dessa forma, um *cluster* não é limitado a ser representado por apenas um valor em um determinado atributo mas sim por um

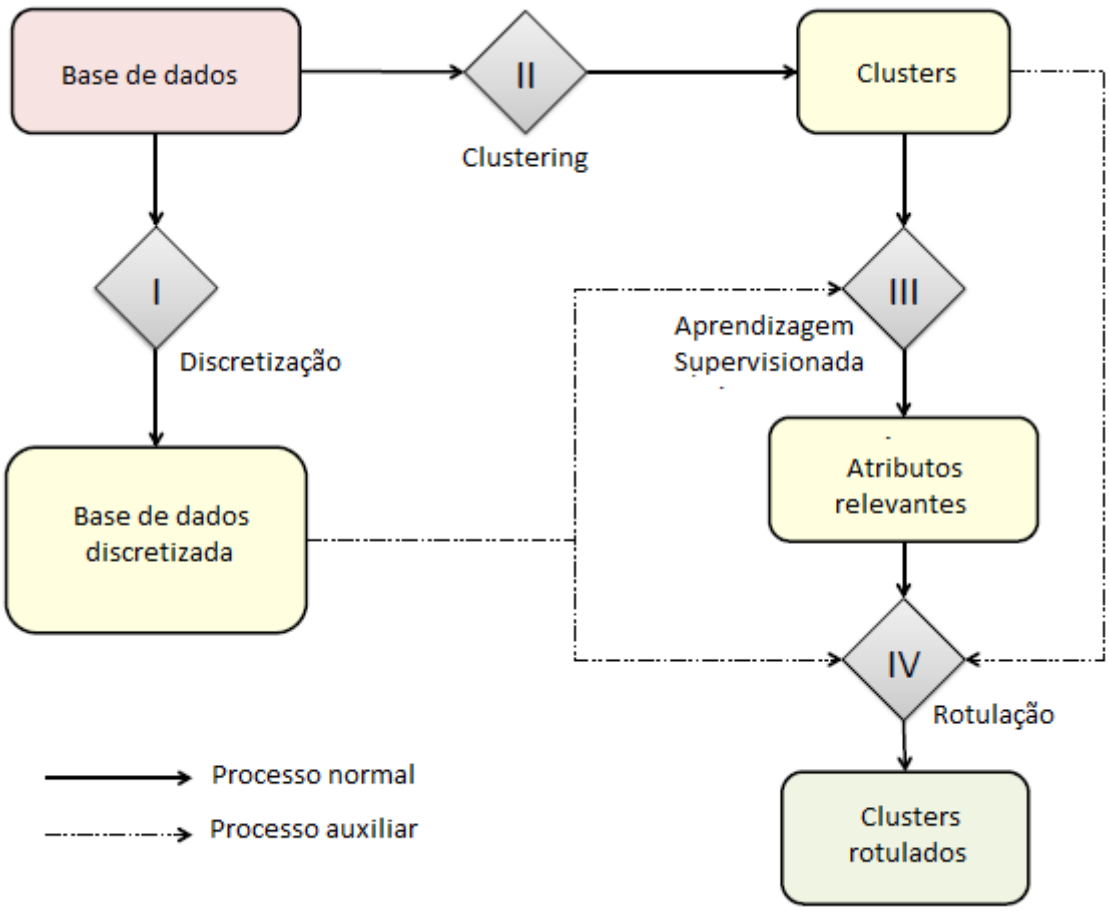

Figura 3 – Fluxograma do modelo proposto por Lopes, Machado e Rabelo (2014).

intervalo de valores. Os autores afirmam que a escolha do método de discretização pode alterar significativamente os resultados gerados. Para aplicar esse procedimento foram utilizadas duas abordagens não supervisionadas: Discretização por Larguras Iguais (*Equal Widths Discretization - EWD*), que consiste em dividir um intervalo em medidas iguais; Discretização por Frequências Iguais (*Equal Frequence Discretization - EFD*), que particiona o intervalo de acordo com a frequência dos valores de atributo. Para regular essas abordagens o modelo apresenta o parâmetro R (número de faixas de valores), o qual especifica a quantidade de faixas de valores que cada atributo terá.

Após discretizados, o modelo submete os dados a um algoritmo supervisonado - Rede Neural Artificial (RNA) do tipo *Multilayer Perceptron* (MLP) - que tem o objetivo de extrair os atributos mais relevantes. A RNA tenta inferir o valor de um atributo, dados todos os valores dos outros atributos. O processo de execução das redes neurais é realizado várias vezes, sendo determinado pelo parâmetro M, também chamado de Número de Iterações Por Atributo. Depois desse processo os atributos que tiveram um número maior de valores inferidos de forma correta são ranqueados e selecionados. Esses atributos são escolhidos de acordo com o parâmetro Variação (V). Os atributos que possuem um valor acima desse parâmetro farão parte dos rótulos. Logo depois à seleção dos atributos, os

valores mais presentes neles também são selecionados para fazer parte do rótulo. Caso a base de dados seja discretizada o rótulo é composto do valor que mais se repete. O resultado de saída do modelo consiste em valores ou faixas de valores associados a seus respectivos atributos.

Os trabalhos realizados por Ribeiro (2016) e Machado, Ribeiro e Rabelo (2015) também seguem a mesma linha de pesquisa da proposta defendida neste trabalho. Neles são apresentados um modelo de rotulação capaz de identificar características únicas em cada grupo, podendo assim, facilitar a sua compreensão. As propostas focam na teoria de conjuntos *fuzzy* para encontrar características relevantes nos elementos de cada grupo e modelar faixas de valores que identificam os grupos de forma única. O algoritmo não-supervisionado *Fuzzy C-Means* foi utilizado para formular faixas de valores em cada atributo, verificar a existência de interseções entre as faixas de valores e montar os rótulos de cada grupo com faixas de valores que não possuem interseção.

Para execução do modelo, o *Fuzzy C-Means* é inicializado com a quantidade de grupos a serem formados (três) e como saída o algoritmo retorna uma matriz U. Esta matriz atribui a cada elemento um grau de pertinência em cada um dos grupos formados. O grau de pertinência é atribuído de tal forma que, quanto mais próximo o elemento estiver de um grupo, maior é seu grau de pertinência em relação ao grupo.

Ribeiro (2016) e Machado, Ribeiro e Rabelo (2015) ressaltam em seus trabalhos que além da base de dados o modelo necessita da definição de dois parâmetros: o Grau de Seleção (GS) e o Incremento do Grau de Seleção (IGS). O primeiro consiste em um número que serve de base para a seleção dos elementos mais significativos na formulação do rótulo, ou seja, são escolhidos os elementos que possuem um grau de pertinência maior que o parâmetro GS. Com isto, em cada grupo selecionado são extraídos os valores máximo e o mínimo de cada atributo. Esses valores correspondem às faixas de valores de cada grupo. O segundo, por sua vez, consiste em um valor de incremento do parâmetro GS a cada iteração. Ambos podem variar entre os valores 0 e 1, inclusive. O fluxograma exibido na Figura 4 representa o modelo proposto pelos pesquisadores.

De acordo com Ribeiro (2016) a proposta mostrou-se promissora conseguindo representar com êxito os elementos das bases de dados estudadas. Por fim, acrescenta que em uma análise comparativa com o modelo idealizado por Lopes, Machado e Rabelo (2014), obteve bons resultados utilizando como métrica a média da menor porcentagem de acerto em cada grupo, como também na contagem total de erros.

A Tabela 1 apresenta um resumo dos trabalhos relacionados que abordam a mesma temática desta proposta, formular uma abordagem capaz de rotular grupos com o propósito de auxiliar especialistas no processo de tomada de decisão. Entretanto, apenas a proposta de Lopes, Machado e Rabelo (2014) utiliza algoritmos de aprendizagem de máquina supervisionada e não supervisionada para atingir seus objetivos. Vale acrescentar

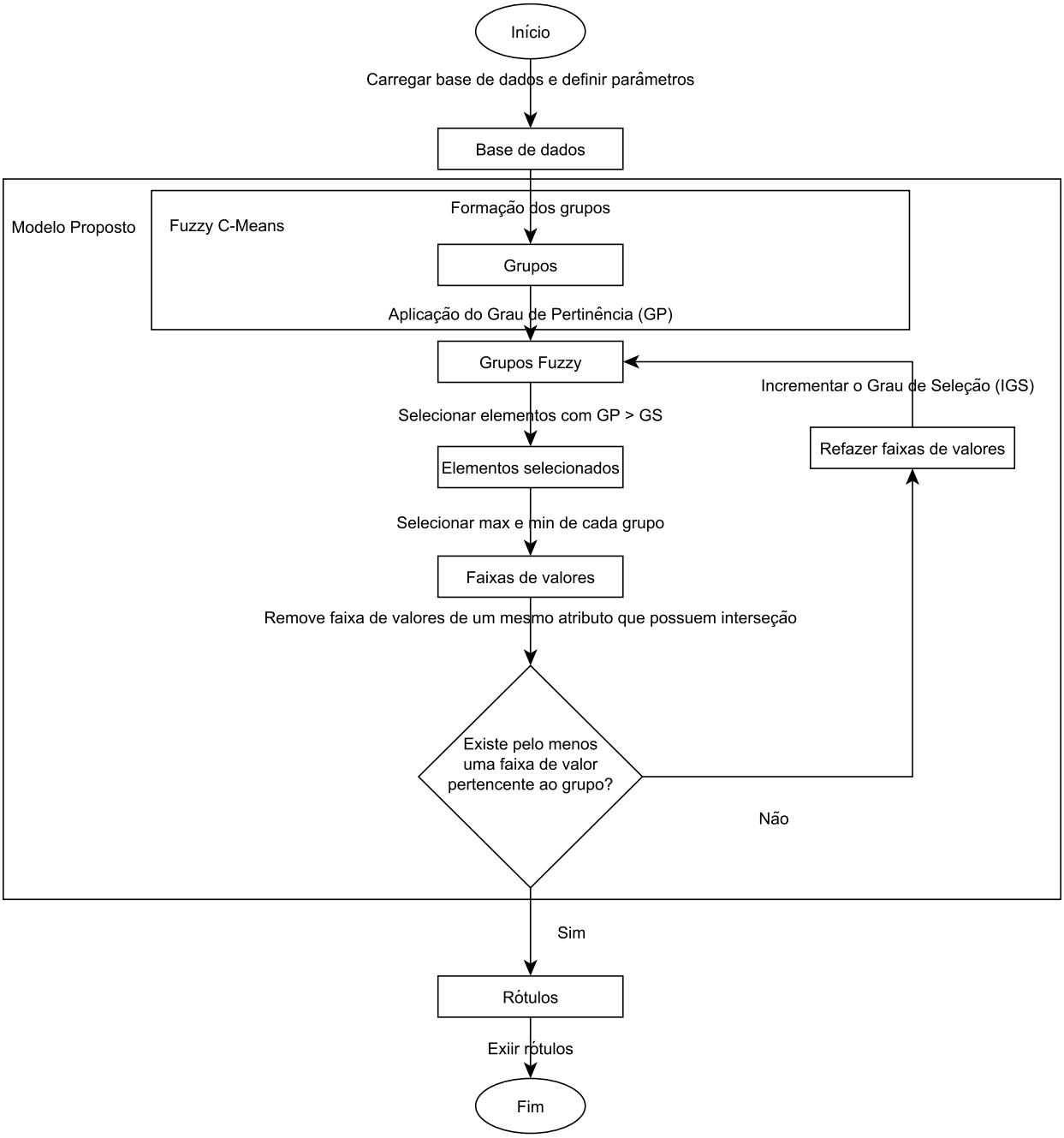

Figura 4 – Fluxograma do modelo proposto por Ribeiro (2016).

que ela também usa, quando necessário, discretização de dados como um processo que antecede a rotulação dos grupos.

Tabela 1 – Trabalhos Relacionados.

TRABALHOS RELACIONADOS	OBJETIVO	METODOLOGIA	ALGORITMO(S) / TÉCNICA(S)	DISCRETIZAÇÃO DOS DADOS
Lopes, Machado e Rabelo (2014)	Apresentar uma abordagem capaz de rotular grupos com base em suas características.	Utilização de um algoritmo não supervisionado para gerar grupos e em seguida rotular conjutos de dados com um algoritmo supervisionado.	K-Means, Rede Neural Artificial e os métodos de Discretização por Larguras Iguais e por Frequências Iguais.	Sim
Ribeiro (2016)	Apresentar um modelo de rotulação capaz de identificar características únicas em cada grupo.	Indentificação de características únicas nos grupos formados.	Fuzzy C-Means.	Não
Modelo proposto	Utilizar um algoritmo de agrupamento de dados baseado em distância e apresentar um modelo capaz de analisar grupos de dados e produzir rótulos.	Transformar saídas baseadas em distância para grau de pertinência, em seguida formular faixas de valores e montar rótulos de cada grupo formado.	K-Means.	Não

Por esses motivos é possível observar que o modelo defendido neste trabalho e a pesquisa idealizada por Ribeiro (2016), levam vantagens em relação ao modelo proposto por Lopes, Machado e Rabelo (2014) quando menciona-se os dois itens descritos anteriormente.

Prováveis justificativas para tal afirmação seriam: a) por utilizar somente uma técnica de aprendizagem de máquina (supervisionada), a submissão de repositórios de dados com grande volume de elementos pode melhorar o desempenho devido ao uso otimizado dos recursos computacionais; e b) a não utilização de processos de discretização minimiza o tempo dispendido na fase de pré-processamento dos dados.

3 Proposta do Modelo de Rotulação

Este capítulo descreve o modelo de rotulação proposto neste trabalho. O objetivo é apresentar uma abordagem capaz de analisar grupos de dados e produzir rótulos para ajudar na compreensão e auxiliar especialistas no processo de tomada de decisão. Para facilitar a compreensão, uma base de dados de testes tendo os valores de seus atributos incluídos no conjunto do números reais é utilizada para demonstrar o funcionamento do modelo.

3.1 Modelo de Rotulação

Através do fluxograma visualizado na Figura 5, o modelo é apresentado e as etapas de seu funcionamento são descritas no decorrer desta seção.

Para maior clareza, a descrição do funcionamento da proposta é feita em duas etapas. A primeira tem como meta transformar saídas baseadas em distância para grau de pertinência e a segunda formular faixas de valores, montar e exibir os rótulos de cada grupo formado.

A inicialização do modelo se dá pelo carregamento da base de dados e definição dos seguintes parâmetros: k, número de grupos a serem formados; Grau de Seleção (GS), valor que serve de base para seleção dos elementos mais significativos na formulação dos rótulos. São escolhidos os elementos que possuem um grau de pertinência maior que o parâmetro GS. Com isto, em cada grupo formado são extraídos os valores máximo e o mínimo de cada atributo. Esses valores correspondem às faixas de valores de cada grupo. Caso exista faixas de valores de um mesmo atributo possuindo interseção, essas faixas são removidas e, nesta situação, o parâmetro Incremento do Grau de Seleção (IGS) é evocado e o processo continua. O IGS consiste em um valor de incremento do parâmetro GS a cada iteração. Sua função é prover a formulação de novas faixas de valores até que seja atingida a condição de parada do modelo: existindo pelo menos uma faixa de valores de um atributo sem interseção em cada grupo, o processo é encerrado e estas faixas de valores servem como rótulo para seus grupos. Os valores para os parâmetros GS e IGS podem variar entre 0 e 1.

Nos testes e validação do modelo foi utilizado como aporte uma base de dados com 30 elementos divididos em 3 classes e os valores dos atributos foram representados por números reais. Por possuir 3 classes, o parâmetro k foi definido com o valor 3. O GS definido como 0,5 por representar um grau de seleção intermediário, no qual elementos que estão abaixo desse valor têm grandes chances de pertencerem a dois grupos; e o IGS

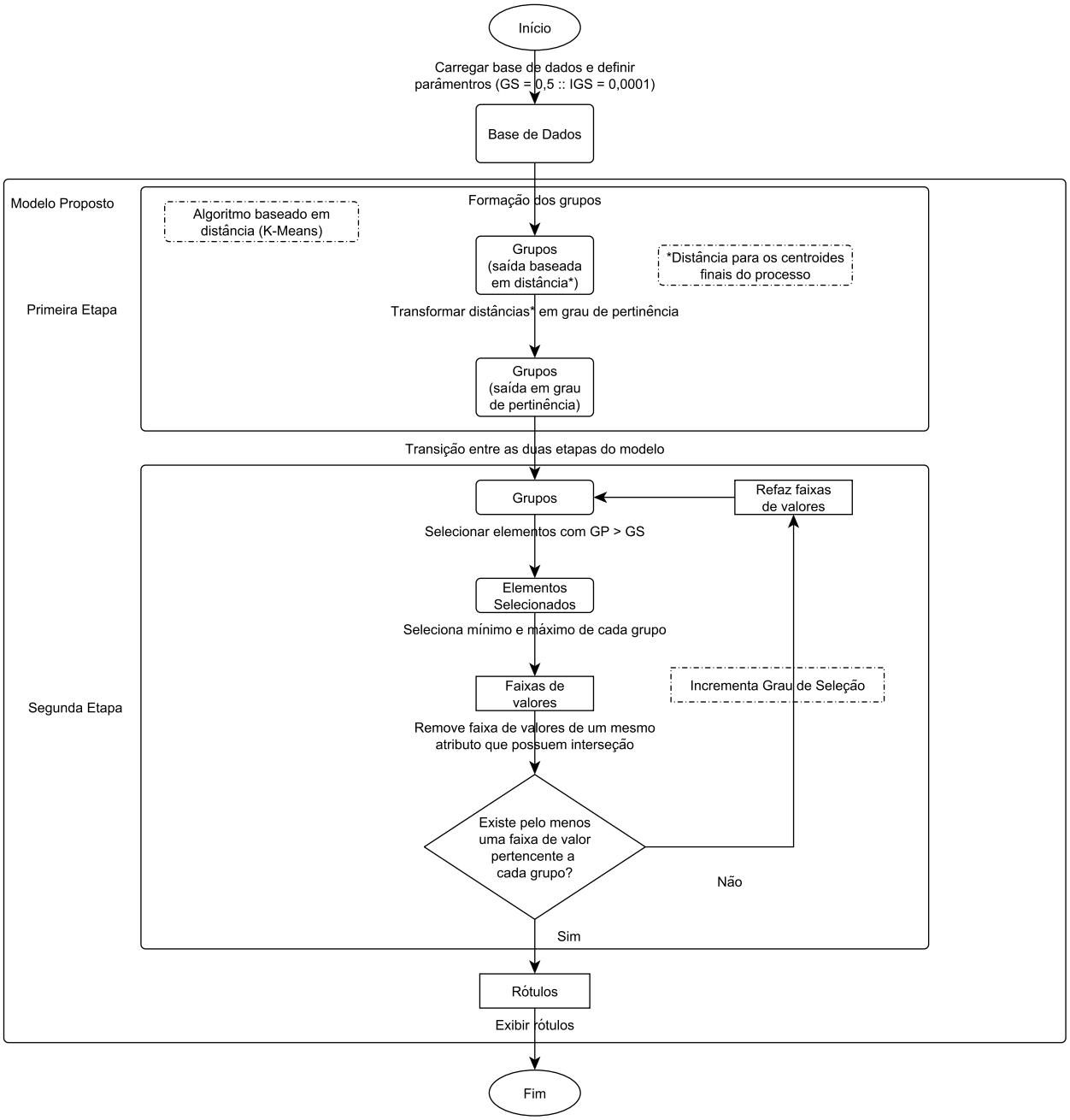

Figura 5 – Fluxograma do modelo de rotulação proposto.

definido como 0,0001. As justificativas para formulação desse último valor são: a) tanto as distâncias geradas pelo algoritmo *K-Means* e a respectiva conversão em grau de pertinência possuem quatro casas decimais; b) o incremento provoca no GS um ajuste na quarta casa decimal, facilitando a seleção dos elementos que compõem as faixas de valores.

3.1.1 Proposta - Primeira Etapa

A primeira etapa da proposta tem como objetivo transformar saídas baseadas em distância para Grau de Pertinência (GP). Paralelamente à formação dos grupos, são extraídas as distâncias de cada elemento para cada um dos grupos formados. Ao final dessa extração e tendo acesso às distâncias, ocorre a transformação dessas em grau de pertinência.

A base de dados de testes é apresentada na Tabela 2. Os 30 elementos são compostos pelos atributos At.1 e At.2 e possuem valores definidos nos conjuntos dos números reais. Os valores foram escolhidos de forma arbitrária para ilustrar a aplicação da proposta. A coluna Id representa o identificador de cada elemento na base de dados.

Tabela 2 – Base de dados de testes - Fonte: Ribeiro (2016).

Id	**At.1**	**At.2**	**Id**	**At.1**	**At.2**
1	4,3	6,0	**16**	8,9	4,0
2	9,7	6,5	**17**	7,8	7,7
3	4,7	5,7	**18**	7,8	4,6
4	7,0	8,0	**19**	3,9	3,6
5	4,3	4,7	**20**	8,5	5,1
6	4,9	4,8	**21**	7,6	6,9
7	3,1	5,7	**22**	5,9	8,5
8	4,1	5,5	**23**	9,1	6,4
9	5,8	7,5	**24**	7,0	7,3
10	9,4	6,0	**25**	6,3	7,4
11	9,8	3,5	**26**	5,3	3,8
12	8,0	4,0	**27**	7,5	8,2
13	9,0	6,7	**28**	5,9	6,9
14	4,5	3,7	**29**	3,8	4,6
15	8,8	4,5	**30**	6,8	7,9

Na Tabela 3 são informados os centroides e é exibida a saída padrão do algoritmo *K-Means* com os grupos e as respectivas distâncias de cada elemento para cada centroide de cada grupo.

Para encontrar o Grau de Pertinência a partir da distância, são executados alguns passos. Primeiro é necessário obter o inverso das distâncias de cada elemento para cada grupo formado. A Tabela 4 apresenta o resultado desse processo.

Neste trabalho o inverso das disntâncias é denotado por (I). A fórmula para definir I é dada por $\frac{1}{d_i}$, onde d_i é a distância de cada elemento para cada grupo i, com i variando

Tabela 3 – Saída padrão do algoritmo *K-Means*: grupos e distâncias de cada elemento para cada centroide de cada grupo formado.

				Distância		
Id	**At.1**	**At.2**	**Grupo**	**Grupo 1**	**Grupo 2**	**Grupo 3**
1	4,3	6,0	1	1,1900	4,6815	2,9510
2	9,7	6,5	2	5,6678	1,5865	3,1497
3	4,7	5,7	1	0,9799	4,2385	2,8229
4	7,0	8,0	3	4,1857	3,4419	0,4410
5	4,3	4,7	1	0,1105	4,6201	3,8258
6	4,9	4,8	1	0,6101	4,0136	3,3865
7	3,1	5,7	1	1,4860	5,8279	4,1377
8	4,1	5,5	1	0,7157	4,8142	3,4077
9	5,8	7,5	3	3,0848	3,9022	0,9688
10	9,4	6,0	2	5,2467	1,0034	3,1027
11	9,8	3,5	2	5,6636	1,8620	5,1282
12	8,0	4,0	2	3,7974	1,4446	3,8359
13	9,0	6,7	2	5,0751	1,5732	2,4254
14	4,5	3,7	1	1,1297	4,6265	4,5335
15	8,8	4,5	2	4,5206	0,6379	3,7361
16	8,9	4,0	2	4,6806	1,1300	4,2138
17	7,8	7,7	3	4,5467	2,7955	1,0424
18	7,8	4,6	2	3,5163	1,2210	3,2035
19	3,9	3,6	1	1,2713	5,2289	4,9417
20	8,5	5,1	2	4,2200	0,4011	3,0706
21	7,6	6,9	3	3,9146	2,1961	1,1129
22	5,9	8,5	3	4,0259	4,5119	1,2233
23	9,1	6,4	2	5,0660	1,2857	2,6436
24	7,0	7,3	3	3,6802	2,8843	0,4080
25	6,3	7,4	3	3,2784	3,4515	0,5143
26	5,3	3,8	1	1,4284	3,8378	4,0988
27	7,5	8,2	3	4,6686	3,3742	0,9341
28	5,9	6,9	3	2,6382	3,4832	1,1281
29	3,8	4,6	1	0,5331	5,1275	4,2359
30	6,8	7,9	3	3,9810	3,4760	0,2729

1 a k. Após obter os inversos das distâncias do elemento para cada grupo, deve-se somar os resultados de I para encontrar o valor inverso total do elemento em relação às suas distância, definido pela fórmula $D = \frac{1}{d_{j,i}}$. Sendo o número de elementos representado por j, com j variando de 1 a n e i variando 1 a k, o somatório dos inversos das distâncias I_j é dada pela Equação 3.1,

$$I_j = \sum_{i=1}^{k} \frac{1}{d_{j,i}}, \tag{3.1}$$

onde:

I_j = somatório dos inversos das distâncias de cada grupo i para cada elemento j;

j = número de elementos na base de dados, com j variando de 1 a n;

i = quantidade de grupos, com i variando de 1 a k;

$d_{j,i}$ = distância de cada elemento j para cada grupo i.

Tabela 4 – Inverso das distâncias de cada elemento para cada grupo formado.

	Inverso			
Id	**Grupo 1**	**Grupo 2**	**Grupo 3**	**G1+G2+G3**
1	0,8403	0,2136	0,3389	1,3928
2	0,1764	0,6303	0,3175	1,1242
3	1,0205	0,2359	0,3542	1,6107
4	0,2389	0,2905	2,2676	2,7970
5	9,0498	0,2164	0,2614	9,5276
6	1,6391	0,2492	0,2953	2,1835
7	0,6729	0,1716	0,2417	1,0862
8	1,3972	0,2077	0,2935	1,8984
9	0,3242	0,2563	1,0322	1,6126
10	0,1906	0,9966	0,3223	1,5095
11	0,1766	0,5371	0,1950	0,9086
12	0,2633	0,6922	0,2607	1,2163
13	0,1970	0,6356	0,4123	1,2450
14	0,8852	0,2161	0,2206	1,3219
15	0,2212	1,5676	0,2677	2,0565
16	0,2136	0,8850	0,2373	1,3359
17	0,2199	0,3577	0,9593	1,5370
18	0,2844	0,8190	0,3122	1,4155
19	0,7866	0,1912	0,2024	1,1802
20	0,2370	2,4931	0,3257	3,0558
21	0,2555	0,4554	0,8986	1,6094
22	0,2484	0,2216	0,8175	1,2875
23	0,1974	0,7778	0,3783	1,3535
24	0,2717	0,3467	2,4510	3,0694
25	0,3050	0,2897	1,9444	2,5391
26	0,7001	0,2606	0,2440	1,2046
27	0,2142	0,2964	1,0705	1,5811
28	0,3790	0,2871	0,8864	1,5526
29	1,8758	0,1950	0,2361	2,3069
30	0,2512	0,2877	3,6643	4,2032

A partir do somatório dos inversos das distâncias I_j de cada elemento da base (ver Tabela 4, coluna "G1+G2+G3"), é possível encontrar o grau de pertinência. Utiliza-se o valor da distância total invertida $D = \frac{1}{d_{j,i}}$ de cada grupo i de cada elemento j e dividir pela distância total I_j correspondente. O procedimento geral para determinar o grau de pertinência pode ser visualizado na Equação 3.2,

$$P_{j,i} = \frac{D}{I_{j,i}}, P_{j,i} \leq 1, \tag{3.2}$$

onde:

$P_{j,i}$ = Grau de Pertinência;

$I_{j,i}$ = somatório dos inversos das distâncias de cada elemento para cada grupo.

$d_{j,i}$ = distância de cada elemento para cada grupo formado.

Para checar os dados obtidos, basta fazer uma verificação nas Tabelas 3 e 5. Como pode ser observado na primeira linha da Tabela 3 a saída padrão do *K-Means* determina que o primeiro elemento (Id 1), com os seus respectivos atributos, pertence ao Grupo 1. Isso pode ser confirmado, já que a menor distância do elemento dentre todos os grupos formados, é o valor de 1,1900.

Tabela 5 – Grau de Pertinência a partir do somatório dos valores inversos das distâncias de cada elemento para cada grupo formado.

	Pertinência			
Id	**Grupo 1**	**Grupo 2**	**Grupo 3**	**Soma**
1	0,6033	0,1534	0,2433	1
2	0,1569	0,5607	0,2824	1
3	0,6336	0,1465	0,2199	1
4	0,0854	0,1039	0,8107	1
5	0,9498	0,0227	0,0274	1
6	0,7507	0,1141	0,1352	1
7	0,6195	0,1580	0,2225	1
8	0,7360	0,1094	0,1546	1
9	0,2010	0,1589	0,6401	1
10	0,1263	0,6602	0,2135	1
11	0,1943	0,5911	0,2146	1
12	0,2165	0,5691	0,2143	1
13	0,1583	0,5106	0,3312	1
14	0,6696	0,1635	0,1669	1
15	0,1076	0,7623	0,1302	1
16	0,1599	0,6624	0,1776	1
17	0,1431	0,2327	0,6242	1
18	0,2009	0,5786	0,2205	1
19	0,6665	0,1620	0,1715	1
20	0,0775	0,8159	0,1066	1
21	0,1587	0,2829	0,5583	1
22	0,1929	0,1721	0,6349	1
23	0,1458	0,5747	0,2795	1
24	0,0885	0,1130	0,7985	1
25	0,1201	0,1141	0,7658	1
26	0,5812	0,2163	0,2025	1
27	0,1355	0,1874	0,6771	1
28	0,2441	0,1849	0,5709	1
29	0,8131	0,0845	0,1023	1
30	0,0598	0,0684	0,8718	1

Com relação ao grau de pertinência ocorre o contrário. Quanto mais próximo de 1 for o valor atribuído ao grupo, mais próximo o elemento vai estar deste grupo. Como pode ser verificado na Tabela 5 (elemento Id 1), percebe-se que o maior valor (0,6033) está relacionado ao Grupo 1, confirmando o processo de transformação de distâncias em grau de pertinência da primeira etapa do modelo proposto.

3.1.2 Proposta - Segunda Etapa

A segunda e última etapa da proposta tem como objetivo formular faixas de valores para cada atributo em cada grupo formado, verificar a existência de interseções entre faixas de valores e, por fim, montar e exibir rótulos de cada grupo com faixas de valores que não possuem interseção.

Nesta etapa o modelo de rotulação proposto seleciona elementos que possuem um GP maior que o parâmetro GS. É importante ressaltar que o GS tem seu valor inicialmente definido como 0,5. Com isto, em cada grupo selecionado são extraídos os valores máximo e o mínimo de cada atributo. Esses valores correspondem às faixas de valores de cada grupo.

Os elementos escolhidos durante a primeira iteração do modelo podem ser vistos nas Tabelas 6, 7 e 8. As células destacadas nas tabelas mostram os valores máximos e mínimos de cada atributo utilizado na formação das faixas de valores.

Tabela 6 – Grupo 1: Elementos selecionados durante a primeira iteração (GS = 0,5).

			Grau de Pertinência		
Id	**At.1**	**At.2**	**Grupo 1**	**Grupo 2**	**Grupo 3**
1	4,3	6,0	0,6033	0,1534	0,2433
3	4,7	5,7	0,6336	0,1465	0,2199
5	4,3	4,7	0,9499	0,0227	0,0274
6	4,9	4,8	0,7507	0,1141	0,1352
7	3,1	5,7	0,6195	0,1580	0,2225
8	4,1	5,5	0,7360	0,1094	0,1546
14	4,5	3,7	0,6696	0,1635	0,1669
19	3,9	3,6	0,6665	0,1621	0,1715
26	5,3	3,8	0,5812	0,2163	0,2025
29	3,8	4,6	0,8131	0,0845	0,1023

Na Tabela 6 é possível observar que todos os elementos do Grupo 1 têm seus GPs maiores que os GPs dos outros grupos e, consequentemente, como determina o modelo de rotulação proposto, maiores que o GS definido inicialmente como 0,5. Essa mesma regra vale para os elementos dos Grupos 2 e 3, visualizados nas Tabelas 7 e 8, respectivamente.

A Tabela 9 apresenta as faixa de valores (rótulos) correspondentes a cada grupo de acordo com os valores máximo e mínimo de cada atributo, tendo por base os elementos selecionados durante a primeira iteração do modelo.

Tabela 7 – Grupo 2: Elementos selecionados durante a primeira iteração (GS = 0,5).

			Grau de Pertinência		
Id	**At.1**	**At.2**	**Grupo 1**	**Grupo 2**	**Grupo 3**
2	9,7	6,5	0,1569	0,5607	0,2824
10	9,4	6,0	0,1263	0,6602	0,2135
11	9,8	3,5	0,1943	0,5911	0,2146
12	8,0	4,0	0,2165	0,5692	0,2143
13	9,0	6,7	0,1583	0,5106	0,3312
15	8,8	4,5	0,1076	0,7623	0,1302
16	8,9	4,0	0,1599	0,6624	0,1776
18	7,8	4,6	0,2009	0,5786	0,2205
20	8,5	5,1	0,0776	0,8159	0,1066
23	9,1	6,4	0,1458	0,5747	0,2795

Tabela 8 – Grupo 3: Elementos selecionados durante a primeira iteração (GS = 0,5).

			Grau de Pertinência		
Id	**At.1**	**At.2**	**Grupo 1**	**Grupo 2**	**Grupo 3**
4	7,0	8,0	0,0854	0,1039	0,8107
9	5,8	7,5	0,2010	0,1589	0,6401
17	7,8	7,7	0,1431	0,2327	0,6242
21	7,6	6,9	0,1587	0,2829	0,5583
22	5,9	8,5	0,1929	0,1722	0,6349
24	7,0	7,3	0,0885	0,1130	0,7985
25	6,3	7,4	0,1201	0,1141	0,7658
27	7,5	8,2	0,1355	0,1874	0,6771
28	5,9	6,9	0,2441	0,1849	0,5710
30	6,8	7,9	0,0598	0,0685	0,8718

Tabela 9 – Rótulos gerados tendo como base os elementos selecionados em cada grupo: Iteração #1 - GS = 0,5.

	Grupo 1	**Grupo 2**	**Grupo 3**
At. 1	3,1 ∼ 5,3	7,8 ∼ 9,8	5,8 ∼ 7,8
At. 2	3,6 ∼ 6,0	3,5 ∼ 6,7	6,9 ∼ 8,5

Na sequência, é verificado se existem interseções entre as faixas de valores pertencentes a um mesmo atributo. Caso exista interseção entre as faixas de valores, como destacado nos intervalos **7,80 ∼ 9,80** e **5,80 ∼ 7,80** da Tabela 9, que compartilham o valor **7,80** em comum nas duas faixas, estas são descartadas e a análise parte para outro conjunto de faixas de valores. O descarte é necessário pois as faixas de valores que compõem a interseção são ambíguas, impossibilitando que se obtenha um rótulo único capaz de representar cada grupo formado.

Caso nenhum atributo possua pelo menos uma faixa de valor capaz de representar cada um dos grupos, como mostrado na Tabela 9, o parâmetro GS é incrementado pelo parâmetro IGS e o processo de seleção de elementos é refeito utilizando um novo valor para GS. Por fim, são geradas novas faixas de valores a serem analisadas. Este processo é

necessário para remover interseções entre as faixas de valores, tornado-as únicas. Os valores únicos apresentam características capazes de distinguir cada um dos grupos, representando assim os seus rótulos.

Existindo pelo menos uma faixa de valores de um atributo sem interseção em cada grupo, como exibido na Tabela 10, o processo é encerrado e estas faixas de valores servem como rótulo para seus grupos.

Conforme visualizado na Tabela 10 após 787 iterações, utilizando-se de um grau de seleção de $0,5786$, as faixas de valores criadas ainda possuem interseção (representadas nas celulas destacadas), porém cada grupo possui pelo menos uma faixa de valor que não tem interseção, satisfazendo, desta forma, a condição de parada do modelo.

Tabela 10 – Rótulos gerados tendo como base os elementos selecionados em cada grupo: Iteração #787 - GS = 0,5786.

	Grupo 1	**Grupo 2**	**Grupo 3**
At. 1	3,1 ∼ 5,3	8,5 ∼ 9,8	5,8 ∼ 7,8
At. 2	3,6 ∼ 6,0	3,5 ∼ 6,0	7,3 ∼ 8,5

Finalmente, os rótulos finais são exibidos na Tabela 11. Esta tabela mostra as faixas de valores correspondentes a cada grupo e atributo. Podem existir várias faixas de valores relacionadas a um grupo, porém não deve existir interseção entre faixas de valores de um mesmo atributo.

Tabela 11 – Rótulos finais: Iteração #787.

	Grupo 1	**Grupo 2**	**Grupo 3**
At. 1	3,1 ∼ 5,3	8,5 ∼ 9,8	5,8 ∼ 7,8
At. 2	-	-	7,3 ∼ 8,5

Os rótulos exibidos na Tabela 11 apresentam somente faixas de valores únicas em relação a um atributo. O conjunto das faixas de valores compõe uma identificação para os grupos e representa a maioria dos elementos contidos nele.

Após atingida a condição de parada, a Tabela 12 apresenta os grupos associados aos seus respectivos rótulos, enfatizando a quantidade de elementos rotulados, o percentual de acertos e a quantidade de erros. O percentual de acertos está relacionado ao nível de precisão dos rótulos e tem como parâmetro a quantidade de elementos que estão presentes nos intervalos das faixas de valores formuladas. A quantidade de erros refere-se ao número de elementos que não obedecem aos rótulos formados.

Como pode ser percebido nas Tabelas 11 e 12, a abordagem defendida neste trabalho é capaz de formular rótulos, identificando faixas de valores únicas para os atributos em cada grupo formado e, desta forma, demonstra seu potencial no sentido de criar descrições que podem auxiliar no processo de tomada de decisão.

Tabela 12 – Porcentagem de acertos: Rótulos produzidos pelo modelo de rotulação proposto.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**Elementos**	**Acertos (%)**	**Erros**
1	At.1	3,1 $\sim$ 5,3	10	100	0
2	At.1	8,5 $\sim$ 9,8	8	80	2
3	At.1	5,8 $\sim$ 7,8	10	100	0
	At.2	7,3 $\sim$ 8,5	8	80	2

Por fim, com uma porcentagem média de acertos de $93,33\%$ e 02 erros no universo de 30 elementos, a abordagem apresentada mostrou-se promissora na formulação de faixas de valores e representação de rótulos para os elementos da base de dados de testes analisada.

4 Implementação e Testes

Neste capítulo serão apresentados os detalhes da implementação do modelo de rotulação abordado neste trabalho, bem como as ferramentas e os parâmetros utilizados com seus respectivos valores. O capítulo também contempla uma análise comparativa com outra proposta de rotulação e a aplicação desta abordagem em bases de dados digitais conhecidas na literatura.

4.1 Detalhes da Implementação

A proposta foi aplicado em bases de dados do repositório digital *UCI Machine Learning*[1], mais especificamente nas bases *Iris Data Set*[2], *Seed Data Set*[3], *Glass Identification Data Set*[4] e *Breast Cancer Wisconsin (Diagnostic) Data Set*[5]. Os valores dos atributos de todas as bases de dados analisadas possuem valores inseridos no conjunto dos números reais

Para saber o nível de precisão dos rótulos verificou-se a quantidade de elementos que estavam presente nos intervalos das faixas de valores formuladas. Uma vez feito isto, o elemento era atribuído ao grupo correspondente a sua faixa de intervalo. As quantidades de elementos associados aos seus respectivos rótulos podem ser verificadas na coluna "Elementos" das Tabelas 16, 21, 23 e 26.

4.2 Base de dados *Iris Data Set*

O modelo foi aplicado na base de dados *Iris Data Set*, disponível no repositório digital *UCI Machine Learning*. A base de dados refere-se a amostras de plantas e contém 150 elementos, cada um deles possui 4 atributos definidos por valores reais. Os atributos são: Comprimento da Sépala (CS), Largura da Sépala (LS), Comprimento da Pétala (CP) e Largura da Pétala (LP). A base de dados possui informações coletados de 3 classes de plantas Iris a saber: Setosa, Versicolor e Virginica. Segundo o trabalho de Fisher (1987), as classes podem ser divididas em:

[1] UCI Machine Learning Repository. Disponível em: <http://archive.ics.uci.edu/ml/index.php>. Acesso em: 22 nov. 2017.
[2] Iris Data Set. Disponível em: <http://archive.ics.uci.edu/ml/datasets/Iris>. Acesso em: 22 nov. 2017.
[3] Seed Data Set. Disponível em: <https://archive.ics.uci.edu/ml/datasets/seeds>. Acesso em: 22 nov. 2017.
[4] Glass Identification Data Set. Disponível em: <https://archive.ics.uci.edu/ml/datasets/glass+identification>. Acesso em: 02 jan. 2018.
[5] *Breast Cancer Wisconsin (Diagnostic) Data Set*. Disponível em: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. Acesso em: 26 jan. 2018.

1. 50 elementos da classe Iris Setosa;
2. 50 elementos da classe Iris Virginica;
3. 50 elementos da classe Iris Versicolor.

Como o modelo proposto tem como base a utilização do algoritmo de agrupamento *K-Means*, o atributo classe foi ignorado e o resultado dos grupos formados são descritos abaixo. Vale ressaltar que a cada execução do algoritmo *K-Means* os elementos podem ser inseridos em grupos diferentes dependendo da posição dos centroides iniciais. Esta colocação é válida para todas as bases de dados avaliadas nesta proposta.

De posse dessas informações, a primeira iteração do modelo sobre a base de dados *Iris Data Set* pode ser visualizada na Tabela 13. Nesta tabela são gerados os rótulos tendo como foco os elementos selecionados em cada grupo, partindo do princípio que o grau de pertinência dos elementos selecionados são maiores que o valor do parâmetro grau de seleção inicial (GS = 0,5).

Tabela 13 – *Iris Data Set* - Rótulos gerados tendo como base os elementos selecionados em cada grupo: Iteração #1.

	Grupo 1	**Grupo 2**	**Grupo 3**
CS	4,3 ∼ 5,8	4,9 ∼ 6,8	6,1 ∼ 7,9
LS	2,3 ∼ 4,4	2,0 ∼ 3,4	2,5 ∼ 3,8
CP	1,0 ∼ 1,9	3,0 ∼ 5,1	4,7 ∼ 6,9
LP	0,1 ∼ 0,6	1,0 ∼ 2,4	1,4 ∼ 2,5

Como observado na Tabela 13, o modelo detectou interseções entre faixas de valores de um mesmo atributo (células destacadas) em grupos distintos. Por essa razão, estas faixas são descartadas e a análise parte para outro conjunto de faixas de valores.

Após 568 iterações, com $GS = 0,5568$, a Tabela 14 exibe o resultado dos rótulos após execução do modelo na base de dados *Iris Data Set*.

Tabela 14 – *Iris Data Set* - Rótulos após iteração #568 - $GS = 0,5568$.

	Grupo 1	**Grupo 2**	**Grupo 3**
CS	4,3 ∼ 5,8	4,9 ∼ 6,6	6,2 ∼ 7,9
LS	2,3 ∼ 4,4	2,2 – 3,4	2,5 ∼ 3,8
CP	1,0 ∼ 1,9	3,5 ∼ 5,0	5,1 ∼ 6,9
LP	0,1 ∼ 0,6	1,0 ∼ 2,0	1,6 ∼ 2,5

Os rótulos visualizados na Tabela 15, após descartadas as interseções entre faixas de valores dos atributos CS e LS (Grupos 1, 2 e 3) e LP (Grupos 2 e 3), apresentam faixas de valores em relação aos grupos e atributos da base de dados estudada. É importante

destacar que podem existir várias faixas de valores relacionadas a um grupo, porém não deve existir interseção entre faixas de valores de um mesmo atributo.

Os rótulos exibidos na Tabela 15 mostram somente faixas de valores únicas em relação a um atributo. O conjunto das faixas de valores compõe uma identificação para os grupos e representa a maioria dos elementos contidos nele.

Tabela 15 – Rótulos Finais - Faixas de valores únicas.

	Grupo 1	**Grupo 2**	**Grupo 3**
CS	-	-	-
LS	-	-	-
CP	1,0 $\sim$ 1,9	3,5 $\sim$ 5,0	5,1 $\sim$ 6,9
LP	0,1 $\sim$ 0,6	-	-

Em consonância com as informações alcançadas depois da condição de parada do modelo, a Tabela 16 apresenta os grupos associados aos seus respectivos rótulos, enfatizando a quantidade de elementos que obedecem aos rótulos formados.

Tabela 16 – Grupos e elementos associados aos respectivos rótulos.

Grupo	**Rótulos**		**Quant. de Elementos**
	Atributos	**Intervalos**	
1	CP LP	1,0 $\sim$ 1,9 0,1 $\sim$ 0,6	50
2	CP	3,5 $\sim$ 5,0	52
3	CP	5,1 $\sim$ 6,9	36

Constata-se na Tabela 16 que o atributo Comprimento da Pétala (CP) está presente em todos os grupos. Essa condição demonstra que os três grupos se diferem pela faixa de valor desse atributo. Um especialista que eventualmente quiser atribuir um novo elemento a um grupo qualquer teria no comprimento da pétala o principal diferencial para a identificação do novo elemento. Os demais atributos e suas faixas de valores representam características secundárias. Estas, por sua vez, em conjunto com a(s) característica(s) principal(is), podem representar os grupos de forma única.

Ao verificar o total de elementos inseridos em cada faixa de valores, percebe-se que 12 não foram rotulados em nenhum dos grupos. Os intervalos com seus respectivos atributos que não obedeceram aos rótulos gerados pelo modelo, podem ser visualizados na Tabela 17.

Os intervalos dispostos na Tabela 17 possuem valores nos atributos CP e LP que não existem em nenhuma das faixas de valores formuladas pelo modelo de rotulação proposto (ver Tabela 15). Isso ocorre devido ao fato de existirem nos intervalos interseções entre faixas de valores dos atributos em grupos distintos. Com isto o modelo conseguiu rotular $92,52\%$ dos elementos baseando-se nas características existentes de cada grupo.

Tabela 17 – Elementos não rotulados na base *Iris Data Set*.

Atributos	**Intervalos**	**Elementos**	**Grupo**
CP	4,7 ∼ 5,0	9	3
LP	1,0 ∼ 2,4	1	2
	1,4 ∼ 2,3	2	3

Portanto, a aplicação do modelo de rotulação proposto mostrou-se eficiente na formulação de faixas de valores para representação das classes de plantas existentes na base de dados *Iris Data Set*.

4.3 Base de dados *Seed Data Set*

A base de dados *Seed Data Set* contém informações sobre tipos de sementes de trigo. Através desta base é possível analisar 210 elementos, onde cada um possui 7 atributos: área (A), perímetro (P), densidade (C), comprimento da semente (LK), largura da semente (WK), coeficiente de assimetria (AC) e comprimento do sulco da semente (LKG). De acordo com Małgorzata et al. (2010), as amostras de sementes são classificadas em três tipos:

1. 70 elementos da classe Kama;
2. 70 elementos da classe Rose;
3. 70 elementos da classe Canadian.

Por utilizar um algoritmo de agrupamento de dados, da mesma forma como ocorreu com a base *Iris Data Set*, o atributo classe foi ignorado e o resultado da análise é demonstrada a seguir.

Depois da primeira iteração do modelo sobre a base de dados, os rótulos gerados podem ser observados na Tabela 18. Ela expressa os rótulos de acordo com os elementos selecionados em cada grupo. Ressalta-se que o grau de pertinência dos elementos selecionados na primeira iteração são maiores que o valor inicial do parâmetro grau de seleção (GS = 0,5).

A Tabela 18 apresenta interseções entre faixas de valores de um mesmo atributo (células destacadas) em grupos distintos. Por essa razão, estas faixas são descartadas e a análise parte para outro conjunto de faixas de valores.

Na sequência, a Tabela 19 exibe o resultado dos rótulos gerados após condição de parada do modelo aplicado à base de dados *Seed Data Set*. A condição de parada é atingida quando não existe interseções entre faixas de valores em pelo mesmo um atributo dentre todos os grupos formados.

Tabela 18 – Seed Data Set - Rótulos gerados tendo como base os elementos selecionados em cada grupo: Iteração #1.

	Grupo 1	**Grupo 2**	**Grupo 3**
Área	12,78 ∼ 16,63	16,53 ∼ 21,18	10,59 ∼ 14,28
Perímetro	13,50 ∼ 15,46	15,34 ∼ 17,25	12,41 ∼ 14,17
Densidade	0,8527 ∼ 0.9153	0,8452 ∼ 0.9108	0,8081 ∼ 0,9183
Comprimento da Semente	5,1380 ∼ 6,0530	5,7910 ∼ 6,6750	4,8990 ∼ 5,5410
Largura da Semente	2,9750 ∼ 3,5820	3,4030 ∼ 4.0330	2,6300 ∼ 3,3830
Coeficiente de Assimetria	0,7651 ∼ 5,5930	1,4720 ∼ 6,6820	1,4150 ∼ 8.4560
Comprimento do Sulco da Semente	4,6070 ∼ 5,9220	5,4840 ∼ 6,5500	4,5190 ∼ 5.4910

Tabela 19 – Seed Data Set - Rótulos gerados após condição de parada do modelo.

	Grupo 1	**Grupo 2**	**Grupo 3**
Área	13,50 ∼ 16.44	17,26 ∼ 21,18	10,59 ∼ 13.37
Perímetro	13,85 ∼ 15,27	15,66 ∼ 17,25	12,41 ∼ 13,95
Densidade	0,8527 ∼ 0,9153	0,8452 ∼ 0,9108	0,8081 ∼ 0.8923
Comprimento da Semente	5,2050 ∼ 5.9200	5,7910 ∼ 6,6750	4,8990 ∼5,4950
Largura da Semente	3,1130 ∼ 3,5820	3,4030 ∼ 4,0330	2,6300 ∼ 3,1280
Coeficiente de Assimetria	0,7651 ∼ 4,1850	1,4720 ∼ 6,6820	2,2210 ∼ 7,5240
Comprimento do Sulco da Semente	4,6490 ∼ 5,8790	5,6610 ∼ 6,5500	4,6050 ∼ 5,4910

Os rótulos disponíveis na Tabela 20, após descartadas as interseções entre faixas de valores dos atributos Perímetro (Grupos 1 e 3) e Densidade, Comprimento da Semente, Largura da Semente, Coeficiente de Assimetria e Comprimento do Sulco (em todos os grupos), apresentam faixas de valores distintas em relação aos atributos e grupos gerados.

Tabela 20 – Rótulos Finais - Faixas de valores únicas

	Grupo 1	**Grupo 2**	**Grupo 3**
Área	13,50 ∼ 16.44	17,26 ∼ 21,18	10,59 ∼ 13.37
Perímetro	-	15,66 ∼ 17,25	-

A Tabela 20 revela faixas de valores únicas em relação a um atributo para cada grupo investigado. O conjunto de faixas de valores compõem uma forma de identificar os grupos representando a maioria do elementos neles contidos.

Os grupos associados aos seus respectivos rótulos, enfatizando a quantidade de elementos (coluna "Elementos") que obedecem aos rótulos formados na base de dados *Seed Data Set*, estão disponíveis na Tabela 21.

Tabela 21 – Grupos e elementos associados aos respectivos rótulos.

Grupos	**Rótulos**		**Elementos**
	Atributos	**Intervalos (cm)**	
1	Área	13,5 ∼ 16,44	61
2	Área Perímetro	17,26 ∼ 21,18 15,66 ∼ 17,25	54 55
3	Área	10,59 ∼ 13,37	80

Ao explorar a Tabela 21 verifica-se que o atributo Área está presente em todos os grupos. Essa condição demonstra que os três grupos se diferem pela faixa de valor desse atributo. Um especialista que eventualmente quiser atribuir um novo elemento a um grupo qualquer teria na característica Área da semente o principal diferencial para a identificação do novo elemento. Os demais atributos e suas faixas de valores representam características secundárias. Estas, por sua vez, em conjunto com a(s) característica(s) principal(is), podem representar os grupos de forma única.

Ao verificar o total de elementos inseridos em cada faixa de valores (ver Tabela 21, considerando o atributo Área), percebe-se que 15 não foram rotulados em nenhum dos grupos. Os intervalos com os respectivos atributos que não obedeceram os rótulos gerados pelo modelo podem ser visualizados na Tabela 22.

Tabela 22 – Elementos não rotulados na base *Seed Data Set*

Atributos	**Intervalos**	**Elementos**	**Grupos**
Área	12,78 ∼ 13,45 13,63 ∼ 13,63	5 1	1
Área Perímetro	16,53 ∼ 17,12 15,34 ∼ 15,65	7 6	2
Área	13,99 ∼ 14,28	2	3

Os intervalos dispostos na Tabela 22 possuem valores nos atributos Área e Perímetro que não existem em nenhuma das faixas de valores apresentadas pela proposta abordada neste trabalho (ver Tabela 21).

Por fim, considerando que a base de dados *Seed Data Set* possui 210 elementos e o modelo proposto rotulou corretamente 195 elementes $(92,86\%)$, é possível observar que a abordagem mostrou-se eficiente na formulação de faixas de valores para os tipos de sementes existentes na base de dados foco da análise.

4.4 Base de dados *Glass Identification Data Set*

A base de dados *Glass Identification Data Set*, está disponível no repositório digital *UCI Machine Learning*. As amostras referem-se a tipos de vidros e são utilizadas como evidência para investigação criminal. Na cena de um crime o tipo de vidro pode ser usado como prova caso esteja corretamente identificado. A base contém 214 amostras, sendo que cada uma possui 9 atributos nomeados da seguinte forma: Índice de Refração (IR), porcentagem de óxido do elementos Sódio (Na), Magnésio (Mg), Alumínio (Al), Silício (Si), Potássio (K), Cálcio (Ca), Bário (Ba) e Ferro (Fe).

De acordo com o repositório *Glass Identification Data Set*, as amostras podem ser distribuídas em duas classes conforme descritas a seguir:

1. **Vidros de janelas** (163)
    - processados/temperados (87)
        - janelas de construção (70)
        - janelas de veículos (17)
    - não processados (76)
        - janelas de construção (76)
        - janelas de veículos (0)
2. **Vidros não-janelas** (51)
    - recipientes (13)
    - louças (9)
    - faróis (29)

Da mesma forma como nas bases anteriores, o atributo classe foi ignorado e o resultado da análise é demonstrada a seguir.

A Tabela 23 apresenta os grupos associados aos seus respectivos rótulos, enfatizando a quantidade de elementos e os erros que não obedecem aos rótulos formados na base de dados *Glass Identification Data Set*.

O atributo Magnésio (Mg) está presente nos rótulos finais descritos na Tabela 23. Esse fato o caracteriza como o mais significativo para identificação de um elemento na base de dados. Um especialista que eventualmente quiser atribuir um novo elemento a um grupo qualquer, teria no percentual de óxido do Mg o principal diferencial para identificação do elemento. De acordo com os rótulos gerados, o Grupo 1 é constituído por amostras de vidros de janelas e o Grupo 2 de amostras de recipientes, louças e faróis.

Tabela 23 – Gupos e elementos associados aos respectivos rótulos - Interação #403 - GS = 0,5403

Grupos	**Rótulos**		**Elementos**	**Erros**
	Atributos	**Intervalo**		
1	Mg	2,19 ∼ 4,49	163	3
2	Mg	0 ∼ 1,88	51	0

Portanto, considerando que a base de dados analisada possui 214 elementos e o modelo proposto rotulou corretamente 211 (99, 08%), observa-se que a abordagem mostrou-se eficiente na formulação de faixas de valores para identificação de tipos de vidros existentes na base de dados *Glass Identification Data Set*.

4.5 Base de dados *Breast Cancer Wisconsin (Diagnostic) Data Set*.

A base de dados *Breast Cancer Wisconsin (Diagnostic) Data Set*, foi idealizada nos Hospitais da Universidade de Wisconsin, Madison, USA. Ela contém amostras de diagnóticos de câncer de mama totalizando 699 elementos. Cada elemento possui 11 atributos, conforme descrito na Tabela 24 (MANGASARIAN; WOLBERG, 1990). Para efeito de análise somente 9 atributos foram considerados relevantes. Dentre os atributos ignorados estão Id e Diagnóstico.

Tabela 24 – Atributos com os possíveis domínios da base *Breast Cancer Wisconsin Data Set*

	Atributos	**Significado**	**Domínio**
1	Id	Número do código da amostra	-
2	EA	Espessura do Aglomerado	1 - 10
3	UTC	Uniformidade do Tamanho da Célula	1 - 10
4	UFC	Uniformidade da Forma da Célula	1 - 10
5	AM	Aderência Marginal	1 - 10
6	TUCE	Tamanho Único da Célula Epitelial	1 - 10
7	ND	Núcleo Descoberto	1 - 10
8	CS	Cromatina Suave	1 - 10
9	NN	Nucleulus Normal	1 - 10
10	Diagnóstico	Atributo Classe	2: Benigno 4: Maligno

Os elementos da base de dados estão distribuídos da seguinte forma: 458 amostras benignas (65,5%) e 241 malignas (34,5%). Dentre as amostras 16 apresentavam valores nulos para o atributo Núcleo Descoberto (ND). Levando essa peculiaridade em consideração, o modelo defendido neste trabalho analisou somente 683 elementos.

A Tabela 25 apresenta os parâmetros utilizados pelo modelo para a formulação de rótulos para a base de dados *Breast Cancer Wisconsin (Diagnostic) Data Set.*

Tabela 25 – Parâmetros do modelo de rotulação aplicados a base *Breast Cancer Wisconsin Data Set.*

Parâmetros	**Valores**
Número de Grupos	2
Grau de Seleção inicial	0,5
Incremento do Grau de Seleção	0,0001

Os rótulos produzidos, após 2.191 iterações e GS = $0,7195$, estão presentes na Tabela 26.

Tabela 26 – Gupos e elementos associados aos respectivos rótulos - Iteração #2195 - GS = 0,7195.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**Elementos**	**Acertos (%)**	**Erros**
1	EA	$4 \sim 10$	231	88,31	27
2	EA	$1 \sim 3$	452	98,01	9

De acordo com o modelo, o atributo Espessura do Aglomerado (EA) é o mais significativo para identificação de amostras acometidas com câncer de mama. Segundo Mangasarian e Wolberg (1990), o valor para o atributo EA, como os demais atributos analisados, é designado quando a área com suspeita de malignidade é examinada por um profissional através do uso de um microscópio. O procedimento de observação física da área afetada é chamado de Aspiração por Agulha Fina (AAF). O domínio para esse atributo é imputado pelo profissional observador, podendo variar em uma escala de $1 - 10$. Quanto maior for o valor para essa característica, maior a possibilidade de incidência da doença. Essa informação pode auxiliar um especialista na detecção da enfermidade e, consequentemente, prescrever um tratamento mais adequado às peculiaridades de cada caso.

Verificando a Tabela 26, é possível destacar dois pontos. a) as amostras contidas no Grupo 1 (33,82%) representam diagnósticos referentes à condição maligna. Enquanto o Grupo 2 representa amostras benignas (66,18%) para a doença câncer de mama; e b) 36 elementos não foram rotulados corretamente. Os intervalos com aos respectivos atributos que não obedeceram os rótulos gerados, podem ser conferidos na Tabela 27.

A métrica da soma total de erros representa o total de valores de atributos que não existem nos intervalos das faixas de valores geradas pelos experimentos realizados. A Tabela 28 exibe os resultados obtidos referentes à soma total de erros alcançados.

Tabela 27 – Elementos não rotulados na base *Breast Cancer Wisconsin Data Set*.

Atributos	**Intervalo**	**Elementos**	**Grupos**
EA	$1 \sim 3$	27	1
EA	$4 \sim 5$	9	2

Tabela 28 – Resultado da métrica soma total de erros para a *Breast Cancer Wisconsin Data Set*.

Métrica	**RBD**
Média da porcentagem de acertos em cada grupo (%)	93,16
Total de erros	36

Finalizando a análise sobre a base *Breast Cancer Wisconsin Data Set*, verifica-se que os resultados alcançados foram muito significativos. Do universo de 683 amostras o modelo proposto neste trabalho conseguiu rotular corretamente 647, representando uma média de acertos de $93,16\%$. Os dados comprovam que os rótulos presentes na Tabela 26, representados pelo intervalo $4 \sim 10$ (Grupo 1) auxiliam o profissional na identificação de enfermidades em amostras referentes a diagnósticos de pacientes com câncer de mama.

A Figura 6 apresenta uma síntese dos resultados alcançados após análise das bases de dados *Iris Data Set*, *Seed Data Set*, *Glass Identification Data Set* e *Breast Cancer Wisconsin Data Set*. Esta figura enfatiza a média de porcentagem de acertos e a soma total de erros encontrados pela proposta de rotulação.

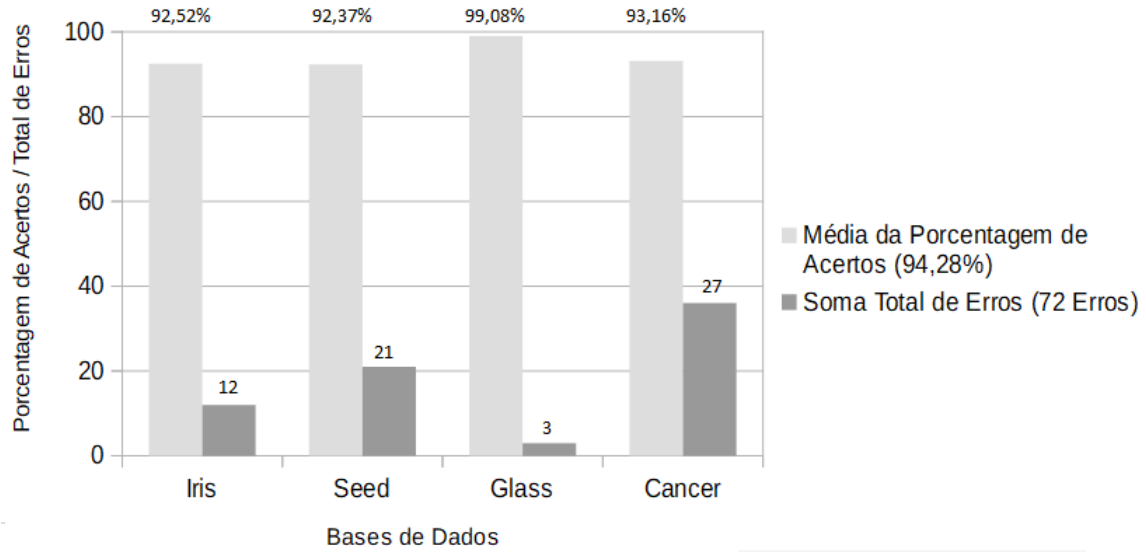

Figura 6 – Síntese dos resultados alcançados pelo modelo proposto.

Na Figura 6, observa-se que a média de porcentagem de acerto em todas as base de dados ultrapassa a casa dos 90%, destacando-se a base *Glass Identification Data Set* que atingiu mais de 99% de acertos.

Com 76 erros encontrados $(6,05\%)$, de um total de 1257 amostras em todas as bases de dados é possível cogitar que a quantidade de erros é pequena considerando-se o universo de amostras disponíveis nos repositórios digitais.

Os resultados apresentados na Figura 6, sustentam a eficiência do modelo na formulação de rótulos para identificar características relevantes nos elementos de cada grupo e identificá-los de forma única. Estes fatores são importantes para a interpretação dos elementos, pois desta forma é factível saber o que torna um elemento pertencente a um grupo e quais são as diferenças e similaridades entre os grupos. De posse dessas informações é razoável afirmar que o especialista tem subsídios para o processo de tomada de decisão.

Por fim, o modelo idealizado nesta dissertação de mestrado foi testado em outras bases de dados digitais e os resultados podem ser verificados no Apêndice A.

5 Análise Comparativa

Este capítulo descreve uma análise comparativa entre a abordagem defendida neste trabalho e o modelo idealizado por Ribeiro (2016) e por Machado, Ribeiro e Rabelo (2015). Estes últimos também abordam o problema da rotulação possibilitando a análise comparativa. Os objetivos são: coletar dados das aplicações dos modelos de rotulação sobre as bases de dados *Iris Data Set*, *Seed Data Set* e *Glass Identification Data Set*, calcular métricas utilizando os dados disponíveis e realizar discussões sobre os resultados obtidos.

O modelo de Ribeiro (2016) foi descrito na Seção 2.4. A referida seção expõe a proposta do pesquisador, os parâmetros e a utilização do algoritmo *Fuzzy C-Means* para agrupar e rotular os dados. O modelo foco deste trabalho utiliza como base o algoritmo *K-Means* para agrupar, em seguida transformar a saída padrão em distância para grau de pertinência e, posteriormente, rotula os dados das bases de dados estudadas. É importante ressaltar que ambas as técnicas são classificadas como algoritmos de aprendizagem de máquina não supervisionada e os modelos analisados possuem um comportamento muito similar.

Para facilitar o entendimento, os modelos propostos nos trabalhos de Ribeiro (2016) e Machado, Ribeiro e Rabelo (2015) foram chamados de Rotulação *Fuzzy* e o modelo descrito nesse trabalho foi chamado de Rotulação Baseada em Distância (RBD).

Os dados sobre o modelo Rotulação *Fuzzy* foram coletados a partir dos trabalhos de Ribeiro (2016) e Machado, Ribeiro e Rabelo (2015). Os dados referentes ao modelo RBD foram gerados a partir da aplicação da abordagem defendida nesta proposta nas mesmas bases de dados referenciadas pelo modelo de Rotulação *Fuzzy*. Os dados coletados foram: os parâmetros utilizados em cada base de dados; o número de elementos contidos em cada grupo; os rótulos gerados pelos modelos, sendo compostos pelos atributos e suas respectivas faixas de valores; a porcentagem de acertos, formulada a partir da quantidade de elementos que possui valores existentes em uma faixa de valor e do número de elementos em cada grupo e a quantidade de erros encontrados, isto é, o número de elementos que não possuem valores existentes em uma determinada faixa de valor.

Para realizar a comparação foram utilizadas as três métricas expostas no trabalho de Lopes, Machado e Rabelo (2014). A primeira pode ser descrita como a média da menor porcentagem de acerto em cada grupo. Esta métrica é calculada a partir da porcentagem de acertos, levando em consideração a quantidade de elementos em cada grupo. Portanto, ela é fortemente influenciada pelo algoritmo de agrupamento utilizado.

A segunda métrica consiste na soma total de erros, representando o total de valores de atributos que não existem no intervalo das faixas de valores. A terceira métrica analisa a

quantidade de parâmetros utilizados em cada base de dados. As métricas foram calculadas a partir dos dados coletados e descritos anteriormente.

5.1 Base *Iris Data Set*

Os parâmetros utilizados pelos modelos para a formação de rótulos para a base de dados *Iris Data Set* estão disponíveis na Tabela 29. O número de grupos disposto na tabela leva em consideração o trabalho de Fisher (1987), onde o autor especifica que exitem três classes de plantas no repositório digital foco da análise.

Tabela 29 – Parâmetros utilizados pelos Modelos de Rotulação *Fuzzy* e RBD.

Parâmetros	**Valores**
Número de Grupos	3
Grau de Seleção inicial	0,5
Incremento do Grau de Seleção	0,0001

Os trabalhos em análise comparativa utilizaram a mesma quantidade e os mesmos valores para os parâmetros descritos nas respectivas pesquisas. Por conta dessa especificidade, a Tabela 29 é suficiente para representar os parâmetros e valores utilizados nos dois modelos. Nas Tabelas 30 e 31 são exibidos os rótulos produzidos pelos modelos de rotulação para a base *Iris Data Set*.

Tabela 30 – Porcentagem de acertos: rótulos produzidos pelo modelo de Rotulação *Fuzzy* para a base *Iris Data Set*.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos (cm)**	**Elementos**	**Acertos (%)**	**Erros**
1	CP LP	1,0 ∼ 1,9 0,1 ∼ 0,6	50	100	0
2	CP	5,1 ∼ 6,9	40	92	3
3	CP	3,5 ∼ 5,0	60	86,66	8

Tabela 31 – Porcentagem de acertos: rótulos produzidos pelo modelo de RBD para a base *Iris Data Set*.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos (cm)**	**Elementos**	**Acertos (%)**	**Erros**
1	CP LP	1,0 ∼ 1,9 0,1 ∼ 0,6	50	100	0
2	CP	3,5 ∼ 5,0	61	85,25	9
3	CP	5,1 ∼ 6,9	39	92,31	3

Nas Tabelas 30 e 31 observa-se que cada grupo é representado por faixas de valores associadas a seus respectivos atributos, formando assim os rótulos. É importante ressaltar que após a formulação dos rótulos, fica evidenciado que o atributo CP distingue cada um dos grupos, facilitando, desta forma, a interpretação dos rótulos, configurando uma característica determinante na classificação dos elementos.

Por utilizarem algoritmos não supervisionados distintos, a disposição dos grupos com suas respectivas faixas de valores diferem em sua localização em dois grupos a saber: no modelo Rotulação *Fuzzy* o Grupo 2 foi produzido com o rótulo 5,1 $\sim$ 6,9 e o Grupo 3 com o rótulo 3,5 $\sim$ 5,0. No modelo RBD ocorre o inverso: rótulos 3,5 $\sim$ 5,0 e 5,1 $\sim$ 6,9 para os Grupos 2 e 3, respectivamente.

A métrica da soma total de erros representa o total de valores de atributos que não existem nos intervalos das faixas de valores geradas pelos modelos de rotulação analisados. A Tabela 32 demonstra os resultados obtidos referentes à métrica soma total de erros.

Tabela 32 – Resultado da métrica soma total de erros para a base *Iris Data Set*.

Métricas	**Rotulação Fuzzy**	**RBD**
Média da porcentagem de acertos em cada grupo (%)	92,88	92,52
Total de erros	11	12

Com os valores apresentados na Tabela 32, constata-se que ambos os modelos atingiram valores muito próximos nas métricas, com uma pequena vantagem em relação a proposta defendida no modelo de Rotulação *Fuzzy*. Nesta abordagem, o rótulo representando o intervalo 3,5 $\sim$ 5,0 (Grupo 3), com 86,66%, foi superior ao mesmo intervalo no RBD. Em contrapartida, o modelo proposto teve uma pequena vantagem quando observa-se o rótulo representado pelo intervalo 5,1 $\sim$ 6,9 (Grupo 3), com 92,31%, em relação a Rotulação *Fuzzy*. Este último também obteve um número de erros menor (um elemento) em referência à Rotulação Baseada em Distância.

5.2 Base *Seed Data Set*

A Tabela 29 disponibiliza as parâmetros utilizados pelos modelos para a formação de rótulos para a base de dados *Seed Data Set*.

Os rótulos produzidos pelos modelos de rotulação para a base *Seed Data Set* podem ser visualizados nas Tabelas 33 e 34. Nestas tabelas é possível observar que cada grupo é representado por faixas de valores associadas a seus respectivos atributos, formando assim os rótulos. É importante ressaltar que após a formulação dos rótulos, fica evidenciado que o atributo Área distingue cada um dos grupos, facilitando, desta forma, a interpretação dos rótulos, configurando uma característica determinante na classificação dos elementos.

Tabela 33 – Porcentagem de acertos: rótulos produzidos pelo modelo de Rotulação Fuzzy para a base Seed Data Set.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**Elementos**	**Acertos (%)**	**Erros**
1	Área	13,50 ∼ 16,14	72	75	18
2	Área	10,59 ∼ 13,37	77	97,4	2
3	Área Perímetro	17,26 ∼ 21,18 15,66 ∼ 17,25	61	88,52 90,16	7 6

Tabela 34 – Porcentagem de acertos:rótulos produzidos pelo modelo de RDB para a base *Seed Data Set*.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**elementos**	**Acertos (%)**	**Erros**
1	Área	13,50 ∼ 16,44	67	91,04	6
2	Área Perímetro	17,26 ∼ 21,18 15,66 ∼ 17,25	61	88,52 90,16	7 6
3	Área	10,59 ∼ 13,57	82	97,56	2

Por utilizarem algoritmos não supervisionados distintos, a disposição dos grupos formados com suas respectivas faixas de valores diferem em sua localização em dois grupos a saber: no modelo de Rotulação *Fuzzy* o Grupo 2 foi produzido com o rótulo 10,59 ∼ 13,37 e o Grupo 3 com os rótulos 17,26 ∼ 21,18 e 15,66 ∼ 17,25 para os atributos Área e Perímentro, respectivamente. No modelo RBD ocorre o inverso: os rótulos 17,26 ∼ 21,18 e 15,66 ∼ 17,25 representando os atributos Área e Perímentro do Grupo 2 e o rótulo 10,59 ∼ 13,37 representado o atributo Área do Grupo 3.

O total de valores de atributos que não existem nos intervalos das faixas de valores geradas pelos modelos de rotulação analisados, representam a métrica da soma total de erros. A Tabela 35 demonstra os resultados obtidos referentes a métrica soma total de erros.

Tabela 35 – Resultado da métrica soma total de erros para a base *Seed Data Set*.

Métricas	**Rotulação Fuzzy**	**RBD**
Média da porcentagem de acertos em cada grupo (%)	86,97	92,37
Total de erros	33	21

Com as informações visualizadas na Tabela 35, apura-se que o modelo RBD obteve resultados superiores (92, 37% - 21 erros) aos obtidos pelo modelo de Rotulação *Fuzzy* (86, 97% - 33 erros). Com os resultados dispostos na Tabela 35, finaliza-se a análise

comparativa entre os modelos de Rotulação *Fuzzy* e RBD sobre a base de dados *Seed Data Set*.

5.3 Base *Glass Identification Data Set*

Os parâmetros utilizados pelos modelos para a formação de rótulos para a base de dados *Glass Identification Data Set* são visíveis na Tabela 36.

Tabela 36 – Parâmetros dos Modelos de Rotulação *Fuzzy* e RBD aplicados a base *Glass Identification Data Set*.

Parâmetros	**Valores**
Número de Grupos	2
Grau de Seleção inicial	0,5
Incremento do Grau de Seleção	0,0001

Os trabalhos em análise comparativa utilizaram a mesma quantidade e os mesmos valores para os parâmetros descritos na Tabela 36. Os rótulos produzidos pelos modelos de rotulação podem ser visualizados nas Tabelas 37 e 38.

Tabela 37 – Porcentagem de acertos: rótulos produzidos pelo modelo de Rotulação Fuzzy para a base *Glass Identification Data Set*.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**Elementos**	**Acertos (%)**	**Erros**
1	Mg	2,19 ∼ 4,49	53	98,11	1
2	Mg	0 ∼ 1,88	161	98,75	2

Tabela 38 – Porcentagem de acertos: rótulos produzidos pelo modelo de RBD para a base *Glass Identification Data Set*.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**Elementos**	**Acertos (%)**	**Erros**
1	Mg	2,19 ∼ 4,49	163	98,16	3
2	Mg	0 ∼ 1,88	51	100	0

Nas Tabelas 37 e 38, verifica-se que os grupos são representados pelas mesmas faixas de valores associadas a seus respectivos atributos, formando assim os rótulos. Ressalta-se que após a formulação dos rótulos fica evidenciado que o atributo Magnésio (Mg) é utilizado para representar os grupos, característica determinante na classificação dos elementos.

Outro detalhe a destacar é que mesmo os modelos utilizando as mesmas faixas de valores para representar os grupos, a quantidade de elementos e a taxa de acerto são diferenciadas. A disposição dos grupos formados com suas respectivas faixas de valores

diferem na quantidade de elementos rotulados. No modelo de Rotulação Fuzzy o Grupo 1 foi produzido com o rótulo 2,19 ∼ 4,49 com 53 elementos e o Grupo 2 com o rótulo 0 ∼ 0,188, com 161 elementos. O modelo RBD com as mesmas faixas de valores produziu 163 e 51 elementos para os Grupos 1 e 2, respectivamente.

A métrica da soma total de erros representa o total de valores de atributos que não existem nos intervalos das faixas de valores geradas pelos modelos de rotulação analisados. A Tabela 39 demonstra os resultados obtidos referentes à soma total de erros alcançados.

Tabela 39 – Resultado da métrica soma total de erros para a base *Glass Identification Data Set*.

Métrica	**Rotulação Fuzzy**	**RBD**
Média da porcentagem de acertos em cada grupo (%)	98,43	99,08
Total de erros	3	3

Com as informações apresentadas na Tabela 39 certifica-se que o modelo RBD obteve resultados superiores (99, 08%) aos obtidos pelo modelo de Rotulação Fuzzy (98, 43%), ambos com 03 erros detectados.

5.4 Análise Comparativa Geral

Esta seção apresenta em forma gráfica o resultado geral da média da porcentagem de acertos e da soma total de erros referente à análise comparativa entre os modelos de Rotulação *Fuzzy* e Rotulação Baseada em Distância, para as bases de dados *Iris Data Set*, *Seed Data Set* e *Glass Identification Data Set*. As Figuras 7 e 8 sintetizam os resultados finais alcançados pelos modelos.

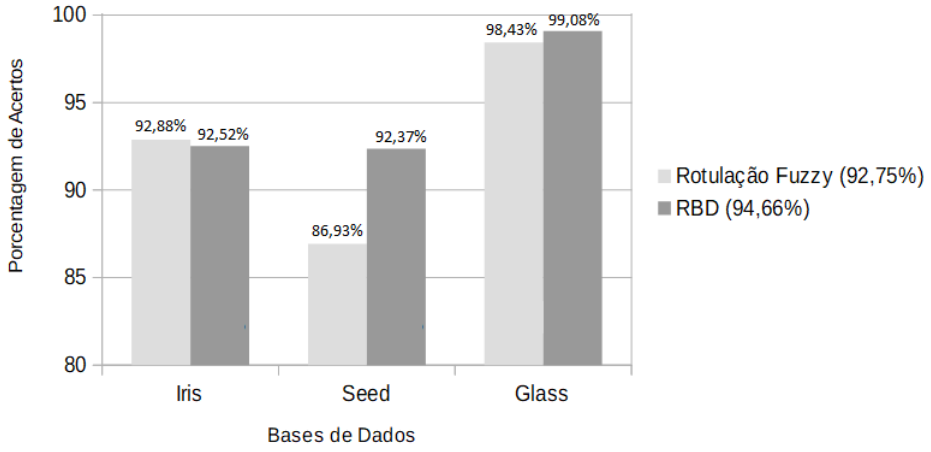

Figura 7 – Análise Comparativa: média da porcentagem de acertos.

Visualizando a Figura 7, constata-se que o modelo RDB obteve resultado superior no item média de porcentagem de acertos (94,66%) em comparação aos resultados alcançados pelo modelo de Rotulação Fuzzy (92,75%).

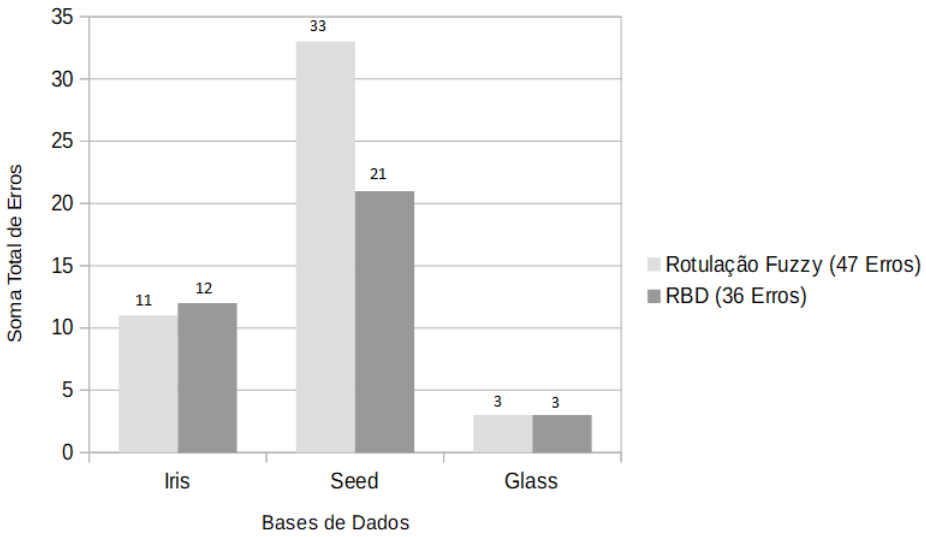

Figura 8 – Análise Comparativa: Soma total de erros.

A Figura 8, exibe o resultado geral referente a métrica soma total de erros para os modelos alvo da comparação. Mais uma vez o modelo RDB destacou-se, tendo menor número de erros (36) em relação aos resultados alcançados pelo modelo de Rotulação Fuzzy (47).

Com as informações disponíveis nas Figuras 7 e 8, ressalta-se que o modelo proposto neste trabalho apresentou uma abordagem capaz de analisar grupos de dados e produzir rótulos para ajudar na compreensão e auxiliar especialistas no processo de tomada de decisão.

É importante frisar que os resultados finais obtidos pelo modelo RBD são frutos de uma única execução do algoritmo *K-Means* sobre as bases de dados analisadas. O ideal seria executar o modelo diversas vezes, já que a cada execução o algoritmos *K-Means* pode vir a gerar grupos com elementos diferentes. Como os resultados obtidos pelo modelo de Rotulação *Fuzzy* são oriundos de apenas uma execução, achamos que seria mais justo comparar também com uma única execução com a abordagem RBD.

6 Conclusões e Trabalhos Futuros

Este capítulo inicia-se expondo as conclusões obtidas com a aplicação do modelo proposto nas bases *Iris Data Set*, *Seed Data Set* e *Glass Identifications Data Set*. Em seguida comenta-se os resultados obtidos após análise comparativa com outro modelo de rotulação, como também os resultados de análises sobre outras bases de dados. Para finalizar, apresenta sugestões de trabalhos futuros que podem contribuir para a expansão da presente proposta.

6.1 Conclusões

Esta qualificação idealizou um modelo que utiliza um algoritmo de aprendizagem de máquina não supervisionada baseado em distância capaz de elaborar rótulos e, desta forma, conseguir representar os dados contidos nos grupos. A definição dos rótulos se dá pela detecção de faixas de valores dos atributos para cada grupo formado. As faixas de valores são associadas a atributos capazes de distinguir cada grupo de forma única. Os rótulos gerados contribuem para o entendimento dos grupos e podem ser utilizados por um especialista no processo de tomada de decisão.

Ambos os modelos de rotulação analisados - Rotulação *Fuzzy* e Rotulação Baseada em Distância (RBD) - mostraram-se promissores na produção de rótulos levando em consideração as amostras das bases de dados, foco da análise comparativa. Começando pelas amostras vinculadas à base de dados *Iris Data Set*, o resultado apresentado pelo modelo de Rotulação *Fuzzy* conseguiu apresentar uma média de percentual de acertos de $92,88\%$ (com 11 elementos rotulados incorretamente), enquanto o modelo centrado neste trabalho atingiu média de percentual de acertos de $92,52\%$ (com 12 elementos rotulados incorretamente). Com isso, percebe-se que houve uma diferença positiva de $0,36\%$ a favor do modelo de Rotulação Fuzzy. Portanto, taxas médias de acertos bem próximas formulando poucas faixas de valores para representar os grupos de forma única.

Quando o foco da análise foram as amostras disponíveis na base de dados *Seed Data Set*, o modelo Rotulação *Fuzzy* conseguiu formular média de percentual de acertos $86,97\%$ (com 33 elementos rotulados incorretamente), enquanto o modelo RDB atingiu média de percentual de acertos de $92,37\%$ (com 21 elementos rotulados incorretamente). Essas informações demonstram uma diferença positiva de $5,4\%$ na média de percentual de acertos a favor ao modelo de Rotulação Baseada em Distância.

Referenciando a análise comparativa sobre a base de dados *Glass Identification Data Set*, o modelo apresentado na Rotulação *Fuzzy* conseguiu formular média de percentual

de acertos $98,43\%$, enquanto o modelo RDB atingiu média de percentual de acertos de $99,08\%$, ambos com 3 elementos rotulados incorretamente. Portanto, uma diferença positiva de $0,65\%$ a favor do modelo de Rotulação Baseada em Distância.

Como resultado geral da análise comparativa, o modelo RBD obteve média superior em porcentagem de acertos $(94,66\%)$ em comparação aos resultados alcançados pelo modelo de Rotulação Fuzzy $(92,75\%)$. Da mesma forma, o modelo proposto neste trabalho destacou-se quando o critério da análise foi referente à métrica soma total de erros (36) em relação aos resultados alcançados pelo modelo de Rotulação Fuzzy (47).

Na análise comparativa conclui-se que a proximidade dos resultados obtidos pelos modelos investigados ocorre não somente pela similaridade dos parâmetros utilizados, mas também pela forma como os modelos são inicialmente implementados. A divergência crucial está no fato de o modelo abordado neste trabalho, antes de formular as faixas de valores e, consequentemente, produzir os rótulos, transforma a saída padrão do algoritmo não supervisionado baseado em distância *K-Means* para grau de pertinência. A partir deste ponto, há similaridade entre os modelos avaliados. Por esse motivo, é possível verificar que com a utilização de outros algoritmos de aprendizagem de máquina não supervisionada baseado em distância é factível elaborar definições (rótulos) para o entendimento de grupos em bases de dados submetidas ao modelo proposto.

Posteriormente à análise comparativa, quatro bases de dados foram submetidas à proposta deste trabalho. A Tabela 40 apresenta de forma sumarizada os resultados gerais obtidos referentes às amostras analisadas das bases de dados *Breast Cancer Wisconsin (Diagnostic) Data Set*, *Heart Disease Data Set* e *Wine Quality Data Set*.

Tabela 40 – Resultado geral da média da porcentagem de acertos e da soma total de erros referente à análise em quatro bases de dados.

Bases de Dados	**Amostras**	**Média de porcentagem de acertos em cada grupo (%)**	**Erros**
Breast Cancer	683	93,16	36 (6,84%)
Heart Disease	297	98,56	4 (1,44%)
Wine Quality (Vinhos Vermelho)	1599	87,49	200 (12,51%)
Wine Quality (Vinhos Branco)	4898	88,83	547 (11,17%)
Resultado Geral	**7477**	**92,01**	**787 (10,52%)**

Os resultados alcançados (ver Tabela 40) foram bastante significativos, levando a crer que o modelo RDB pode ser utilizado para formulação de rótulos em amostras de outras bases de dados.

Por fim, vale ressaltar que o modelo RBD atingiu média de percentual de acertos

abaixo de 90% na formulação de rótulos quando a base de dados é muito desbalanceada e o valor do parâmetro K (número de grupos) é alto. Com relação a quantidade de atributos e o número de elementos nas bases de dados, a proposta não teve redução de sua eficácia quando da formulação dos rótulos, sempre alcançando media de percentual de acertos acima de 90%.

Com os resultados acima, observa-se que os modelos comparados conseguiram atingir os objetivos inerentes às suas proposta. Constata-se que ambos têm capacidade de formular rótulos baseando-se nas diferenças de cada grupo, facilitando a análise sob o ponto de vista de um especialista.

6.2 Trabalhos Futuros

Como trabalhos futuros sugere-se: a) submeter ao modelo proposto a outras bases de dados com o valor do parâmetro K (número de grupos) alto, na tentativa de identificar e melhorar a eficiência na formulação de rótulos; b) submeter o modelo a outras bases de dados quem contenham atributos do tipo não numérico e/ou não possuam atributos classe; e c) utilizar um segundo algoritmo de aprendizagem de máquina não supervisionada baseado em distância para enriquecer e validar a proposta apresentada nesta investigação.

Referências

AGGARWAL, C. C.; REDDY, C. K. *Data Clustering: Algorithms and Applications*. 1. ed. [S.l.]: Chapman & Hall/CRC. ISBN 1466558210, 9781466558212, 2013. Citado na página 1.

ANNA, L. B.; ERHAN, G. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials. DOI: 10.1109/COMST.2015.2494502*, v. 18, p. 1153 – 1176, 2016. Citado na página 5.

ATILGAN, C.; NASIBOV, E. A memory efficient distributed fuzzy joint points clustering algorithm. *IEEE 10th International Conference on Application of Information and Communication Technologies (AICT). DOI: 10.1109/ICAICT.2016.7991729*, n. 10, p. 1–5, 2016. Citado na página 6.

CHANG, K.-S.; PEN, Y.-W.; CHEN, W.-M. Density-based clustering algorithm for gpgpu computing. *In IEEE: International Conference on Applied System Innovation (ICASI). DOI: 10.1109/ICASI.2017.7988545*, p. 774–777, 2017. Citado na página 6.

COPPIN, B. *Inteligência Artificial: tradução e revisão técnica Jorge Duarte Pires Valério*. 1. ed. [S.l.]: Rio de Janeiro: LCT. ISBN 9788521617297, 2010. Citado 2 vezes nas páginas 1 e 5.

CORTEZ, P. et al. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, v. 47, n. 4, p. 547–553, 2009. Citado 2 vezes nas páginas 18 e 58.

DETRANO, R. International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, v. 64, p. 304–310, 1989. Citado na página 55.

EBRU, A.; AKCAYOL, M. A. A comprehensive survey for sentiment analysis tasks using machine learning techniques. *IEEE: International Symposium on INnovations in Intelligent SysTems and Applications (INISTA). DOI: 10.1109/INISTA.2016.7571856*, p. 1 – 7, 2016. Citado na página 5.

FISHER, D. Improving inference through conceptual clustering. *In: Proceedings of the Sixth National Conference on Artificial Intelligence. AAAI Press. ISBN 0-934613-42-7*, v. 2, n. AAA'87, p. 461–465, 1987. Citado 2 vezes nas páginas 25 e 38.

GIMéNEZ-GóMEZ, P. et al. Analysis of free and total sulfur dioxide in wine by using a gas-diffusion analytical system with ph detection. *Food Chemistry*, v. 228, p. 518–525, 2017. Citado na página 59.

GONG, J.; KUANG, X.-H.; LIU, Q. Survey on software vulnerability analysis method based on machine learning. *IEEE First International Conference on Data Science in Cyberspace (DSC). DOI: 10.1109/DSC.2016.33*, p. 642 – 647, 2016. Citado na página 5.

GUERRERO, R. F.; CANTOS-VILLAR, E. Demonstrating the efficiency of sulphur dioxide replacements in wine: A parameter review. *Trends in Food Science & Technology*, v. 42, p. 27–43, 2015. Citado na página 59.

HANEN, A.; RIDHA, B. Exploiting machine learning strategies and rssi for localization in wireless sensor networks: A survey. *IEEE: 13th International Wireless Communications and Mobile Computing Conference (IWCMC). DOI: 10.1109/IWCMC.2017.7986447*, p. 1150 – 1154, 2017. Citado na página 5.

HOPFER, H.; HEYMANN, H. Judging wine quality: Do we need experts, consumers or trained panelists? *Food Quality and Preference - ISSN: 0950-3293*, v. 32, p. 221(13), 2014. Citado na página 62.

KUN, L. X. Z. et al. Protein function detection based on machine learning: Survey and possible solutions. *15th International Symposium on Parallel and Distributed Computing (ISPDC). DOI: 10.1109/ISPDC.2016.78*, p. 227–333, 2016. Citado na página 5.

LINDER, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, n. 4, p. 18–36, 2009. Citado 2 vezes nas páginas 6 e 9.

LOPES, L.; MACHADO, V.; RABELO. Automatic cluster labeling through artificial neural networks. *In: IEEE International Joint Conference on Neural Networks (IJCNN). [S.l.: s.n.]*, p. 762–769, 2014. Citado 7 vezes nas páginas 15, 9, 10, 11, 13, 37 e 55.

MACHADO, V. P.; RIBEIRO, V. P. F.; RABELO, R. A. L. Rotulação de grupos utilizando conjuntos fuzzy. *Simpósio Brasileiro de Automação Inteligente - SBAI*, 2015. Citado 3 vezes nas páginas 2, 11 e 37.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967. p. 281 – 297. Disponível em: <https://projecteuclid.org/euclid.bsmsp/1200512992>. Citado 2 vezes nas páginas 6 e 8.

MANGASARIAN, O. L.; WOLBERG, W. H. Cancer diagnosis via linear programming. *SIAM News*, v. 23, p. 1–18, 1990. Citado 2 vezes nas páginas 32 e 33.

MAŁGORZATA, C. et al. A complete gradient clustering algorithm for features analysis of x-ray images. *In: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg*, p. 15–24, 2010. Citado na página 28.

MEHMET, D. A survey of machine learning applications for energy-efficient resource management in cloud computing environments. *IEEE 14th International Conference on Machine Learning and Applications (ICMLA). DOI: 10.1109/ICMLA.2015.205*, p. 1185 – 1190, 2015. Citado na página 5.

OIV: International organisation of vine and wine. 2015. Disponível em: <http://www.oiv.int/en/technical-standards-and-documents>. Acesso em: 12 fev. 2018. Citado na página 59.

PARTH, M. et al. Survey of unsupervised machine learning algorithms on precision agricultural data. *IEEE: International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). DOI: 10.1109/ICIIECS.2015.7193070*, p. 1 – 8, 2015. Citado na página 5.

RASIM, A. et al. Batch clustering algorithm for big data sets. *IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, p. 1–4, 2016. Citado na página 1.

RIBEIRO, V. P. F. *Rotulação de grupos utilizando conjuntos fuzzy.* Dissertação — Universidade Federal do Piauí, 2016. Citado 8 vezes nas páginas 15, 17, 1, 2, 11, 12, 13 e 37.

RUSSEL, S. J.; NORVIG, P. *Inteligência Artificial*. 3. ed. [S.l.]: Rio de Janeiro: Elsevier Editora Ltda. ISBN 9788535237016, 2013. Citado na página 5.

SILVA, J. de N.; FILHO, A. de C.; SILVA, A. a. a. C. Automatic Detection of Masses in Mammograms Using Quality Threshold Clustering, Correlogram Function, and SVM. *J Digital Imaging*, v. 28, p. 323 – 337, 2015. Disponível em: <https://doi.org/10.1007/s10278-014-9739-3>. Citado na página 5.

Apêndices

APÊNDICE A – RESULTADOS

Outras bases de dados foram submetidas ao processo de análise de grupos de dados e produção de rótulos realizado pelo modelo de Rotulação Baseado em Distância (RDB). É relevante mencionar que todas as bases estão disponíveis no repositório digital UCI Machine Learning.

Os métricas aplicadas para análise dos resultados das rotulações foram as mesmas descritas no trabalho de Lopes, Machado e Rabelo (2014), detalhadas no Capítulo 5. Como o modelo apresentado neste trabalho utiliza um algoritmo de Aprendizagem de Máquina não Supervisionada, os atributos classe de todas as bases foram ignorados e os resultados das análises são descritos a seguir.

A.1 Base de dados *Heart Disease Data Set*

A base de dados *Heart Disease Data Set*[1], contempla amostras de registros relacionados a pacientes com doenças cardíacas. O repositório apresenta dados referentes a quatro instituições: Instituto de Cardiologia Húngaro, Budapeste; Hospital Universitário, Zurique, Suíça; Hospital Universitário, Basileia, Suíça e; Fundação Clínica *Cleveland* e Centro Médico V.A., *Long Beach*. Para efeito de análise somente a base de dados da última instituição foi utilizada.

A base de dados da Fundação Clínica *Cleveland* e Centro Médico V.A., possui 303 elementos e cada elemento contem 76 atributos incluindo o atributo classe (DETRANO, 1989). Excluindo o atributo classe, a análise realizada neste trabalho utilizou somente 13 atributos, pois segundo Detrano (1989) apenas estes eram relevantes e foram utilizados para publicação de seus experimentos. Por esse motivo, resolveu-se utilizar o mesmo subconjunto de atributos nesta pesquisa. Tomando como base as informações disponíveis no repositório, a Tabela 41 apresenta o significado e os valores possíveis para cada atributo.

De acordo a Tabela 41 o atributo classe (Num) determina o estado angiográfico da doença cardíaca. O valor 0 ($< 50\%$) indica que o paciente não tem problemas cardíacos e 1 ($> 50\%$) indica que o paciente tem algum enfermidade relacionada ao coração. Ressalta-se que dos 303 elementos inseridos na base de dados 6 foram ingnorados, pois apresentavam valores nulos para os atributos CA (Número de grandes vasos) e Thal (Frenquência cardíaca), com 4 e 2 elementos respectivamente.

Os parâmetros utilizados pelo modelo para a formulação de rótulos para a base de

[1] Heart Disease Data Set. Disponível em: <http://archive.ics.uci.edu/ml/ datasets/Heart+Disease>. Acesso em: 22 jan. 2018.

Tabela 41 – informações sobre o significado e valores possíveis para cada atributo para a base *Heart Disease Data Set*.

	Atributos	Significado	Domínio
1	Age	Idade em anos	
2	Sex	Sexo	0: Feminino 1: Masculino
3	CP	Chest pain type (tipo de dor no peito) (A angina define a dor no peito causada pelo enfraquecimento dos músculos do coração)	1: Angina tipica 2: Angina atípica 3: Sem dor anginal 4: Assintomático
4	Trestbps	Pressão arterial em descanso (em mm Hg na entrada do hospital)	
5	Chol	Colesterol (soro) em mg/dl	
6	FBS	Açúcar no sangue em jejum >120 mg / dl	0: Falso 1: Verdadeiro
7	Restecg	Resultados eletrocardiográficos em repouso	0: Normal 1: Anormalidade da onda ST-T (inversão de onda T e/ou elevação do segmento ST ou depressão >0,05 mV) 2: Provável ou definitiva hipertrofia ventricular esquerda segundo os critérios de Romhilt-Estes
8	Thalach	Frequência cardíaca máxima atingida	
9	Exang	Angina induzida ao exercício	0: Não 1: Sim
10	Oldpeak	Depressão ST induzida pelo exercício em relação ao descanso	
11	Slope	Inclinação do segmento ST (esforço máximo ao exercício)	
12	CA	Número de grandes vasos (0-3) coloridos por fluoroscopia	
13	Thal	Frequência cardíaca	3: Normal 6: Defeito fixado 7: Defeito reversível
14	Num (class)	Diagnóstico da doença cardíaca (estado angiográfica da doença)	0: <50% - estreitamento do diâmetro 1: >50% - estreitamento do diâmetro

dados *Heart Disease Data Set*, são apresentados na Tabela 42.

Tabela 42 – Parâmetros do modelo de rotulação aplicados a base *Heart Disease Data Set*.

Parâmetros	Valores
Número de Grupos	2
Grau de Seleção inicial	0,5
Incremento do Grau de Seleção	0,0001

De posse dos valores dos parâmetros disponíveis na Tabela 42, os rótulos produzidos pelo modelo de rotulação para a base *Heart Disease Data Set*, podem ser verificados na Tabela 43.

É possível identificar na Tabela 43 que pacientes que estão com o nível de colesterol no intervalo de $259mg/dl$ a $564mg/dl$ tem algum problema relacionado a doença cardíaca. Por outro lado, indivíduos que estão com o nível de colesterol no intervalo de $126mg/dl$ a $257mg/dl$ não tem problemas relacionado à saúde do coração. De acordo com o modelo proposto, o atributo Chol (Colesterol) é o mais relevante para identificação de pacientes com doenças relacionadas ao coração. Portanto, um especialista de posse de exames clínicos

Tabela 43 – Gupos e elementos associados aos respectivos rótulos - Interação #113 - GS = 0,5113.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**Elementos**	**Acertos (%)**	**Erros**
1	Chol	126 ∼ 257	186	98,92	2
2	Chol	259 ∼ 564	111	98,20	2

apropriados pode associar o nível de colesterol na corrente sanguínea de um individuo como um indício para possíveis problemas cardíacos.

A Tabela 43 também denota que 4 elementos não foram rotulados corretamente. Isso significa que eles não foram inseridos em nenhum intervalo de valores formulados pelo modelo RBD. Os intervalos com os respectivos atributos que não obedeceram os rótulos gerados, podem ser conferidos na Tabela 44.

Tabela 44 – Elementos não rotulados na base *Heart Disease Data Set*.

Atributos	**Intervalo (mg/dl)**	**Elemento**	**Grupos**
Chol	258 ∼ 260	2	1
Chol	258 ∼ 258	2	2

A Tabela 45 exibe os resultados obtidos referentes a métrica soma total de erros encontrados na base de dados analisada.

Tabela 45 – Resultado da métrica soma total de erros para a base *Heart Disease Data Set*.

Métrica	**RBD**
Média da porcentagem de acertos em cada grupo (%)	98,56
Total de erros	4

Com os valores alcançados, certifica-se que o modelo RBD atingiu percentual de acerto bastante expressivo. De um total de 297 amostras analisados, a proposta rotulou corretamente 293 elementos representando um percentual de acertos de $98,56\%$.

A.2 Base de dados *Wine Quality Data Set*

A base de dados *Wine Quality Data Set*[2], está subdividida em duas bases de dados contendo exemplos de tipos de vinhos vermelho e variantes do vinho branco Português "Vinho Verde", produzidos no norte de Portugal. A primeira contem 1599 e a segunda 4898

[2] Wine Quality Data Set. 2009. Disponível em: <http://archive.ics.uci.edu/ml/ datasets/Wine+Quality>. Acesso em: 17 jan. 2018.

elementos. O objetivo deste repositório é modelar a qualidade dos vinhos com base em propriedades físico-químicas (CORTEZ et al., 2009). De acordo com o autor as entradas incluem valores como, por exemplo, pH e teor alcoólico dos vinhos, e a saída é baseada em dados sensoriais (média de pelo menos 3 avaliações feitas por especialistas em vinhos). Os *experts* classificaram a qualidade dos vinhos em uma escala que varia de 0 (muito ruim) a 10 (muito excelente).

Cada elemento das bases de dados é formado por 12 atributos conforme descritos na Tabela 46. O atributo Qualidade é designado como atributo classe, baseia-se em dados sensoriais, variando de 0 a 10, e para efeito de análise foi ignorado.

Tabela 46 – Atributos com as propriedades físico-químicos da base *Wine Quality Data Set* - Adaptado de (CORTEZ et al., 2009).

	Atributos	**Significado**	**Propriedades/Domínios**
1	AF	Acidez Fixa	(g(ácido tartárico)/dm^3)
2	AV	Acidez Volátio	(g(ácido acético)/dm^3)
3	AC	Ácido Cítrico	(g/dm^3)
4	AR	Açucar Residual	(g/dm^3)
5	NaCl	Cloreto de Sódio	(g(cloreto de sódio)/dm^3)
6	DEL	Dióxido de Enxofre Livre	(mg/dm^3)
7	DET	Dióxido de Enxofre Total	(mg/dm^3)
8	D	Densidade	(g/cm^3)
9	pH	pH	-
10	K	Sulfato de Potássio	(g(sulfato de potássio)/dm^3)
11	A	Álcool	(% teor alcoólico)
12	Qualidade	Atributo classe	Intervalo: 0 - 10

A base de dados de vinhos vermelho não possuí amostras classificadas nos intervalos de [0 a 2 e 9 a 10]. A base de vinhos branco, por sua vez, não possuí amostras no intervalo de [0 a 2 e 10]. Por não terem amostras representando todos os intervalos de valores que determinam a qualidade dos vinhos, os parâmetros k e GS submetidos ao modelo receberam os valores fixados na Tabela 47.

Tabela 47 – Parâmetros do modelo de rotulação aplicados a base *Wine Quality Data Set*.

Parâmetros	**Valores**
Número de Grupos	6
Grau de Seleção inicial	0,3
Incremento do Grau de Seleção	0,0001

De acordo com testes realizados quando a base de dados é muito desbalanceada e o valor de k é alto, menor é o grau de pertinência atribuído aos elementos em relação aos seus grupos. Esse fato ocorre devido a proximidade que os pontos ficam em relação aos centroides atribuídos pelo algoritmo *K-Means*. Portanto, para rotular os elementos foi preciso reduzir o valor do grau de seleção, pois nenhum elemento das bases de dados após

a conversão de distância para grau de pertinência, tiveram o valor do GS maior ou igual a 0,5.

Os resultados das análises para as bases de dados contidas no repositório *Wine Quality Data Set*, podem ser verificados a seguir.

A.2.1 Base de dados Vinhos Vermelho

De acordo com os valores dos parâmetros disponibilizados na Tabela 47, os rótulos produzidos pelo modelo de rotulação para a base de dados de vinhos vermelhos, são exibidos na Tabela 48.

Tabela 48 – Gupos e elementos associados aos respectivos rótulos - Interação #1133 - GS = 0,4133.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos (mg/dm^3)**	**Elementos**	**Acertos (%)**	**Erros**
1	DET	42 ∼ 57	302	83,77	49
2	DET	61 ∼ 78	189	80,42	37
3	DET	87 - 110	168	74,40	43
4	DET	25 ∼ 40	405	87,90	49
5	DET	124 - 165	68	77,94	15
6	DET	6 ∼ 24	467	98,50	7

Ao explorar a Tabela 48 verifica-se que o atributo Dióxido de Enxofre Total (DET) está presente em todos os grupos. Essa condição demonstra que os grupos se diferem pela faixa de valores desse atributo. Um especialista que eventualmente quiser atribuir um novo elemento a um grupo qualquer teria na característica DET do vinho o principal diferencial para a identificação do novo elemento.

Apesar de apresentar riscos a saúde humana, o dióxido de enxofre (SO_2) é o conservador mais utilizado na indústria do vinho. Suas propriedades antioxidantes e antimicrobianas torna-o essencial na produção de vinhos (GIMéNEZ-GóMEZ et al., 2017). O SO_2 tem sido usado para inibir a atividade da polifenoloxidase durante a vinificação, bem como para controlar o aparecimento de fermentações indesejáveis, como a fermentação acética ou malolática (GUERRERO; CANTOS-VILLAR, 2015).

Altas concentrações de dióxido de enxofre afetam a qualidade final do vinho, principalmente o cheiro e o sabor e podem inibir a fermentação malolática. O nível máximo de dióxido de enxofre total e livre é fixado na Comunidade Europeia pela Organização Internacional da Vinha e do Vinho (OIV) e depende do tipo de vinho (até 150 mg L^1 para vinhos tintos e até 400 mg L^1 para vinhos branco doce) (GIMéNEZ-GóMEZ et al., 2017). Se o teor total de dióxido de enxofre exceder 10 mg L^1, deve ser expresso no rótulo da garrafa de vinho (OIV, 2015).

Conforme observado na Tabela 48, com exceção do Grupo 5, os níveis de SO_2 presente nos intervalos estão dentro do padrão definido pela Organização Internacional da Vinha e do Vinho. Isso significa que o modelo RBD conseguiu rotular as amostras, qualificando-as de acordo com nível de SO_2 presente em cada elemento.

A Tabela 48 também expõe que alguns elementos (200, representando $12,51\%$) não foram rotulados corretamente. Isso significa que eles não foram inseridos em nenhum intervalo de valores formulados pelo modelo RBD. Os intervalos com os respectivos atributos que não obedeceram os rótulos gerados, podem ser conferidos na Tabela 49.

Tabela 49 – Elementos não rotulados na base Vinhos Vermelho.

Atributos	**Intervalos (mg/dm^3)**	**Elementos**	**Grupos**
DET	$35 \sim 41$ $58 \sim 62$	49	1
DET	$58 \sim 60$ $79 \sim 83$	37	2
DET	$84 \sim 86$ $111 \sim 119$	43	3
DET	$22 \sim 24$ $41 \sim 45$	49	4
DET	$115 \sim 123$ $166 \sim 289$	15	5
DET	$25 \sim 25$	7	6

Por fim, a Tabela 50 exibe a métrica da soma total de erros alcançada pelo modelo RBD para a base Vinhos Vermelho.

Tabela 50 – Resultado da métrica soma total de erros para a base Vinhos Vermelho.

Métrica	**RBD**
Média da porcentagem de acertos em cada grupo (%)	87,49
Total de erros	200

Logo, considerando que a base de dados de vinhos vermelho possui 1599 elementos e o modelo proposto rotulou corretamente 1399 elementos (87,49%), observa-se que a aplicação formulou faixas de valores expressivas para identificação da qualidade de vinhos existentes na base de dados.

A.2.2 Base de dados Vinhos Branco

Tomando como referencia os valores dos parâmetros disponibilizados na Tabela 47, os rótulos produzidos pelo modelo de rotulação para a base de dados de vinhos branco são exibidos na Tabela 51.

Tabela 51 – Gupos e elementos associados aos respectivos rótulos - Interação #767 - GS = 0,3767.

Grupos	**Rótulos**		**Análise**		
	Atributos	**Intervalos (mg/dm^3)**	**Elementos**	**Acertos (%)**	**Erros**
1	DET	206 ∼ 256	354	88,98	39
2	DET	140 ∼ 166	1086	83,89	175
3	DET	169 ∼ 200	847	86,78	112
4	DET	84 ∼ 111	991	94,15	58
5	DET	30 ∼ 83	385	92,72	28
6	DET	113 ∼ 137	1235	89,07	135

Da mesma forma como ocorreu com a base Vinhos Vermelho (Tabela 48), o atributo Dioxido de Enxofre Total (DET) é o que melhor representa os grupos formados na base Vinhos Branco. Esta condição pode auxiliar na interpretação dos rótulos, configurando uma característica determinante na classificação dos elementos.

É possível verificar na Tabela 51 que os níveis de SO_2 presentes nos intervalos estão dentro do limite fixado pela Organização Internacional da Vinha e do Vinho. Isso significa que o modelo RBD conseguiu rotular as amostras, qualificando-as de acordo com nível de SO_2 presente em cada elemento.

Como observado na Tabela 51, várias amostras (547, representando 11,17%) não foram rotuladas corretamente. Os atributos com os respectivos intervalos não rotulados, são visíveis na Tabela 52.

Tabela 52 – Elementos não rotulados na base Vinhos Branco.

Atributos	**Intervalos (mg/dm^3)**	**Elementos**	**Grupos**
DET	198 ∼ 205 257 ∼ 440	39	1
DET	129 ∼ 139 167 ∼ 178	175	2
DET	152 ∼ 168 201 ∼ 210	112	3
DET	78 ∼ 83 112 ∼ 114	58	4
DET	9 ∼ 29 84 ∼ 86	28	5
DET	106 ∼ 112 138 ∼ 145	135	6

Por fim, a Tabela 53 exibe a métrica da soma total de erros alcançada pelo modelo RBD para a base Vinhos Branco.

Assim sendo, considerando que a base de dados de vinhos branco possui 4898

Tabela 53 – Resultado da métrica soma total de erros para a base Vinhos Branco.

Métrica	**RBD**
Média da porcentagem de acertos em cada grupo (%)	88,83
Total de erros	547

elementos e o modelo proposto rotulou corretamente 4351 elementos (88,83%), tal qual como na base de dados de vinhos vermelho, percebe-se que a aplicação formulou faixas de valores expressivas para identificação da qualidade de vinhos existentes na base de dados.

Finalmente, dimensões intrínsecas à qualidade de vinhos podem ser definidas pela experiência do apreciador, independente do seu nível de conhecimento técnico, incluindo fatores como prazer, aroma, sabor e sensação na boca, aparência, bem como características que em muitos casos são mais importantes, como origem, variedade e tipicidade do vinho. Os fatores extrínsecos incluem o cultivo de uvas e a vinificação e, em um nível mais baixo, a "correção técnica", incluindo a definição mais básica de qualidade do vinho como a ausência de falhas e / ou a capacidade sensorial do degustador (HOPFER; HEYMANN, 2014).