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

Rotulação de grupos utilizando conjuntos fuzzy

Vilmar Pereira Ribeiro Filho

Teresina-PI, 29 de fevereiro de 2016

Vilmar Pereira Ribeiro Filho

Rotulação de grupos utilizando conjuntos fuzzy

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), 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: Prof. Dr. Vinícius Ponte Machado

Teresina-PI

29 de fevereiro de 2016

Vilmar Pereira Ribeiro Filho
Rotulação de grupos utilizando conjuntos *fuzzy*/ Vilmar Pereira Ribeiro Filho. – Teresina-PI, 29 de fevereiro de 2016-
35 p. : il ; 30 cm.

Orientador: Prof. Dr. Vinícius Ponte Machado

Dissertação (Mestrado) – Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação, 29 de fevereiro de 2016.

1. *Cluster*. 2. *Fuzzy*. I. Vinícius Ponte Machado. II. Universidade Federal do Piauí. III. Rotulação de grupos utilizando conjuntos *fuzzy*

CDU 02:141:005.7

Vilmar Pereira Ribeiro Filho

Rotulação de grupos utilizando conjuntos fuzzy

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

Trabalho aprovado. Teresina-PI, 29 de fevereiro de 2016:

Prof. Dr. Vinícius Ponte Machado
Orientador

Prof. Dr. Ricardo de Andrade Lira Rabêlo

Prof. Dr. Rodrigo de Melo Souza Veras

Prof. Dr. Ricardo Augusto Souza Fernandes

Teresina-PI
29 de fevereiro de 2016

Aos meus pais Rozângela Maria e Vilmar Ribeiro,
por sempre estarem comigo em todos os momentos.

Agradecimentos

Agradeço em primeiro lugar, a Deus, e a minha família.

Agradeço a meus pais, Vilmar Ribeiro e Rozângela Maria, por todo o carinho, atenção, amor, confiança, ensino e inspiração em toda a minha vida.

Agradeço ao meu orientador Vinícius Ponte Machado pela paciência e confiança. Foi um grande prazer e uma honra tê-los durante essa jornada. Meu eterno agradecimento.

Agradeço a todo os novos e velhos amigos do PPgCC, especialmente ao Lucas, Jonathas, Thiago e Kalyf.

Agradeço a minha irmã, Sara, à minha namorada, Nathylla, pela compreensão, carinho e paciência.

Agradeço a cada um de meus professores, de colégio e faculdade, pelos conhecimentos adquiridos, em especial aos professores Francisco Araújo, José Ferreira, Ricardo Sekeff, Ricardo Queiroz e Harilton Araújo, Rosianni e Amélia pela confiança e pelos ensinamentos além do curso.

Agradeço a todos que contribuíram direta ou indiretamente com a realização deste trabalho.

"A persistência é o caminho do êxito."
(Charles Chaplin)

Resumo

O agrupamento (*clustering*) de dados tem sido considerado como um dos tópicos mais relevantes dentre aqueles existentes na área de aprendizagem de máquina não supervisionada. Embora o desenvolvimento e aprimoramento de algoritmos que tratam esse problema tenham sido o principal foco de muitos pesquisadores, a compreensão dos grupos (*clusters*) é tão importante quanto sua formação. Definir um grupo pode ajudar na sua compreensão. Por exemplo, ao se encontrar uma definição para grupos consumidores é possível saber quais as principais diferenças entre os grupos e tomar decisões direcionadas para cada um deles. Frente ao problema de encontrar definições, também chamadas de rótulos, capazes de identificar cada grupo de forma fácil, este trabalho descreve um modelo que elabora rótulos utilizando a 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. Para avaliar o desempenho, o modelo produziu rótulos para grupos de três bases de dados e foi submetido a uma análise comparativa com um modelo de rotulação. Os rótulos produzidos conseguiram representar um grande número de elementos, apresentando assim um bom resultado. Na comparação, o modelo conseguiu produzir rótulos mais fáceis de serem compreendidos. Os experimentos realizados demonstram que o modelo proposto é capaz de construir rótulos para a identificação dos grupos, melhorando assim a compreensão dos grupos fornecidos.

Palavras-chaves: Aprendizagem. *Cluster*. *Fuzzy*. Rotulação.

Abstract

The clustering of data has been regarded as one of the most relevant topics among those existing in unsupervised machine learning area. Although the development and improvement of algorithms that address this issue have been the main focus of many researchers, understanding the clusters is as important as your training. Define a cluster can help in your understanding, for example, to find a definition for consumer groups is possible to know what the main differences between the cluster and make decisions directed to each of them. Facing the problem of finding definitions also called labels, able to identify each easily cluster, this paper describes a model that produces labels using the theory fuzzy sets to find relevant characteristics of the elements of each cluster and model ranges values that uniquely identify the clusters. To evaluate the performance of the model produced labels for clusters of three databases and was subjected to a comparative analysis with a labeling template. The labels produced managed to represent a large number of elements, thus presenting a good result. In comparison, the model was able to produce labels easier to be understood. The experiments demonstrate that the model is capable to build labels for the identification of clusters, thereby enhancing the understanding of the provided clusters.

Keywords: Learning. *Cluster*. *Fuzzy*, Labeling.

Lista de ilustrações

Figura 1 – Fluxograma do Modelo de Rotulação. 11
Figura 2 – Gráfico de dispersão da base de dados fictícia. O gráfico exibe os elementos que obedecem cada um dos rótulos e aqueles que não foram compatíveis com nenhum rótulo. 35

Lista de tabelas

Tabela 1 – Base de Dados Fictícia. 12
Tabela 2 – Matriz U, saída do algoritmo *Fuzzy C-Means*. 13
Tabela 3 – Elementos escolhidos no grupo 1. 14
Tabela 4 – Elementos escolhidos no grupo 2. 14
Tabela 5 – Elementos escolhidos no grupo 3. 15
Tabela 6 – Rótulos da Iteração #1. 15
Tabela 7 – Rótulos da Iteração #319. 16
Tabela 8 – Rótulos Finais. 16
Tabela 9 – Resultados da base de dados Iris. 18
Tabela 10 – Elementos não rotulados da base de dados Iris. 18
Tabela 11 – Resultados da base de dados Seed. 19
Tabela 12 – Resultados da base de dados Glass. 20
Tabela 13 – Parâmetros do Modelo Rotulação Neural aplicados a base de dados Iris. 21
Tabela 14 – Parâmetros do Modelo Rotulação *Fuzzy* aplicados a base de dados Iris. 21
Tabela 15 – Rótulos da base de dados Iris produzidas pelo Modelo Rotulação Neural. 22
Tabela 16 – Rótulos da base de dados Iris produzidas pelo Modelo Rotulação *Fuzzy*. 22
Tabela 17 – Resultado das métricas para a base de dados Iris. 22
Tabela 18 – Parâmetros do Modelo Rotulação Neural aplicados a base de dados Glass. 23
Tabela 19 – Parâmetros do Modelo Rotulação *Fuzzy* aplicados a base de dados Glass. 23
Tabela 20 – Rótulos da base de dados Glass produzidas pelo Modelo Rotulação Neural. 24
Tabela 21 – Rótulos da base de dados Glass produzidas pelo Modelo Rotulação *Fuzzy*. 24
Tabela 22 – Resultado das métricas para a base de dados Glass. 25
Tabela 23 – Parâmetros do Modelo Rotulação Neural aplicados a base de dados Seed. 25
Tabela 24 – Parâmetros do Modelo Rotulação *Fuzzy* aplicados a base de dados Iris. 25
Tabela 25 – Rótulos da base de dados Seed produzidas pelo Modelo Rotulação Neural. 26
Tabela 26 – Rótulos da base de dados Seed produzidas pelo Modelo Rotulação *Fuzzy*. 26
Tabela 27 – Resultado das métricas para a base de dados Seed. 26

Lista de abreviaturas e siglas

At	Atributo
GP	Grau de Pertinência
GS	Grau de Seleção
IGS	Incremento do Grau de Seleção

Sumário

Introdução 1
Motivação 1
Proposta 1
Objetivos 2
Estrutura Organizacional 2

1 REFERENCIAL TEÓRICO 3
1.1 Aprendizagem de máquina 3
1.2 Agrupamento (Clustering) 3
1.3 Hard C-means 4
1.4 Fuzzy C-means 6
1.5 Trabalhos relacionados 8

2 MODELO DE ROTULAÇÃO PROPOSTO 11
2.1 Modelo de Rotulação 11

3 RESULTADOS 17
3.1 Detalhes da Implementação 17
3.2 Base de Dados Iris 17
3.3 Base de Dados Seed 19
3.4 Base de Dados Glass 19
3.5 Análise Comparativa 20
3.5.1 Base de dados Iris 21
3.5.2 Base de dados Glass 23
3.5.3 Base de dados Seed 25

Conclusão e Trabalhos Futuros 29

REFERÊNCIAS 31

APÊNDICE A – GRÁFICO DE DISPERSÃO DA BASE DE DADOS FICTÍCIA 35

Introdução

Com o surgimento crescente de novas tecnologias, o número de dados produzido é cada vez maior. Uma das maneiras de lidar com essa enorme quantidade de dados é por meio de agrupamentos. As pessoas buscam agrupar dados com a finalidade de extrair características que sejam capazes de descrevê-los e ainda compará-los (XU; WUNSCH II, 2005). O problema básico do agrupamento pode ser declarado como segue: dado um conjunto de dados, particionar em grupos os elementos que são tão semelhantes quanto possível. Note que esta é uma definição muito simples, e as variações na definição do problema podem ser significativas, dependendo do modelo utilizado (AGGARWAL; REDDY, 2013).

Motivação

O agrupamento (*clustering*) de dados tem sido considerado como um dos tópicos mais relevantes dentre aqueles existentes na área de aprendizagem de máquina e mineração de dados (AGGARWAL; REDDY, 2013). Assim, embora o desenvolvimento e aprimoramento de algoritmos que enfrentam esse problema tenham sidos foco de muitos pesquisadores, poucos trabalhos lidam explicitamente com a interpretação dos *clusters*.

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. Como exemplo, podemos citar uma base de dados de clientes, na qual a análise para saber quais as características determinantes em um grupo pode servir para a tomada de decisões na empresa.

Proposta

Este trabalho descreve um modelo capaz de analisar os *clusters* e produzir definições que também podem ser chamadas de rótulos. Na formação dos rótulos o modelo utiliza o algoritmo não supervisionado *Fuzzy C-means* para elaborar a matriz U. Esta matriz é composta pelos elementos da base de dados associados aos seus graus de pertinência em cada grupo. O modelo utiliza esta matriz para selecionar os elementos relevantes em cada grupo. Estes elementos 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, o modelo a cada iteração seleciona elementos com grau de pertinência maiores, com o objetivo de selecionar elementos mais relevantes e eliminar as interseções entre as faixas de valores, identificando assim cada grupo.

Objetivos

O objetivo desse trabalho consiste em apresentar um modelo de rotulação capaz de identificar características únicas em cada grupo, facilitando assim a sua compreensão. Para isto é necessário aplicar o algoritmo *Fuzzy C-Means*, 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. Este trabalho também tem como objetivo, avaliar o modelo por meio da quantidade de elementos que cada rótulo é capaz de representar e realizar uma análise comparativa com o trabalho proposto por Lopes, Machado e Rabelo (2014).

Estrutura Organizacional

Este trabalho está organizado como segue: o Capítulo 1 apresenta as principais teorias utilizadas no modelo e os trabalhos relacionados, o Capítulo 2 apresenta o modelo proposto. O Capítulo 3 apresenta os detalhes da implementação, os resultados obtidos com o modelo, uma análise comparativa dos rótulos gerados e por fim as conclusões e os trabalhos futuros.

1 Referencial Teórico

Este capítulo discute, primeiramente, sobre aprendizagem de máquina e agrupamento de dados, em seguida descreve os algoritmos *Hard C-means* e *Fuzzy C-means*, por último aborda os trabalhos relacionados com a pesquisa.

1.1 Aprendizagem de máquina

Segundo Mitchell (1997), a área de aprendizagem de máquina está preocupada em construir programas de computador que possam melhorar seus desempenho de forma automática por meio de experiências. Para atingir seus resultados a Aprendizagem de Máquina utiliza-se de outras áreas de conhecimento como: Inteligência Artificial, Probabilidade, Estatística, Psicologia e Biologia.

Os programas de computador que utilizam modelos de Aprendizagem de Máquina são capazes de formular hipóteses por meio de um conjunto de experiências anteriormente adquiridas. O programa deve passar por uma aprendizagem pelo qual ele adquirirá tal experiência. Nós podemos dividir o processo de aquisição da experiência em dois grandes paradigmas: supervisionado e não-supervisionado. Ambos paradigmas realizam a busca por um modelo capaz de generalizar os dados. A aprendizagem supervisionada se destaca pela utilização de dados rotulados enquanto a aprendizagem não-supervisionada utiliza dados não rotulados. Existe também uma abordagem semi-supervisionada na qual consiste em uma tentativa de aprimorar um classificador criado a partir de dados rotulados com o uso de amostras não rotuladas (BARBER, 2012).

1.2 Agrupamento (*Clustering*)

Segundo Oliveira e Pedrycz (2007), *Clustering* é uma tarefa de aprendizagem não-supervisionada que visa a decomposição de um dado conjunto de objetos em subgrupos ou grupos com base na similaridade. O objetivo é dividir os dados de tal maneira que os objetos ou conjunto de dados que pertencem ao mesmo grupo sejam tão semelhantes quanto possível, enquanto objetos pertencentes a diferentes grupos sejam tão diferentes quanto possível.

A análise de agrupamento é principalmente uma ferramenta para descobrir estrutura escondida em um conjunto de objetos. Neste caso, supõe-se que o agrupamento existe nos dados. Ao organizar objetos semelhantes em *clusters* se tenta reconstruir a estrutura desconhecida na esperança de que cada aglomerado encontrado represente um tipo real ou

uma categoria de objetos. Métodos de *Clustering* também podem ser utilizados para fins de redução de dados. Alguns critérios matemáticos podem ser utilizados para decidir sobre a composição de *clusters*, assim como classificar conjuntos de dados automaticamente. Portanto, os métodos de *Clustering* são dotados de funções de distância que medem a dissimilaridade de casos de exemplo apresentados, o que é equivalente a medir a sua semelhança. O agrupamento de dados pode ser realizado utilizando-se de várias técnicas como: Cobweb (FISHER, 1987), *Self Organizing Maps* (SOM) (FIGUEIREDO S. BOTELHO; HAFFELE, 2012), Redes Neurais Artificiais (AZIZ et al., 2012), *K-means* (KANUNGO et al., 2002), *Hard C-means* (PERES et al., 2012), *Fuzzy C-means* (RAMATHILAGA; LEU; HUANG, 2011) entre outras.

Algoritmos de particionamento visam encontrar a melhor partição dos grupos de dados com base na medida de dissimilaridade dada. Métodos de particionamento de agrupamento são diferentes das técnicas hierárquicas. Este último organiza os dados em uma sequência aninhada dos grupos, que podem ser visualizados na forma de um dendrograma ou árvore. Com base em um dendrograma pode-se decidir sobre o número de grupos na qual os dados são melhores representados para uma dada finalidade. Geralmente ao usar um método de particionamento é necessário especificar o número de *clusters* como um parâmetro de entrada. Estimar o número real de *clusters* é, portanto, uma questão importante (OLIVEIRA; PEDRYCZ, 2007).

Um conceito comum em todas as abordagens de agrupamento, é que eles são baseados em protótipos, ou seja, os grupos são representados por protótipos de *cluster*. Os protótipos são usados para capturar a estrutura (distribuição) de dados em cada grupo. Com essa representação dos *clusters* denotamos formalmente como conjunto de protótipos. Cada protótipo é um n-tuplo de parâmetros que consistem no centro do *cluster* (parâmetro de localização) e talvez alguns parâmetros adicionais sobre o tamanho e a forma do agrupamento. Os protótipos serão apresentados na sua forma mais simples. Cada protótipo consiste apenas dos vetores do centro, de tal modo que os pontos de dados (elementos) atribuído a um *cluster* são representados por um ponto no espaço de dados.

Os algoritmos que serão descritos utilizam uma função objetivo, representada por J. A função objetivo deve ser minimizada para obter as melhores soluções para cada *cluster*. Tendo definido um critério desse tipo, a tarefa de agrupamento pode ser formulada como um problema de otimização.

1.3 *Hard C-means*

No modelo *Hard C-means* cada ponto de dados x_j em um conjunto de dados $X = \{x_1, ..., x_n\}$, onde $X \subseteq \mathbb{R}^p$ é atribuído a um *cluster*. Cada *cluster* Γ_i é assim, um subconjunto do conjunto de dados $\Gamma_i \subset X$. O conjunto de *clusters* $\Gamma = \{\Gamma_1, ..., \Gamma_c\}$ é uma

partição dos conjunto de dados X em c conjuntos, não vazios com pares de subconjuntos disjuntos onde $\Gamma_i, 1 < c < n$. No *Hard C-means* uma partição de dados é dito ótima quando a soma dos quadrados das distâncias entre os centros dos grupos e os pontos de dados que lhes são atribuídas é mínima (KRISHNAPURAM; KELLER, 1996). A função objetivo do *Hard C-means* pode ser escrita da seguinte forma:

$$J_h(X, U_h, C) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} d_{ij}^2 \tag{1.1}$$

Onde $C = \{C_i, ..., C_c\}$ é o conjunto de protótipos, d_{ij} é a distância entre x_j e o centro do agrupamento c_i, U_h é uma matriz $c \times n$ binária chamada matriz de partição, tal que:

$$u_{ij} \in \{0, 1\} \tag{1.2}$$

A atribuição de um ponto de dado para os *clusters* é dada como: $u_{ij} = 1$ se o x_j é atribuído ao protótipo C_i, ou seja, $x_j \in \Gamma_i$; e $u_{ij} = 0$ caso contrário. Para assegurar que cada ponto de dados é atribuído exatamente a um conjunto, é necessário que:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1, ..., n\} \tag{1.3}$$

Esta restrição impõe partições completas e também serve o propósito de evitar a solução trivial minimizando J_h, que é, não atribuir dados a *clusters*: $u_{ij} = 0, \forall i, j$. Com isso, uma outra restrição também é necessária. A restrição (1.4) evita que seja possível ter *clusters* vazios.

$$\sum_{j=1}^{n} u_{ij} > 0, \forall i \in \{1, ..., c\} \tag{1.4}$$

O problema de encontrar parâmetros que minimizem a função objetivo é *NP-hard* (DRINEAS et al., 2004). Portanto não é garantido que o ótimo global será alcançado. No caso do *Hard C-means* o regime de otimização iterativa funciona a partir da escolha de c vetores aleatórios, outros métodos de inicialização mais sofisticados também podem ser usados (MCKAY; BECKMAN; CONOVER, 1979). Em seguida, com os parâmetros C definidos, atribuições em U_h são determinadas para minimizar a quantidade de J_h. Nesta etapa, cada ponto de dados é atribuído ao centro do *cluster* mais próximo:

$$u_{ij} = \begin{cases} 1, & \text{se } i = \text{argmin}_{l=1}^{c} d_{lj} \\ 0, & \text{caso contrário} \end{cases} \tag{1.5}$$

Após as atribuições, a partição de dados U_h é definida, e novos centros dos *clusters* são calculados com base nos ponto de dados atribuídos ao *cluster*. O cálculo do centro do *cluster* é indicado mais formalmente como:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij} x_j}{\sum_{j=1}^{n} u_{ij}}. \tag{1.6}$$

Em seguida, o *Hard C-means* termina, dando origem a centros de *clusters* finais e a partição de dados, que possivelmente é um ótimo local. Concluindo o *Hard C-means*, expressa a tendência de ficar preso em mínimos locais, o que torna necessário a realização de várias execuções do algoritmo com diferentes inicializações (DUDA; HART, 1973). Em seguida, o melhor resultado de agrupamentos pode ser escolhido com base nos valores de J_h.

Passaremos agora para as abordagens *Fuzzy*, que relaxam a exigência $u_{ij} \in \{0, 1\}$ que é colocada sobre as atribuições nas abordagens clássicas (*hard*) de *Clustering*.

1.4 *Fuzzy C-means*

O agrupamento *Fuzzy* permite que os dados tenham graus de pertinência em relação a seus *clusters*, variando de $[0, 1]$. Isto dá a flexibilidade para expressar pontos de dados que podem pertencer a mais de um *cluster*. Além disso, estes graus de pertinência podem ser ajustados para fornecer um detalhamento mais fino do modelo de dados. Além de associar o ponto de dados aos *clusters*, os graus de pertinência também podem expressar o quão ambíguo ou diferente um ponto de dados é de um *cluster*. O conceito destes graus de pertinência é definido pela teoria dos conjuntos *Fuzzy* (ZADEH, 1965). Assim, agrupamentos *Fuzzy* permitem espaços de solução ajustáveis, em forma de partições do conjunto de dados. Considerando o conjunto de dados $X = \{x_1, ..., x_n\}$, e o *cluster* Γ_i, de partições formadas por subconjuntos clássicos, a sua representação pelos conjuntos *Fuzzy*, e dada com μ_{Γ_i} dos dados do conjunto X. De acordo com a teoria dos conjuntos *Fuzzy*, u_{ij} é o grau de pertinência de um x_j a um *cluster* Γ_i, de tal forma que: $u_{ij} = \mu_{\Gamma_i}(x_j) \in [0, 1]$. Os métodos de agrupamento *Fuzzy* atribuem um vetor para cada ponto de dados x_j que indica suas associações para os c *clusters*.

A matriz $U_f = (u_{ij}) = (u_1, ..., u_n)$, $c \times n$, é chamada matriz de partição *Fuzzy*. Com base na notação de conjuntos *Fuzzy*, os *clusters* são mais adequados para lidar com ambiguidade.

Sendo $X = \{x_1, ..., x_n\}$ o conjunto de dados e c o número de *clusters* $(1 < c < n)$ representado pelos conjuntos *Fuzzy* $\mu_{\Gamma_i}, (i = 1, ..., c)$, então chamamos $U_f = (u_{ij}) =$

$(\mu_{\Gamma_i}(x_j))$ partição de X se:

$$\sum_{j=1}^{n} u_{ij} > 0, \forall i \in \{1, ..., c\} \tag{1.7}$$

e

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1, ..., n\} \tag{1.8}$$

Os $u_{ij} \in [0, 1]$ são interpretados como o grau de pertinência de um x_j para o *cluster* Γ_i relativo a todos os outros *clusters*.

As restrições (1.7) e (1.8) têm de ser satisfeitas para graus de pertinência em U. A condição (1.7) garante que não exista um agrupamento vazio e evita a solução trivial do problema de minimização, ou seja, $u_{ij} = 0, \forall i, j$. A restrição de normalização (1.8) conduz a uma "distribuição" do grau de pertinência de cada ponto de dados nos diferentes conjuntos. A condição de normalização implementa a propriedade de particionamento conhecido de qualquer algoritmo de agrupamento *Fuzzy*.

Depois de definir partições podemos voltar para o desenvolvimento de uma função objetivo. Certamente, quanto mais próximo um ponto de dados encontra-se do centro de um *cluster*, maior o seu grau de pertinência deve ser para este *cluster*. Seguindo esse raciocínio, pode-se dizer que as distâncias entre os centros dos grupos e os pontos de dados que lhe são atribuídas deve ser mínima. O problema de dividir um conjunto de dados em c *clusters* pode ser indicado como a tarefa de minimizar as distâncias ao quadrado dos pontos de dados para seus centros de *cluster*, ou seja, maximizar os graus de pertinência. A função objetivo J_f baseia-se assim na menor soma das distâncias ao quadrado.

Formalmente, um modelo de grupos *Fuzzy* é dado como um conjunto de dados X em c *clusters* e definido como ótimo quando minimiza a função objetivo:

$$J_f(X, U_f, C) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \tag{1.9}$$

O parâmetro $m, m > 1$, é chamado de *fuzzifier* ou expoente de ponderação. A exponenciação com m em J_f pode ser visto como uma função g nos graus de pertinência, $g(u_{ij}) = u_{ij}^m$. Para o caso de $m = 1$ (tornando J_h e J_f idênticos), as atribuições do *cluster* permanecem semelhando ao *Hard C-means*. (DUNN, 1973). A generalização para exponentes $m > 1$ tem sido proposta em (BEZDEK, 1973). Valores mais elevados para o m tornam os limites entre os grupos mais suaves, com valores menores os grupos se tornam *crisp*. Normalmente $m = 2$ é escolhido. Além da ponderação padrão para os graus de pertinência u_{ij}^m outras funções g *fuzzifiers* podem ser exploradas.

A função objetivo J_f é otimizada por meio das equações (1.10) e (1.11), os graus de pertinência são otimizados com os protótipos definidos, em seguida, os protótipos são otimizados por meio dos graus de pertinência.

Os graus de pertinência têm de ser escolhidos de acordo com a seguinte fórmula, que é independente da medida de distância escolhida (BEZDEK, 1981):

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{ij}^2}{d_{lj}^2} \right)^{\frac{1}{m-1}}} = \frac{d_{ij}^{-\frac{2}{m-1}}}{\sum_{l=1}^{c} d_{lj}^{-\frac{2}{m-1}}} \tag{1.10}$$

No caso de existir um *cluster* i com distância zero de um x_j, $u_{ij} = 1$ e $u_{lj} = 0$ para todos os outros grupos $l \neq i$. A equação (1.10) mostra o caráter relativo do grau de pertinência. Ele não depende apenas da distância do ponto de referência x_j para o *cluster* i, mas também as distâncias entre este ponto de dado e os outros *clusters*.

A fórmula para calcular os centros dos *clusters* é dada como (BEZDEK, 1981):

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}. \tag{1.11}$$

1.5 Trabalhos relacionados

Durante o levantamento bibliográfico alguns trabalhos pesquisados tiveram semelhanças com esta pesquisa e somente um trata do mesmo problema dessa pesquisa.

Os trabalhos de Glover et al. (2002), Chuang e Chien (2004), Popescul e Ungar (2000) abordam o mesmo problema, porém utilizando abordagens diferentes. Eles tratam da questão de extrair tópicos de textos e organizá-los de forma hierárquica. A exemplo disso temos estes algoritmos processando textos de biologia e sendo capazes de organizar na forma de uma árvore hierárquica, tópicos como: ciência, biologia e botânica colocando cada uma dessas palavras em um nível hierárquico diferente, colocando também outros termos encontrado nos textos. Apesar destas pesquisas trabalharem com dados textuais, o que não é o foco desse trabalho, elas mostram a importância de condensar dados em estruturas que possam ser facilmente interpretadas.

Outros trabalhos, como mostrados em Eltoft e Figueiredo (1998) e Chen, Chuang e Chen (2008), se referem à rotulação de *clusters* como o problema de atribuir um elemento desconhecido a um *cluster*. Estes trabalhos apresentam bons resultados ao atribuir um novo elemento a um *cluster*, porém não demostram as regras utilizadas para estas atribuições.

Algoritmos como C4.5 e ID3, propostos por Quinlan (1986), constroem árvores de decisão, uma vez aplicados para encontrar regras para os *clusters*, pode ser utilizado como solução do problema, já que os algoritmos formam árvores de decisão e os caminhos até

as folhas podem ser considerados rótulos, porém esses algoritmos geram extensas árvores contendo várias condições, dificultando a interpretação do *cluster*.

Pode ser visto em Cintra et al. (2011), Liu, Feng e Pedrycz (2013) que alguns trabalhos se propõem a construir árvores de decisão por meio da extração de regras com lógica *Fuzzy*, entre eles os trabalhos de Setnes (1999), que constrói regras de lógica *Fuzzy* para classificar elementos em um *cluster*, este trabalho se assemelha a proposta dessa pesquisa porém gera árvores de decisão, e como já foi dito, grandes árvores de decisão podem dificultar o entendimento do *cluster*.

Os trabalhos Vargas, Bedregal e Filho (2009) e Vargas e Bedregal (2009) utilizam matemática intervalar para estender as funcionalidades do algoritmo *Fuzzy C-Means* porém seu objetivo é agrupar dados que estão no formato de intervalos e não tem o objetivo de expor as regras capazes de definir um *cluster*.

O trabalho Lopes, Machado e Rabelo (2014) aborda a mesma problemática desse trabalho, ou seja, 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 Lopes, Machado e Rabelo (2014) utiliza o algoritmo K-means, porém outros algoritmos de agrupamento podem ser utilizados. A discretização dos dados é realizada caso a base de dados possua valores contínuos, caso contrário, os valores não são alterados. O processo de discretização consiste em converter os valores dos atributos de forma contínua em valores discretos, para isso podem ser utilizadas duas abordagens, são elas: Discretização por Larguras Iguais (EWD), que consistem em dividir um intervalo em medidas iguais; Discretização por Frequências Iguais (EFD), que ira particionar 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) que especifica a quantidade de faixas de valores cada atributo ira ter.

Com os dados discretizados, o modelo submete os dados a uma rede neural, que tem o objetivo de extrair os atributos mais relacionados. A rede neural tenta inferir o valor de um atributo, dado todos os valores dos outros atributos. O processo de execução das redes neurais é realizado várias vezes, sendo determinadas pelo parâmetro M, também chamado de Número de Iterações por Atributo. Após esse processo os atributos que tiveram um número maior de valores inferidos de forma correta são ranqueados e selecionados. Esses atributos são selecionados de acordo com o parâmetro V, chamado Variação. Os atributos que possuem um valor acima desse parâmetro foram parte dos rótulos. Após a seleção dos atributos, os valores que mais se repetem são selecionados para fazer parte do rótulo. Caso a base de dados tenha sido discretizada, o valor que mais se repete ira representar a faixa de valores definida na discretização. O resultado de saída do modelo consiste em valores ou faixas de valores associados a seus respectivos atributos.

Este modelo possui a vantagem de ser aplicado em bases de dados com valores discretos ou contínuos, os grupos fornecidos ao modelo não se restringe a grupos formados por algoritmos não supervisionado. Por um outro lado o modelo possui parâmetros difíceis de serem ajustados exigindo um grande conhecimento da base dados e em alguns casos o modelo produziu rótulos muito semelhantes mesmo utilizando valores deferentes para o parâmetro Variação (V).

2 Modelo de Rotulação Proposto

Este capítulo descreve o funcionamento do modelo de rotulação. O modelo tem como objetivo formular faixas de valores, verificar a existência de interseção entre elas e montar os rótulos com faixas de valores que não possuem interseção. Para facilitar a descrição do modelo é demonstrada a aplicação do modelo em uma base de dados com valores fictícios.

2.1 Modelo de Rotulação

Para facilitar a compreensão do leitor, a descrição do funcionamento do modelo de rotulação é feita paralelamente com a sua aplicação em uma base de dados fictícia .A Figura 1 demonstra o modelo utilizando um fluxograma. As etapas demonstradas na figura são explicadas a seguir.

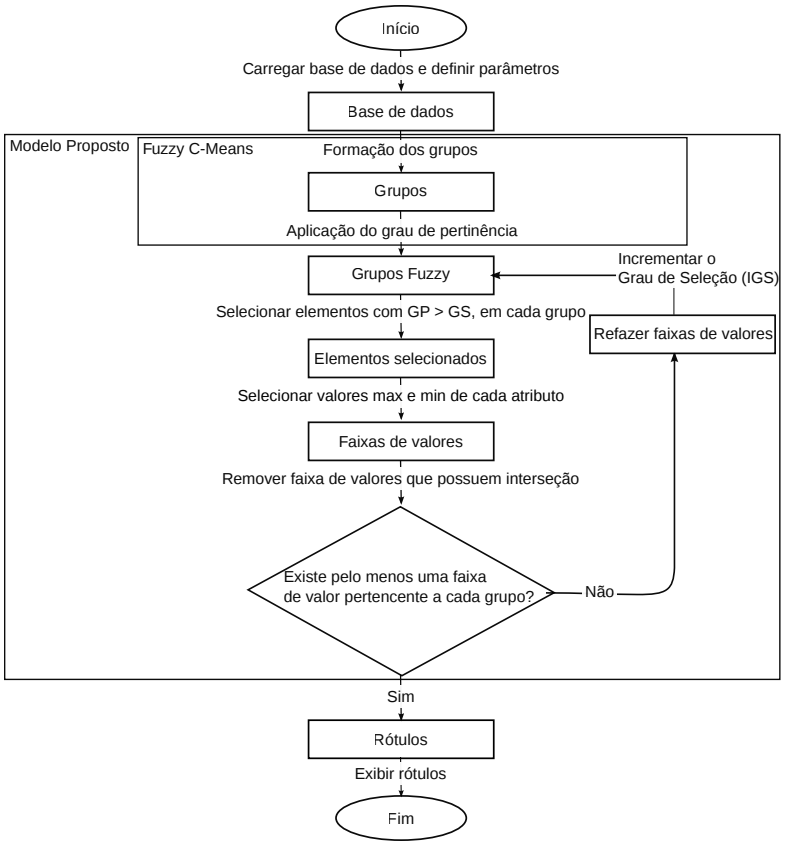

Figura 1 – Fluxograma do Modelo de Rotulação.

O modelo é iniciado com o carregamento da base de dados, a definição dos parâmetros GS (Grau de Seleção) e definição do IGS (Incremento do Grau de Seleção). O

parâmetro GS consiste em um número que serve de base para a seleção dos elementos mais significativos na formulação do rótulo e pode variar entre 0 e 1. O IGS consiste em um valor de incremento do parâmetro GS a cada iteração e também varia entre 0 e 1.

A base de dados fictícia é apresentada na Tabela 1. Os 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 do modelo. O parâmetro GS foi definido como 0,5 por representar um grau de pertinência intermediário, no qual elementos que estão abaixo desse número têm grande chances de pertencer a dois grupos. Uma vez que a matriz U possui 4 casas decimais, o parâmetro IGS foi definido como 0,0001, pois provoca no GS um ajuste na quarta casas decimal, facilitando a seleção dos elementos que compõem as faixas de valores.

Tabela 1 – Base de Dados Fictícia.

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

Inicialmente é aplicado o algoritmo de agrupamento não-supervisionado *Fuzzy C-Means*. Este algoritmo é inicializado com a quantidade de grupos a serem formados. Foi utilizado o valor 3 como parâmetro da quantidade de grupos no algoritmo *Fuzzy C-Means*.

O algoritmo *Fuzzy C-Means* fornece como saída 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 do centróide de um grupo, maior é seu grau de pertinência em relação ao grupo.

A matriz U da base de dados fictícia pode ser vista na Tabela 2. Como exemplo

temos o elemento de *id* igual a 10, que possui grau de pertinência maior no grupo 2. Isso significa que este elemento está mais próximo do centróide do grupo 2. De maneira análoga podemos perceber que este elemento possui grau de pertinência menor no grupo 3, significando que ele está mais distante do centróide do grupo 3.

Tabela 2 – Matriz U, saída do algoritmo *Fuzzy C-Means*.

	Grau de Pertinência		
Id	**Grupo 1**	**Grupo 2**	**Grupo 3**
1	0,1269	0,0535	0,8196
2	0,2383	0,6920	0,0697
3	0,0973	0,0458	0,8569
4	0,9758	0,0140	0,0102
5	0,0009	0,0007	0,9984
6	0,0286	0,0217	0,9497
7	0,1062	0,0565	0,8373
8	0,0401	0,0213	0,9386
9	0,8411	0,0598	0,0991
10	0,1329	0,8230	0,0441
11	0,0983	0,8230	0,0787
12	0,0890	0,8220	0,0890
13	0,3388	0,5891	0,0721
14	0,0549	0,0565	0,8886
15	0,0142	0,9764	0,0094
16	0,0476	0,9148	0,0376
17	0,8693	0,0920	0,0387
18	0,0954	0,8273	0,0773
19	0,0589	0,0561	0,8850
20	0,0135	0,9797	0,0068
21	0,7885	0,1567	0,0548
22	0,8425	0,0681	0,0894
23	0,2381	0,7011	0,0608
24	0,9791	0,0126	0,0083
25	0,9436	0,0261	0,0303
26	0,0944	0,1165	0,7891
27	0,9095	0,0576	0,0329
28	0,7602	0,0859	0,1539
29	0,0162	0,0117	0,9721
30	0,9871	0,0071	0,0058

Com a formação da matriz U, o modelo de rotulação proposto escolhe elementos que possuem um GP (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 as faixas de valores de cada grupo.

Os elementos escolhidos durante a primeira iteração do modelo podem ser vistos nas Tabelas 3, 4 e 5. Nelas estão presentes os valores selecionados com o grau de seleção (GS) igual a 0,5. As células das tabelas de cor cinza mostram os valores máximos e mínimos de cada atributo utilizado na formação das faixas de valores.

Tabela 3 – Elementos escolhidos no grupo 1.

			Grau de Pertinência		
Id	**At.1**	**At.2**	**Grupo 1**	**Grupo 2**	**Grupo 3**
4	7,0	8,0	0,9758	0,0140	0,0102
9	5,8	7,5	0,8411	0,0598	0,0991
17	7,8	7,7	0,8693	0,0920	0,0387
21	7,6	6,9	0,7884	0,1567	0,0548
22	5,9	8,5	0,8425	0,0681	0,0894
24	7,0	7,3	0,9792	0,0126	0,0083
25	6,3	7,4	0,9436	0,0261	0,0303
27	7,5	8,2	0,9095	0,0576	0,0329
28	5,9	6,9	0,7602	0,0859	0,1539
30	6,8	7,9	0,9872	0,0071	0,0058

Tabela 4 – Elementos escolhidos no grupo 2.

			Grau de Pertinência		
Id	**At.1**	**At.2**	**Grupo 1**	**Grupo 2**	**Grupo 3**
2	9,7	6,5	0,2383	0,6920	0,0697
10	9,4	6,0	0,1329	0,8229	0,0441
11	9,8	3,5	0,0983	0,8229	0,0787
12	8,0	4,0	0,0890	0,8220	0,0890
13	9,0	6,7	0,3388	0,5891	0,0721
15	8,8	4,5	0,0142	0,9764	0,0094
16	8,9	4,0	0,0476	0,9148	0,0376
18	7,8	4,6	0,0954	0,8273	0,0773
20	8,5	5,1	0,0135	0,9797	0,0068
23	9,1	6,4	0,2381	0,7011	0,0608

Tabela 5 – Elementos escolhidos no grupo 3.

			Grau de Pertinência		
Id	**At.1**	**At.2**	**Grupo 1**	**Grupo 2**	**Grupo 3**
1	4,3	6,0	0,1269	0,0535	0,8196
3	4,7	5,7	0,0973	0,0458	0,8569
5	4,3	4,7	0,0009	0,0007	0,9984
6	4,9	4,8	0,0286	0,0217	0,9498
7	3,1	5,7	0,1062	0,0565	0,8372
8	4,1	5,5	0,0401	0,0213	0,9386
14	4,5	3,7	0,0549	0,0565	0,8886
19	3,9	3,6	0,0589	0,0561	0,8850
26	5,3	3,8	0,0944	0,1165	0,7891
29	3,8	4,6	0,0162	0,0117	0,9721

A Tabela 6 representa as faixas de valores geradas, tendo por base os elementos selecionados.

Tabela 6 – Rótulos da Iteração #1.

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

Por fim, é verificado se existem interseções entre as faixas de valores pertencentes a um mesmo atributo. Caso exista interseção entre as faixas de valores, como mostrado na Tabela 6 em negrito, pelas faixas 5,80 ∼ 7,80 e 7,80 ∼ 9,80, que compartilham o número 7,80 em comum nas duas faixas, estas faixas são descartadas e a análise parte para outro conjunto de faixas de valores. Isto é necessário pois as faixas de valores que compõem interseção são ambíguas e incapazes de representar um único grupo.

Caso nenhum atributo possua pelo menos uma faixa de valor capaz de representar cada um dos grupos, como mostrado na Tabela 6, 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 a interseção 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.

Caso exista pelo menos uma faixa de valores sem interseção em cada grupo, como mostrado na Tabela 7, o processo é encerrado e estas faixas de valores servem como rótulo

para seus grupos.

Depois de 319 iterações, utilizando-se de um grau de seleção de 0,8290. As faixas de valores criadas ainda possuem interseção, porém cada grupo possui pelo menos uma faixa de valor que não tem interseção, satisfazendo a condição de parada do modelo, como mostra a Tabela 7.

Tabela 7 – Rótulos da Iteração #319.

	Grupo 1	**Grupo 2**	**Grupo 3**
At.1	5,8 ∼ 7,8	8,5 ∼ 8,9	3,1 ∼ 4,9
At.2	7,3 ∼ 8,5	**4,0 ∼ 5,1**	**3,6 ∼ 5,7**

Por fim, os rótulos são exibidos na Tabela 8 que mostra as faixas de valores correspondentes a seu 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.

Os rótulos exibidos na Tabela 8 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 8 – Rótulos Finais.

	Grupo 1	**Grupo 2**	**Grupo 3**
At.1	5,8 ∼ 7,8	8,5 ∼ 8,9	3,1 ∼ 4,9
At.2	7,3 ∼ 8,5	-	-

.

3 Resultados

Este capítulo trata dos detalhes da implementação do modelo de rotulação, como as ferramentas e os parâmetros utilizados. Nesse capítulo também será descrito a aplicação do modelo em três bases de dados e uma análise comparativa entre o modelo proposto nesse trabalho e um outro modelo de rotulação.

3.1 Detalhes da Implementação

O modelo proposto foi implementado na ferramenta MATLAB. Os algoritmos fornecidos pela ferramenta apresentam uma boa performance e um ambiente de desenvolvimento fácil para a prototipação de novos modelos.

Para o procedimento de criação dos *clusters* foi utilizado o algoritmo *Fuzzy C-means*. Este algoritmo tem como saída a matriz U que representa os elementos relacionados com seus respectivos graus de pertinência em cada grupo, formando assim grupos que são representados por conjuntos *fuzzy*. Foram utilizados os parâmetros *default* da ferramenta. Isto corresponde aos valores 2,0 para o expoente m, o valor 100 para o número máximo de iterações e 10^{-5} para a quantidade mínima de melhoria. Para os parâmetros do modelo GS e IGS, foram utilizados 0,5 e 10^{-4} receptivamente.

O modelo de rotulação foi aplicado em três bases de dados do repositório UCI *Machine Learning*[1]. Elas foram escolhidas por serem bases de dados amplamente conhecidas, por apresentarem dados compatíveis com o modelo e por facilitar a comparação com o outro modelo de rotulação. Os dados possuem valores reais, suprindo a restrição para a aplicação do modelo.

Para saber o nível de precisão dos rótulos verificou-se os valores dos elementos que estavam presente nos intervalos das faixas de valores construídas pelo modelo. Uma vez feito isto, o elemento era atribuído ao *cluster* correspondente às faixas obedecidas. Esta quantidade de elementos que se encaixam nos rótulos podem ser vistas na coluna Elementos das tabelas de resultado.

3.2 Base de Dados Iris

O modelo foi aplicado na base de dados "Iris", disponível no repositório 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: o comprimento

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

da sépala (CS), a largura da sépala (LS), o comprimento da pétala (CP), a largura da pétala (LP). Esta base de dados possui dados coletados de 3 classes de plantas, *Setosa*, *Versicolor* e *Virginica*.

Segundo o trabalho Fisher (1987), as classes podem ser divididas em:

1) 50 elementos da classe Iris Setosa;

2) 50 elementos da classe Iris Virginica;

3) 50 elementos da classe Iris Versicolor.

A tabela 9 mostra o resultado da execução do modelo na base de dados Iris. Nesta tabela são mostrados os *clusters* associados aos seus respectivos rótulos. A coluna Elementos mostra a quantidade de elementos que obedecem ao rótulo.

Tabela 9 – Resultados da base de dados Iris.

Cluster	**Rótulos**		**Elementos**
	Atributos	**Intervalos (cm)**	
1	CP	5,1 ∼ 6,9	42
2	CP LP	1,0 ∼ 1,9 0,1 ∼ 0,6	50
3	CP	3,5 ∼ 5,0	55

Ainda também pode-se ver na tabela 9 os rótulos gerados pelo modelo, sendo compostos de faixas de valores associadas a alguns atributos de cada *cluster*. Percebe-se também que o atributo comprimento da pétala (CP) está presente em todos os *clusters*, isso mostra que os três *clusters* 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 a principal característica para a identificação do novo elemento. Os demais atributos e suas faixas de valores representam características secundárias, mas também caracterizam os grupos de forma única.

Ao comparar o total de elementos e a soma dos mesmos, percebe-se que 3 deles não foram rotulados em nenhum dos *clusters*. Os que não obedeceram os rótulos gerados pelo modelo, podem ser vistos na tabela 10.

Tabela 10 – Elementos não rotulados da base de dados Iris.

CS	**LS**	**CP**	**LP**
5,0	2,3	3,3	1,0
5,1	2,5	3,0	1,1
4,9	2,4	3,3	1,0

Os 3 elementos da tabela 10 possuem valores no atributo CP que não existe em nenhuma das faixas de valores. Com isto o modelo conseguiu rotular 98% dos elementos

baseando-se somente em poucas características de cada grupo.

3.3 Base de Dados *Seed*

Esta base de dados se refere a tipos de sementes de trigo e também pode ser encontrada no repositório UCI *Machine Learning* como *Seed Data Set*. Ela contém 210 elementos. Cada elemento possui 7 atributos como: área (A), perímetro (P), densidade (C), comprimento da semente (LK), largura da semente (WK), coeficiente de assimetria (AC), comprimento do sulco da semente (LKG). Cada uma das amostras de sementes são classificadas em 3 diferentes tipos, que são:

1) 70 elementos do tipo Kama;

2) 70 elementos do tipo Rosa;

3) 70 elementos do tipo Canadian.

Tabela 11 – Resultados da base de dados Seed.

	Rótulos		
Cluster	**Atributos**	**Intervalos**	**Elementos**
1	A	10,5 ∼13,3	84
2	A	13,5 ∼16,1	56
3	A P	17,2 ∼21,1 15,6 ∼17,2	54

A tabela 11 mostra que os tipos de sementes são fortemente diferenciadas pela área, já que esse atributo está presente nos rótulos. Somente o grupo 3 apresenta uma faixa de valor no atributo perímetro. i Isto mostra que grande parte dos elementos do grupo 3 tem seus valores de perímetro pertencente ao intervalo da faixa de valor.

Durante os testes foi verificado que 194 elementos (92,38%) tiveram os valores de seus atributos pertencentes ao intervalo das faixas de valores fornecidas pelo modelo, enquanto 16 elementos tiveram alguma discordância com os rótulos.

3.4 Base de Dados *Glass*

O modelo também foi aplicado a base de dados *Glass Identification* que pode ser encontrada no repositório de dados UCI *Machine Learning*. O estudo desta base de dados tem aplicação na área forense onde a identificação do tipo de vidro pode ajudar a solucionar crimes.

A base de dados contém 214 amostras de vidro. Cada amostra e formada pelos atributos: índice de refração (IR), e a porcentagens de óxido dos elementos Na, Mg, Al, Si, K, Ca, Ba e Fe. Os valores dos atributos são contínuos e formam dois grandes grupos

que são: 163 elementos de amostra de vidros de janela e 51 de outros tipos de objetos de vidro. No primeiro grupo temos amostras de janelas de veículos, *float* e *no-float*, e janelas de construção, do tipo *float* e *no-float*. No segundo grupo temos amostras de utensílios de cozinha, recipientes e faróis.

Tabela 12 – Resultados da base de dados Glass.

	Rótulos		
Cluster	**Atributos**	**Intervalos**	**Elementos**
1	Mg	$0 \sim 1{,}8$	53
2	Mg	$2{,}1 \sim 4{,}5$	160

A tabela 12 mostra que a principal característica que diferencia os grupos é a presença de Magnésio (Mg) nas faixas de valores mostradas no rótulo. Pelos rótulos mostrados percebe-se que o grupo 1 consiste nas amostras de utensílios de cozinha, recipientes e faróis e o grupo 2 consiste nas amostras de vidros de janela. Segundo Navarro (2003) a quantidade de Mg na fabricação do vidro é relacionada a resistência mecânica que o vidro possui. Com isto a quantidade de Mg nas amostras de Janela são maiores devido à necessidade dos vidros de janelas serem mais resistentes a impactos, já que sua geometria plana e fina facilita sua quebra. Durante os testes 213 elementos (99,53%) obedeceram aos rótulos e 1 elemento não obedeceu ao rótulo.

3.5 Análise Comparativa

Esta seção descreve uma análise relacionada ao modelo proposto neste trabalho. Ela é realizada por meio da comparação deste trabalho com o proposto em Lopes, Machado e Rabelo (2014). A mesma tem sua importância por ambos os modelos abordarem o problema da rotulação. Com isto, ela tem como objetivos: coletar dados da aplicação dos modelos de rotulação nas bases de dados; calcular métricas utilizando tais dados e realizar discussões sobre os resultados obtidos.

O trabalho Lopes, Machado e Rabelo (2014) é descrito na seção Trabalhos Relacionados. A seção descreve o modelo de rotulação que utiliza rede neural para a formação dos rótulos e o funcionamento dos seus parâmetros. No trabalho de apresentação do modelo são exibidos resultados realizados com a utilização do algoritmo não supervisionado K-means, diferentemente do modelo defendido nesse trabalho que utiliza o algoritmo *Fuzzy C-Means*. Apesar dessa diferença ambos os algoritmos possuem um comportamento muito similar.

Para uma melhor compreensão, o modelo proposto no trabalho Lopes, Machado e Rabelo (2014) foi chamado de Rotulação Neural e o modelo descrito nesse trabalho foi chamado de Rotulação *Fuzzy*.

Os dados sobre o modelo Rotulação Neural foram coletados a partir do trabalho Lopes, Machado e Rabelo (2014). Os dados do modelo Rotulação *Fuzzy* foram gerados a partir da aplicação nas mesmas bases de dados do modelo anterior. Os dados coletados foram: o número de elementos contidos em cada grupo fornecido ao modelo; os rótulos gerados pelos modelo, sendo compostos pelos atributos e suas respectivas faixas de valores; a quantidade de erros cometidos, ou seja, a quantidade de elementos que não possuem valores existente em uma determinada faixa de valor; uma porcentagem de acerto, sendo formulada a partir da quantidade de elementos que possui valores existentes em uma faixa de valor e do número de elementos em cada cluster; os parâmetros do modelo utilizados em cada base de dados.

Para realizar a comparação foram utilizadas duas métricas. A primeira pode ser descrita como a média da menor porcentagem de acerto em cada grupo, sendo exposta no trabalho de Lopes, Machado e Rabelo (2014). Esta métrica é calculada a partir da porcentagem de acertos, que leva em consideração a quantidade de elementos em cada grupo. Portanto, podemos dizer que essa métrica sofre uma forte influência do algoritmo de agrupamento. A segunda métrica consiste na soma total de erros, essa métrica representar o total de valores de atributos que não existem no intervalo das faixas de valores.

3.5.1 Base de dados Iris

A seguir as tabelas 13 e 14 mostram os parâmetros utilizados pelos modelos para a produção de rótulos para a base de dados Iris.

Tabela 13 – Parâmetros do Modelo Rotulação Neural aplicados a base de dados Iris.

Parâmetro	**Valor**
Número de iterações por atributo	10
Número de faixas de valores	3
Tipo de discretização	EFD
Variação	10%

Tabela 14 – Parâmetros do Modelo Rotulação *Fuzzy* aplicados a base de dados Iris.

Parâmetro	**Valor**
Grau de seleção inicial	0,5
Incremento do grau de seleção	0,0001

O modelo Rotulação Neural utilizou os mesmos parâmetros descritos no trabalho Lopes, Machado e Rabelo (2014), assim como o modelo Rotulação *Fuzzy* utilizou os mesmos parâmetros anteriormente discutidos neste trabalho. Pode ser visto aqui que o modelo

Rotulação *Fuzzy* utiliza uma quantidade menor de parâmetros para a elaboração dos rótulos.

As tabelas 15 e 16 exibem os rótulos produzidos pelos dois modelos para a base de dados Iris.

Tabela 15 – Rótulos da base de dados Iris produzidas pelo Modelo Rotulação Neural.

Cluster	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**# Elementos**	**Acertos(%)**	**# Erros**
1	LP	0,1 ∼ 1	50	100	0
	CP	1 ∼ 3,7		100	0
2	CP	5,1 ∼ 6,9	38	92,10	3
	LP	1,7 ∼ 2,5		94,73	2
3	CP	3,7 ∼ 5,1	62	90,32	6

Tabela 16 – Rótulos da base de dados Iris produzidas pelo Modelo Rotulação *Fuzzy*.

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

Como pode ser visto, cada grupo é representado por várias faixas de valores associadas a seus respectivos atributos, formando assim os rótulos. Pode-se perceber que algumas faixas de valores produzidas pelos modelos são semelhantes e que o modelo Rotulação *Fuzzy* representou os grupos utilizando um número menor de faixas de valores. Nessa base de dados os dois modelos conseguiram formar rótulos que separam cada um dos grupos utilizando o atributo CP, isto facilita a interpretação dos rótulos, já que expõe essa característica determinante na classificação dos elementos.

A tabela 17 mostra o cálculo das métricas a partir dos dados da coluna Análise.

Tabela 17 – Resultado das métricas para a base de dados Iris.

Métricas	**Rotulação Neural**	**Rotulação *Fuzzy***
Média da menor porcentagem de acerto em cada grupo (%)	94,13	92,88
Soma total de erros	11	11

Com os valores das métricas podemos ver que ambos os modelos atingiram números praticamente iguais nas duas métricas. Na primeira métrica podemos ver uma pequena diferença, com o modelo Rotulação *Fuzzy* tendo uma porcentagem menor de acertos, justificada unicamente pelo seu grupo 3, já que os outros valores da porcentagem de acertos foram maiores que os do modelo Rotulação Neural.

3.5.2 Base de dados Glass

As tabelas 18 e 19 a lista de parâmetros utilizados pelos modelos para a base de dados Glass.

Tabela 18 – Parâmetros do Modelo Rotulação Neural aplicados a base de dados Glass.

Parâmetro	**Valor**
Número de iterações por atributo	10
Número de faixas de valores	4
Tipo de discretização	EWD
Variação	15%

Tabela 19 – Parâmetros do Modelo Rotulação *Fuzzy* aplicados a base de dados Glass.

Parâmetro	**Valor**
Grau de seleção inicial	0,5
Incremento do grau de seleção	0,0001

Os parâmetros utilizados pelo modelo Rotulação Neural foram diferentes dos utilizados na base de dados anterior, aumentando o número de faixas de valores criadas, alterando o processo de cálculo das mesmas e aumentando a variação. Enquanto os parâmetros do modelo Rotulação *Fuzzy* utilizou os mesmos valores.

Tabela 20 – Rótulos da base de dados Glass produzidas pelo Modelo Rotulação Neural.

Cluster	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**# Elementos**	**Acertos(%)**	**# Erros**
1	Ba	0 ∼ 0,7875	74	100	0
	K	0 ∼ 1,5525		100	0
	Si	72,61 ∼ 74,01		97,29	2
	Na	12,3925 ∼ 14,055		95,94	3
2	Fe	Fe 0 ∼ 0,1275	5	100	0
	Ca	Ca 5,43 ∼ 8,12		100	0
3	K	0 ∼ 1,5525	19	100	0
	Ba	0 ∼ 0,7875		94,73	1
4	K	0 ∼ 1,5525	32	100	0
	Ba	0 ∼ 0,7875		96,87	1
	Ca	8,12 ∼ 10,81		96,87	1
5	Ba	0∼0,7875	56	100	0
	K	0∼1,5525		100	0
	Na	12,3925∼14,055		96,42	2
	Al	1,0925∼1,895		92,85	4
	Mg	3,3675 ∼ 4,49		89,28	6
6	Fe	0 ∼ 0,1275	28	100	0
	K	0 ∼ 1,5525		96,42	1

Tabela 21 – Rótulos da base de dados Glass produzidas pelo Modelo Rotulação *Fuzzy*.

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

Pode ser visto que a quantidade de grupos utilizados pelos modelos foi diferente. Isso se deu devido a uma limitação no modelo Rotulação *Fuzzy*. Esta limitação consiste em um caso específico de posicionamento dos grupos, onde um par de centroides tem valores muito próximos em um atributo, fazendo com que o modelo não consiga encontrar uma faixa de valor capaz de distinguir os grupos e impedindo a formação dos rótulos.

Tabela 22 – Resultado das métricas para a base de dados Glass.

Métricas	**Rotulação Neural**	**Rotulação *Fuzzy***
Média da menor porcentagem de acerto em cada grupo (%)	95,54	98,43
Soma total de erros	21	3

A primeira métrica, que se refere à média da menor porcentagem de acerto em cada grupo, obteve valores muito próximos nos dois modelos. Já a segunda métrica, que consiste na soma total de erros, teve um valor menor para a Rotulação *Fuzzy*, provavelmente devido a redução no número de grupos.

3.5.3 Base de dados Seed

A seguir a tabela 23 exibe os parâmetros utilizados pelos modelos para a produção de rótulos para a base de dados Seed.

Tabela 23 – Parâmetros do Modelo Rotulação Neural aplicados a base de dados Seed.

Parâmetro	**Valor**
Número de iterações por atributo	10
Número de faixas de valores	3
Tipo de discretização	EFD
Variação	5%

Tabela 24 – Parâmetros do Modelo Rotulação *Fuzzy* aplicados a base de dados Iris.

Parâmetro	**Valor**
Grau de seleção inicial	0,5
Incremento do grau de seleção	0,0001

Nesta base de dados os parâmetros correspondem aos mesmos defendidos no trabalho proposto por Lopes, Machado e Rabelo (2014). Os parâmetros do modelo Rotulação *Fuzzy* se mantiveram os das bases de dados anteriores.

Tabela 25 – Rótulos da base de dados Seed produzidas pelo Modelo Rotulação Neural.

Cluster	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**# Elementos**	**Acertos(%)**	**# Erros**
1	Área	12,78 ∼ 16,14	67	88,05	8
	Perímetro	13,73 ∼ 15,18		86,56	9
2	Área	10,59 ∼ 12,78	82	85,36	12
	Perímetro	12,41 ∼ 13,73		87,80	10
3	Perímetro	15,18 ∼ 17,25	61	100	0
	Largura do núcleo	3,465 ∼ 4,033		95,08	3
	Comprimento do núcleo	5,826 ∼ 6,675		98,36	1
	Área	16,14 ∼ 21,18		100	0

Tabela 26 – Rótulos da base de dados Seed produzidas pelo Modelo Rotulação *Fuzzy*.

Cluster	**Rótulos**		**Análise**		
	Atributos	**Intervalos**	**# Elementos**	**Acertos(%)**	**# Erros**
1	Área	13,50 ∼ 16,14	72	75,00	18
2	Área	10,59 ∼ 13,37	77	97,40	2
3	Área	17,26 ∼ 21,18	61	88.52	7
	Perímetro	15,66 ∼ 17,25		90.16	6

Os rótulos gerados mostram uma grande semelhança entre as faixas de valores referente ao atributo Área. Aqui também podemos ver que o modelo Rotulação *Fuzzy* gerou um número menor de faixas de valores para a base de dados Seed.

Tabela 27 – Resultado das métricas para a base de dados Seed.

Métricas	**Rotulação Neural**	**Rotulação *Fuzzy***
Média da menor porcentagem de acerto em cada grupo (%)	89,0	86,97
Soma total de erros	43	33

A tabela 27 mostra que o modelo Rotulação *Fuzzy* obteve um número inferior na primeira métrica, porém cometeu menos erros na rotulagem da base de dados.

Podemos ver em todas as tabelas de parâmetros que o modelo Rotulação Neural utiliza 4 parâmetros. Alguns desses parâmetros exigem do especialista um conhecimento maior da base de dados como por exemplo, a forma como as faixas de valores são previamente determinadas e a quantidade de faixas de valores definidas. Já o modelo Rotulação *Fuzzy* possui somente dois parâmetros, facilitando sua utilização.

Os dois modelos possuem algumas limitações. O modelo Rotulação Neural pode ser aplicado em base de dados contínuas e discretas, porém necessita que o especialista tenha um grande conhecimento da base de dados e dos parâmetros do modelo. O modelo Rotulação *Fuzzy* só pode ser aplicado em bases de dados contínuas e possui uma limitação em relação a determinadas formações dos *clusters*. Em alguns casos o algoritmo de agrupamento produz centroides que possuem valores de atributo muito próximos, isso impede que o modelo Rotulação *Fuzzy* produza faixas de valores adequadas para cada grupo.

A análise mostrou que o modelo Rotulação Neural tem um bom desempenho quando aplicado a métrica descrita no trabalho Lopes, Machado e Rabelo (2014), porém essa métrica leva em consideração a quantidade de elementos em cada grupo, fazendo com que os erros cometidos em grandes grupos tenham um menor impacto, e por isso sendo muito dependente do algoritmo de agrupamento. O modelo Rotulação *Fuzzy* atingiu um bom desempenho na segunda métrica, que consiste na soma total de erros. Esta métrica não tem relação com a quantidade de elementos em cada grupo, fazendo com que cada erro de rotulação seja igualmente importante.

Conclusões

Conclusões e Trabalhos Futuros

Neste trabalho foi proposto um modelo capaz de elaborar definições, também chamadas de rótulos, capazes de representar os dados contidos nos *clusters*. A definição se dá pela elaboração de faixas de valores dos atributos para cada *cluster*. As faixas de valores são associadas a atributos capazes de distinguir cada *cluster*. Os rótulos gerados contribuem para o entendimento dos *clusters*, estes são geralmente formados por algoritmos de aprendizagem não-supervisionada.

O modelo apresentado mostrou-se promissor conseguindo representar em média 96,61% dos elementos das bases de dados, utilizando os rótulos produzidos. Além disso, o modelo conseguiu atingir essa marca elaborando poucas faixas de valores, e encontrando a principal característica que separa os grupos. Com isto, foi observado que o modelo de Rotulação *Fuzzy* tem capacidade de formular rótulos baseando-se nas diferenças de cada grupo.

Na análise comparativa, o modelo proposto por Lopes, Machado e Rabelo (2014) obteve bons resultados utilizando como métrica a média da menor porcentagem de acerto em cada grupo, e o modelo proposto nesse trabalho atingiu bons resultados na contagem total de erros. Devido a essas métricas serem calculadas a partir das faixas de valores individualmente, e não do conjunto de faixas de valores que compõem o rótulo, as métricas não contabilizam a quantidade de elementos que se encaixam nos rótulos integralmente. Devido a isto podemos dizer que essas métricas talvez não sejam as mais adequadas para calcular a representatividade dos rótulos.

Com a análise comparativa também foi possível perceber que o modelo de Rotulação *Fuzzy* produziu uma quantidade inferior de faixas de valores, isso facilita a interpretação dos rótulos. O modelo proposto utiliza-se de uma quantidade inferior de parâmetros para serem regulados, facilitando a aplicação do modelo. Foi possível analisar que em alguns casos o modelo não foi capaz de encontrar diferenças significativas entre os grupos. Contudo, o modelo de Rotulação *Fuzzy* demostrou ser uma boa opção para a geração de rótulos.

Como trabalhos futuros se espera aplicar o modelo em bases de dados que possuem um grande volume de elementos, e com isso analisar o seu custo computacional. Também serão estudadas formas de eliminar as limitações do modelo, como a sua aplicação em outras base de dados, a formação dos rótulos independente dos centróides fornecidos e a adaptação do modelo para outros algoritmos de agrupamento. Uma outra abordagem que utiliza a análise de componentes principais será analisada, uma vez que ela poderá reduzir

a dimensionalidade da base de dados e fornecer os atributos relevantes para a formação das faixas de valores.

Referências

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

AZIZ, D. et al. Initialization of adaptive neuro-fuzzy inference system using fuzzy clustering in predicting primary triage category. In: IEEE. *International Conference on Intelligent and Advanced Systems (ICIAS), 2012 4th*. [S.l.], 2012. v. 1, p. 170–174. Citado na página 4.

BARBER, D. *Bayesian Reasoning and Machine Learning*. New York, NY, USA: Cambridge University Press, 2012. ISBN 0521518148, 9780521518147. Citado na página 3.

BEZDEK, J. *Fuzzy mathematics in pattern classification*. [S.l.]: Cornell University, 1973. Citado na página 7.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981. ISBN 0306406713. Citado na página 8.

CHEN, H.-L.; CHUANG, K.-T.; CHEN, M.-S. On data labeling for clustering categorical data. *IEEE Transactions on Knowledge and Data Engineering*, IEEE Computer Society, Los Alamitos, CA, USA, v. 20, n. 11, p. 1458–1472, 2008. ISSN 1041-4347. Citado na página 8.

CHUANG, S.-L.; CHIEN, L.-F. A practical web-based approach to generating topic hierarchy for text segments. In: *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2004. (CIKM '04), p. 127–136. ISBN 1-58113-874-1. Citado na página 8.

CINTRA, M. E. et al. An approach for the extraction of classification rules from fuzzy formal contexts. *Computer Science and Mathematics Institute Technical Reports*, p. 1–28, 2011. Citado na página 9.

DRINEAS, P. et al. Clustering large graphs via the singular value decomposition. *Machine Learning*, v. 56, p. 9–33, 2004. Citado na página 5.

DUDA, R. O.; HART, P. E. *Pattern Classification and Scene Analysis*. New Yotk, USA: John Willey & Sons, 1973. Citado na página 6.

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, v. 3, n. 3, p. 32–57, 1973. Citado na página 7.

ELTOFT, T.; FIGUEIREDO, R. de. A self-organizing neural network for cluster detection and labeling. In: IEEE. *The IEEE International Joint Conference on Neural Networks Proceedings*. [S.l.], 1998. v. 1, p. 408–412. Citado na página 8.

FIGUEIREDO S. BOTELHO, P. D. M.; HAFFELE, C. Self-organizing mapping of robotic environments based on neural networks. *Brazilian Symposium on Artificial Neural Networks*, p. 136–141, 2012. Citado na página 4.

FISHER, D. Improving inference through conceptual clustering. In: *Proceedings of the Sixth National Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 1987. (AAAI'87, v. 2), p. 461–465. ISBN 0-934613-42-7. Citado 2 vezes nas páginas 4 e 18.

GLOVER, E. et al. Inferring hierarchical descriptions. In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management*. New York, USA: ACM, 2002. p. 507–514. ISBN 1-58113-492-4. Citado na página 8.

KANUNGO, T. et al. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 7, p. 881–892, Jul 2002. ISSN 0162-8828. Citado na página 4.

KRISHNAPURAM, R.; KELLER, J. The possibilistic c-means algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems*, v. 4, n. 3, p. 385–393, Ago 1996. ISSN 1063-6706. Citado na página 5.

LIU, X.; FENG, X.; PEDRYCZ, W. Extraction of fuzzy rules from fuzzy decision trees: An axiomatic fuzzy sets (afs) approach. *Data Knowl. Eng.*, Elsevier Science Publishers B. V., v. 84, p. 1–25, 2013. ISSN 0169-023X. Disponível em: <http://dx.doi.org/10.1016/j.datak.2012.12.001>. Citado na página 9.

LOPES, L.; MACHADO, V.; RABELO, R. Automatic cluster labeling through artificial neural networks. In: *IEEE International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2014. p. 762–769. Citado 7 vezes nas páginas 2, 9, 20, 21, 25, 27 e 29.

MCKAY, M. D.; BECKMAN, R. J.; CONOVER, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, Taylor & Francis, Ltd. on behalf of American Statistical Association and American Society for Quality, v. 21, n. 2, p. pp. 239–245, 1979. ISSN 00401706. Citado na página 5.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072. Citado na página 3.

NAVARRO, J. *El vidrio*. Consejo Superior de Investigaciones Científicas, 2003. (Textos Universitários). ISBN 9788400081584. Disponível em: <https://books.google.com.br/books?id=4GsNCPQRaTwC>. Citado na página 20.

OLIVEIRA, J. V. d.; PEDRYCZ, W. *Advances in Fuzzy Clustering and Its Applications*. New York, USA: John Wiley & Sons, Inc., 2007. ISBN 0470027606. Citado 2 vezes nas páginas 3 e 4.

PERES, S. M. et al. Tutorial sobre fuzzy-c-means e fuzzy learning vector quantization: Abordagens híbridas para tarefas de agrupamento e classificação. *Revista de Informática Teórica e Aplicada*, v. 19, n. 1, p. 120–163, 2012. Citado na página 4.

POPESCUL, A.; UNGAR, L. H. Automatic labeling of document clusters. Unpublished MS, U. Pennsylvania. 2000. Disponível em: <http://www.cis.upenn.edu/~popescul/Publications/popescul00labeling.pdf>. Citado na página 8.

QUINLAN, J. Induction of decision trees. *Machine Learning*, Kluwer Academic Publishers, v. 1, n. 1, p. 81–106, 1986. ISSN 0885-6125. Citado na página 8.

RAMATHILAGA, S.; LEU, J.-Y.; HUANG, Y.-M. Adapted mean variable distance to fuzzy-cmeans for effective image clustering. In: IEEE. *First International Conference on Robot, Vision and Signal Processing (RVSP), 2011*. [S.l.], 2011. p. 48–51. Citado na página 4.

SETNES, M. Supervised fuzzy clustering for rule extraction. In: *Fuzzy Systems Conference Proceedings, 1999*. [S.l.: s.n.], 1999. v. 3, p. 1270–1274 vol.3. ISSN 1098-7584. Citado na página 9.

VARGAS, R.; BEDREGAL, B. Uma extensão intervalar do algoritmo fuzzy c-means. In: *Proceedings of CNMAC 2009 (32th Brazilian Conference on Applied and Computational Math)*. [S.l.: s.n.], 2009. Citado na página 9.

VARGAS, R.; BEDREGAL, B.; FILHO, I. O. Algoritmo fuzzy c-means adaptado para aplicações com dados intervalares simbólicos. In: *Proceedings of ERMAC 2009 (IX Encontro Regional de Matemática Aplicada e Computacional)*. [S.l.: s.n.], 2009. Citado na página 9.

XU, R.; WUNSCH II, D. Survey of clustering algorithms. *Transactions on Neural Networks*, IEEE Press, Piscataway, NJ, USA, v. 16, n. 3, p. 645–678, 2005. ISSN 1045-9227. Citado na página 1.

ZADEH, L. A. Fuzzy sets. *Information and control*, Elsevier, v. 8, n. 3, p. 338–353, 1965. Citado na página 6.

APÊNDICE A – Gráfico de dispersão da base de dados fictícia

Figura 2 – Gráfico de dispersão da base de dados fictícia. O gráfico exibe os elementos que obedecem cada um dos rótulos e aqueles que não foram compatíveis com nenhum rótulo.