Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 27, Num. 1 (2020) 48-61

RESEARCH ARTICLE

Group Labeling Methodology Using Distance-based Data Grouping Algorithms

Metodologia de Rotulação de Grupos Utilizando Algoritmos de Agrupamento de Dados Baseados em Distância

Francisco das Chagas Imperes Filho[1*], Vinicius Ponte Machado[2], Rodrigo de Melo Souza Veras[2], Kelson Rômulo Teixeira Aires[2], Aline Montenegro Leal Silva[3]

Abstract: Clustering algorithms are often used to form groups based on the similarity of their members. In this context, understanding a group is just as important as its composition. Identifying, or labeling groups can assist with their interpretation and, consequently, guide decision-making efforts by taking into account the features from each group. Interpreting groups can be beneficial when it is necessary to know what makes an element a part of a given group, what are the main features of a group, and what are the differences and similarities among them. This work describes a method for finding relevant features and generate labels for the elements of each group, uniquely identifying them. This way, our approach solves the problem of finding relevant definitions that can identify groups. The proposed method transforms the standard output of an unsupervised distance-based clustering algorithm into a Pertinence Degree (GP), where each element of the database receives a GP concerning each formed group. The elements with their GPs are used to formulate ranges of values for their attributes. Such ranges can identify the groups uniquely. The labels produced by this approach averaged 94.83% of correct answers for the analyzed databases, allowing a natural interpretation of the generated definitions.
Keywords: Data Labelling — Clustering — Machine Learning

Resumo: Algoritmos de agrupamento de dados são utilizados com frequência para formação de grupos com base na similaridade de seus membros. Neste caso, a compreensão dos grupos é tão importante quanto a sua composição. Definir, ou rotular, grupos pode auxiliar na interpretação e, consequentemente, direcionar esforços para tomada de decisão levando em consideração as peculiaridades de cada grupo. 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 semelhanças entre os grupos, entre outras situações. Devido ao problema relacionado a encontrar definições capazes de identificar facilmente os grupos, este trabalho descreve uma metodologia que elabora rótulos para encontrar características relevantes nos elementos de cada grupo e identificá-los de forma única. A proposta transforma a saída padrão de um algoritmo de agrupamento não supervisionado baseado em distância em Grau de Pertinência (GP), onde cada elemento da base de dados recebe um GP em relação a cada grupo formado. Os elementos com seus respectivos GPs são utilizados para formular faixas de valores para os atributos. Estes, por sua vez, são capazes de identificar grupos de forma única. Os rótulos produzidos pela proposta deste trabalho atingiram média de percentual de acertos de $94,83\%$ nas bases de dados analisadas, permitindo uma fácil interpretação das definições geradas.
Palavras-Chave: Rotulação de Dados — Agrupamento de Dados — Aprendizagem de Máquina

[1] *Curso de Sistemas de Informação - CSHNB, Universidade Federal do Piauí, Brasil*
[2] *Departamento de Computação, Universidade Federal do Piauí, Brasil*
[3] *Centro de Educação Aberta e a Distância, Universidade Federal do Piauí, Brasil*
*__Corresponding author__: fcoimperes@ufpi.edu.br
DOI: http://dx.doi.org/10.22456/2175-2745.91414 • **Received:** 28/03/2019 • **Accepted:** 30/09/2019

1. Introdução

A capacidade de produção de informações geradas através dos mais variados meios tem dificultado o processo de interpretação de dados para muitos especialistas, que têm 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 [1].

Uma possível forma de tratar grandes volumes de dados é através da utilização de algoritmos de Aprendizagem de Máquina (AM). Técnicas que se enquadram na categoria de AM tentam construir modelos a partir de entradas específicas e usam essas entradas para fazer previsões, ao invés de seguir o conjunto fixo de instruções definidas pelo usuário [2].

Basicamente a AM divide-se em quatro grandes sub-categorias, supervisionada, não supervisionada, semi-supervisionada e aprendizagem por reforço. A primeira utiliza o recurso de um atributo especial nomeado atributo classe. Ela busca mapear entradas em saídas definidas de acordo com o valor do atributo classe de cada elemento da base de dados [3]. A segunda busca encontrar características semelhantes nos valores dos atributos que possam agrupar os elementos de acordo com um padrão de similaridade [4].

O método semi-supervisionada é um meio termo entre aprendizado supervisionado e não supervisionado. Nessa técnica, dados rotulados e não rotulados são usados no processo de aprendizado. Esses algoritmos são usados porque, em muitas tarefas, a geração de dados rotulados costuma ser cara [5]. A aprendizagem por reforço trata de situações onde um agente aprende por tentativa e erro ao atuar sobre um ambiente dinâmico [6]. O presente trabalho foca no uso da técnica de AM não supervisionada.

Algoritmos que utilizam métodos de AM não supervisionada aprendem sem qualquer intervenção humana [7]. Eles selecionam um conjunto de dados e agrupa-os 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 [8].

O Agrupamento de Dados (AD) tem sido considerado um dos tópicos mais relevantes dentre aqueles existentes na área de AM e Mineração de Dados (MD) [9]. A MD é definida como o procedimento que localiza dados valiosos a partir de conjuntos de informações brutas, investigando e compactando-os considerando pontos de vista alternativos [10]. O desenvolvimento e aperfeiçoamento de algoritmos que tratam do AD tornaram-se o centro de muitas pesquisas, porém poucos trabalhos visam especificamente o estudo da rotulação, notadamente as definições e compreensões dos grupos.

A rotulação consiste em nomear os grupos de acordo com suas principais características, o que significa apresentar uma clara identificação dos agrupamentos formados. Uma boa definição de cada grupo pode torná-lo mais compreensível para um especialista enquanto estuda ou interpreta dados [11]. Isto posto, o problema que compreende a rotulação de grupos pode ser definido como segue.

Dado um conjunto de grupos $C = \{c_1, ..., c_k \,|\, K \geq 1\}$, de modo que cada grupo contém um conjunto de elementos $c_i = \vec{e}_1, ..., \vec{e}_{n^{(c_i)}} \mid n^{(c_i)} \geq 1$ que pode ser representado por um vetor de atributos definido em $\mathbb{R}^m$ e expresso por $\vec{e}_j^{(c-i)} = (a_1, ..., a_m)$ e ainda que $c_i \cap c_{j'} = \{\emptyset\}$ com $1 \leq i, i' \leq k$ e $i \neq i'$, objetivo consiste em apresentar um conjunto de rótulos $R = \{r_{c_1}, ..., r_{c_k}\}$ no qual cada rótulo específico é dado por um conjunto de pares de valores, atributo e seu respectivo intervalo, $r_{c_i} = \{(a_1, [p_1, q_1]), ..., (a_{m^{c_i}},]P_{m^{(c_i)}}, Q_{m^{(c_i)}}])\}$ capaz de melhor expressar o grupo c_i associado [12].

A fim de esclarecimento, tem-se que: K número de grupos; c_i um grupo qualquer; $n^{(c_i)}$ número de elementos do grupo c_i; $\vec{e}_j^{(c-i)}$ se refere ao j-ésimo elemento pertencente ao grupo c_i; m a dimensão do problema; r_{c_i} rótulo referente ao grupo c_i; $]P_{m^{(c_i)}}, Q_{m^{(c_i)}}]$ representa o intervalo de valores do atributo $a_{m^{c_i}}$ onde $P_{m^{(c_i)}}$ é o limite inferior e $Q_{m^{(c_i)}}$ é o limite superior; e $m^{(c_i)}$ é a quantidade de atributos presente em um rótulo referente ao grupo c_i.

Diante desta problemática, esse trabalho tem como objetivo utilizar um algoritmo de agrupamento de dados baseado em distância e apresentar uma metodologia capaz de analisar grupos de dados e produzir rótulos para ajudar na compreensão e auxiliar especialistas no processo de tomada de decisão.

2. Trabalhos Relacionados

A seguir, são apresentados os trabalhos relacionados que seguem a mesma linha de pesquisa desta proposição.

Rotulação de Grupos Utilizando Conjuntos *Fuzzy*

O trabalho de Machado, Ribeiro e Rabelo (2015)[13] apresenta um modelo de rotulação capaz de identificar características únicas em cada grupo, podendo assim, facilitar a sua compreensão. A proposta foca na teoria de conjuntos *fuzzy* para encontrar atributos 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 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.

Os autores ressaltam em seu trabalho que além da base de dados o modelo necessita da definição dos parâmetros 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 isso, 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 consiste em um valor de incremento do parâmetro GS a cada iteração. Ambos

podem variar entre os valores 0 e 1, inclusive.

Os autores apontam que a proposta mostrou-se promissora conseguindo rotular com êxito os grupos das bases de dados estudadas. Por fim, acrescentam que em uma análise comparativa com o modelo idealizado por [11], obteve bons resultados utilizando como métrica a média da menor porcentagem de acerto em cada grupo, como também na contagem total de erros.

Automatic labelling of clusters of discrete and continuous data with supervised machine learning

O foco deste trabalho é formular rótulos para um conjunto de grupos fornecidos. Os rótulos têm o objetivo de representar os elementos, facilitando a compreensão dos grupos [11].

A metodologia da proposta é descrita como segue. De posse de uma base de dados como entrada, um algoritmo de aprendizagem de máquina não supervisionada é aplicado com o objetivo de formar grupos a partir dos elementos inicialmente fornecidos. Neste primeiro momento, o modelo utiliza o algoritmo *K-Means* para agrupar os dados, porém, como os autores do trabalho afirmam, outros algoritmos de agrupamento de dados podem ser utilizados. Em seguida, para cada grupo formado um segundo algoritmo, desta vez com aprendizagem de máquina 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.

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 grupo não é limitado a ser representado por apenas um valor de um determinado atributo mas sim por um intervalo de valores. Os autores acrescentam que a escolha do método de discretização pode alterar significativamente os resultados gerados.

Para aplicar o procedimento de discretização foram utilizadas duas abordagens não supervisionadas. Discretização por Larguras Iguais e Discretização por Frequências Iguais, mesmas técnicas utilizadas em Araújo et al. (2016) [14]. 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 a relevância de um atributo, dadas todas as relevâncias dos outros atributos.

Logo após a 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.

Automatic Cluster Labeling Based on Phylogram Ana- lysis

Este artigo propõe o uso de métodos de aprendizado de máqui- na não supervisionado e supervisionado para tarefas de agrupamento e rotulagem de dados, respectivamente [14]. Para agrupar os dados foi utilizado o algoritmo *DAta MIning of COde REpository* (DAMICORE) e para rotular o método *Automated Labeling Method* (ALM) [11]. Segundo os autores duas etapas foram adicionadas para melhorar o resultado do processo de armazenamento dos grupos. Antes de serem agrupados, os conjuntos de dados foram submetidos a um pré-processamento, composto por fases de discretização e codificação.

Os atributos contínuos foram discretizados usando dois métodos: Discretização de largura igual (*Equal Widths Discretization* - EWD), cuja faixa de valores assumida pelo atributo é dividida em faixas de largura iguais; e Discretização de Frequência Igual (*Equal Frequence Discretization* - EFD), que divide o intervalo de valores dos atributos para alocar a mesma quantidade de elementos em cada intervalo resultante. Essa tarefa é importante para a etapa de codificação, que permite ao algoritmo de compactação diferenciar valores mais facilmente.

Os autores concluíram que o método ALM foi utilizado para rotular grupos formados pelo algoritmo *DAMICORE*. Acrescentam que os resultados obtidos foram comparados com os apresentados em Lopes et al (2018) [11]. A análise mostrou que os resultados da aplicação ALM obtidos nos grupos formados pelo *DAMICORE* têm mais precisão em comparação com o alcançado pela aplicação no agrupamento *K-Means*. A eficácia da marcação obtida para o agrupamento *DAMICORE* é devida ao número de agrupamentos resultantes. Com um número maior de grupos, há uma maior especificidade das características de seus respectivos elementos, devido ao menor grau de generalização. Ressalta-se também que a técnica de agrupamento é determinante para a qualidade do rótulo atribuído pelo ALM, uma vez que quanto maior a similaridade intra-cluster, maior a precisão dos rótulos encontrados.

A Tabela 1 apresenta um resumo dos trabalhos relacionados que tratam a mesma temática desta proposta, formular uma abordagem capaz de rotular grupos com o propósito de auxiliar especialistas no processo de tomada de decisão.

De acordo com a Tabela 1, apenas as propostas de Araújo et al. (2018) [14] e Lopes et al. (2016) [11], utilizam algoritmos de aprendizagem de máquina supervisionada para descobrir atributos relevantes. Vale ressaltar que ambas usam, quando necessário, discretização de dados como um processo que antecede a rotulação dos grupos.

Por esses motivos é possível observar que o modelo defendido neste trabalho e a pesquisa proposta por Machado, Ribeiro e Rabêlo (2015) [13], levam vantagens em relação aos dois primeiros modelos inserido na Tabela 1, quando

Tabela 1. Trabalhos Relacionados.

Trabalhos Relacionados	Metodologia	Algoritmos(s) / Técnicas(s)	Discretização dos Dados
Machado, Ribeiro e Rabêlo (2015) [13]	Encontrar atributos relevantes nos elementos de cada grupos e modelar faixas de valores que identifique os grupos de forma única.	*Fuzzy C-Means*.	Não
Lopes et al. (2016) [11]	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
Araújo et al. (2018) [14]	Utilização de um algoritmo não supervisionado para agrupar dados e para rotular conjuntos de dados uma técnica supervisionado.	*DAMICORE*, ALM e os métodos de Discretização por Larguras Iguais e por Frequências Iguais.	Sim
Modelo proposto neste trabalho	Transformar saídas baseadas em distância para grau de pertinência, em seguida formular faixas de valores e montar rótulos de cada grupo formado.	Algoritmo de Agrupamento baseado em distância	Não

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. Materiais e Métodos

A metodologia descrita neste trabalho utiliza um algoritmo de aprendizagem de máquina não supervisionada baseado em distância como uma etapa para rotulação de grupos de dados. O algoritmo *K-Means* é usado como exemplo para demonstrar a abordagem proposta. Também apresenta uma sistemática capaz de analisar grupos de dados e produzir rótulos para ajudar na compreensão, e auxiliar especialistas no processo de tomada de decisão.

3.1 Aprendizagem de Máquina

A AM é uma subárea da Inteligência Artificial (IA) e esta, por sua vez, é o processo de indução de uma hipótese a partir de experiências passadas [15]. A AM deve ter a capacidade de se adaptar a novas circunstâncias, detectar e extrapolar padrões [16]. Portanto, a AM surgiu da percepção de criar programas computacionais que aprendem um determinado comportamento ou padrão de forma independente, eliminando ou reduzindo a necessidade de intervenção humana.

Algoritmos de AM podem ser retratados como mecanismos que extraem um padrão de comportamento a partir de exemplos (dados). Tais algoritmos têm sido utilizados em várias áreas do conhecimento, como imagens médicas [17], segurança e detecção de intrusão [18], bioinformática [19], redes de sensores sem fio [20], análise de vulnerabilidade de software [21], análise de sentimentos [22] e identificação de riscos de evasão de estudantes matriculados na modalidade Educação à Distância [23].

A AM apresenta alguma relação com o aprendizado humano, onde seres humanos são capazes de generalizar (aprender) a partir de exemplos ou observações. Logo, a utilização de abordagens que usam ferramentas computacionais introduzindo técnicas de AM são relevantes para o processo de detecção de padrões de forma automática.

3.2 Agrupamento de Dados

O AD é uma técnica de análise que agrupa dados que possuem atributos semelhantes [24]. Algoritmos que utilizam essa técnica dividem os elementos de uma base de dados em grupos, de modo que a similaridade intragrupo (elementos de um mesmo grupo) é maximizada e a similaridade intergrupo (elementos de grupos distintos) é minimizada. Um método de agrupamento é caracterizado principalmente por sua escolha de medida de similitude [25].

O objetivo do agrupamento é encontrar uma atribuição de consenso em cada grupo, combinando padrões de semelhança. 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 agrupar dados.

O conceito de agrupamento é naturalmente utilizado quan-

do existe a necessidade de organizar conjuntos de objetos heterogêneos com base em alguma medida de similaridade. Assim, o agrupamento é centrado em torno de um objetivo intuitivo: dado um conjunto de objetos, pode-se particioná-lo em uma coleção de grupos nos quais os objetos do mesmo grupo estão próximos, enquanto os objetos de diferentes grupos estão distantes um do outro [26].

3.3 *K-Means*

O algoritmo de agrupamento de dados não supervisionado *K-Means* é descrito como um processo de particionamento de uma população N-dimensional em k conjuntos de dados distintos [27].

O *K-Means* pode ser descrito 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 1 [28],

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

onde P é a distância do ponto ao centro mais próximo; χ o conjunto de k centros; n é o número de pontos; e d a distância entre um ponto e um conjunto de grupos.

O *K-Means* pode ser implementado utilizando-se de diferentes medidas de distâncias para estabelecer o particionamento N-dimensional de amostras de dados. As mais difundidas são as distâncias Euclidiana e Manhattan. A primeira, representada pela Equação 2, é definida como a soma da raiz quadrada da diferença entre x e y (pontos em um espaço bidimensional) em suas respectivas dimensões,

$$d(x, y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{2}$$

A Distância de Manhattan é caracterizada pela soma das diferenças entre x e y em cada dimensão, como na Equação 3,

$$d(x, y) = |x_1 - x_2| + |y_1 - y_2|. \tag{3}$$

Em ambos os casos, a complexidade computacional do algoritmo *K-Means* é expressa pela Equação 4 [29],

$$O(\text{k*n}), \tag{4}$$

onde k é o número de grupos e n representa o número de pontos. Este trabalho evidencia a utilização do algoritmo *K-Means* implementando a Distância Euclidiana.

Com o *K-Means* é viável processar grandes amostras de dados. Algumas possíveis aplicações que podem utilizar a eficiência desse algoritmo incluem técnicas de mineração de textos [30], segmentação de imagens de impressões digitais [31], mapeamento de áreas de risco de incêndios terrestres [32] e recomendação de produtos baseados nas características de compras de clientes na *web* [33].

A necessidade de um parâmetro k definido de forma explícita, costuma ser um problema tendo em vista que nem sempre se sabe quantos grupos existem a priori. Mesmo existindo outras possibilidades para descrever este método de particionamento de um conjunto de dados, de forma genérica o algoritmo *K-Means* pode ser descrito conforme fluxograma exibido na Figura 1 [34].

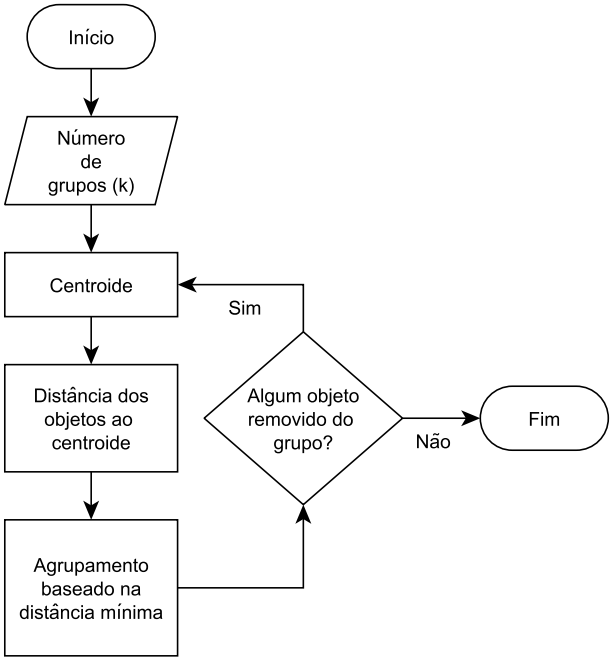

Figura 1. Fluxo de execução do *K-Means*. Fonte: adaptado de Shabari, Shetty e Siddappa (2017) [34].

O *K-Means* tende a convergir para uma configuração estável, na qual nenhum elemento está designado para um grupo cujo centro não lhe seja o mais próximo. Um exemplo da execução do algoritmo é apresentado na Figura 2. A Figura 2(a), representa um conjunto de dados em um espaço bidimensional. Na Figura 2(b), são atribuídos centroides aos grupos de forma aleatória ($k = 3$). Na Figura 2(c), é calculada a distância entre os pontos de dados e os centroides, e cada elemento é designado a um grupo específico. A Figura 2(d), demonstra novos centros de grupos obtidos a partir do cálculo da média das distâncias dos pontos de dados para cada grupo. Por fim, para haver a convergência exibida na Figura 2(e), as etapas demonstradas na Figura 2(c) e 2(d) são repetidas até que os centroides dos grupos não sejam alterados ou o número máximo de iterações seja atingido.

A convergência estável ocasionalmente pode gerar um problema que enfatiza a questão da homogeneidade da separação dos grupos. Isto pode impactar na formação dos grupos

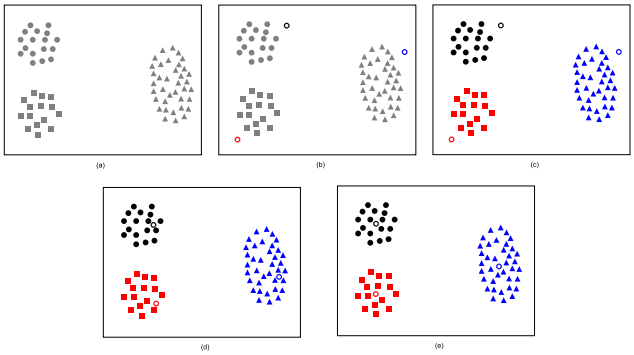

Figura 2. Exemplo de execução do algoritmo *K-Means*.

no caso de uma má inicialização dos centroides, inicialização esta que é feita de forma aleatória no início da execução.

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.

4. Metodologia Proposta

Como demonstra o fluxograma da Figura 3, a metodologia proposta nesta abordagem é descrita em duas etapas. A primeira utiliza um algoritmo de agrupamento para particionar um conjunto de dados e transformar saídas baseadas em distância para grau de pertinência; a segunda formula faixas de valores, define e exibe os rótulos de cada grupo formado.

Ao submeter a base de dados ao agrupador a inicialização do método se dá pela definição dos valores do parâmetro k, Grau de Seleção (GS) e Incremento do Grau de Seleção (IGS). O primeiro refere-se à quantidade de grupos a serem formados pelo algoritmo de agrupamento de dados; o segundo a um valor que serve de base para seleção dos elementos mais significativas na formulação dos rótulos; e o terceiro refere-se a um valor de incremento do parâmetro GS, a cada iteração do processo de definição dos rótulos finais. Sua função é prover a formulação de novas faixas de valores até que seja atingida a condição de parada do modelo. Caso exista 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.

Concluída a primeira etapa, são selecionados os elementos que possuem GP maior que o parâmetro GS. Com isto, em cada grupo são extraídos os valores máximo e o mínimo de cada atributo. Esses valores correspondem às faixas de valores de cada grupo. Caso existam faixas de valores de um mesmo atributo possuindo interseção, essas faixas são descartadas e, nesta situação, o parâmetro IGS é incrementado e o processo continua até que seja atingida a condição de parada definida para a metodologia proposta.

4.1 Métricas de verificação de resultados

As métricas quantidade de parâmetros (QP), total de erros (TE), média da taxa de acertos (MTA) [35] e desvio padrão (DP) foram usadas para verificação dos resultados. A métrica QP define os parâmetros k, GS e IGS com seus respectivos valores a serem utilizados nas base de dados analisadas neste trabalho. Essa métrica é necessária pois bases de dados distintas podem requerer valores distintos para a definição dos valores dos parâmetros.

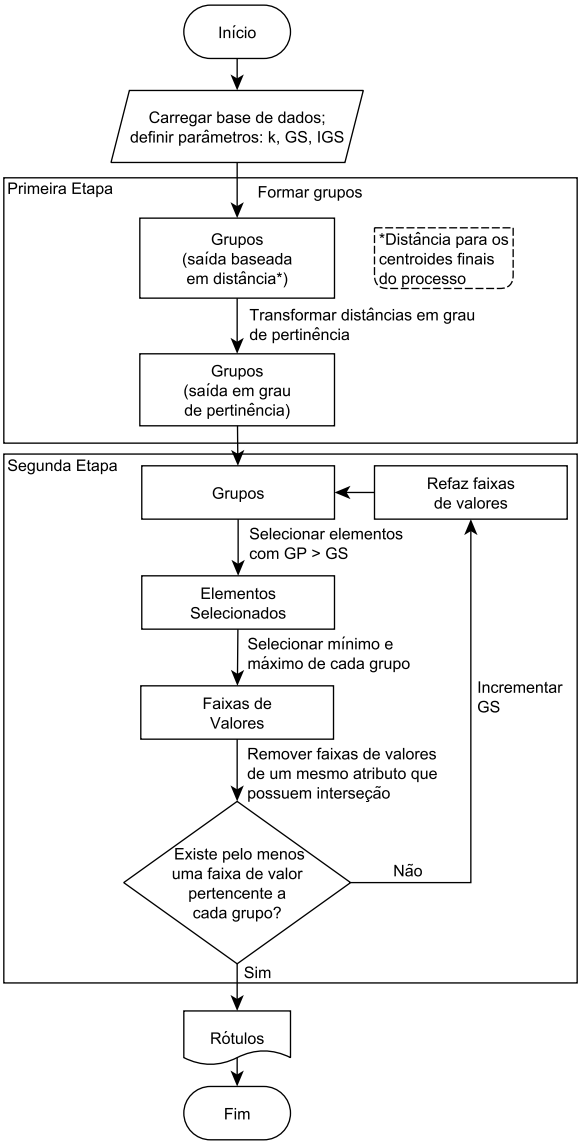

Figura 3. Fluxograma da metodologia proposta.

A métrica TE, definida pela Equação 5, representa o total de valores de atributos não rotulados no intervalo das faixas de valores de cada grupo,

$$TE = n - TA, \tag{5}$$

onde n é o número de elementos em cada grupo; e TA o total

de acertos, tipificando os elementos rotulados corretamente nos intervalos das faixas de valores de cada grupo.

A métrica MTA, descrita pela Equação 6, equivale a média da taxa de acerto em cada grupo. Ela é calculada levando em consideração a quantidade de elementos de cada grupo e os respectivos elementos não rotulados dentro do grupo. Portanto, ela é fortemente influenciada pelo algoritmo de agrupamento utilizado,

$$MTA = \frac{(n - TE).100}{n}, \tag{6}$$

onde n representa o número de elementos em cada grupo; e TE o total de erros, caracterizando o número de elementos não rotulados em cada grupo.

As métricas apresentadas são importantes porque servem de critérios para demonstrar a eficiência da metodologia proposta.

A metodologia de agrupamento funciona independente do número de grupos gerados pelo agrupador. Optamos por utilizar os mesmos números de classes (k) das bases estudadas para poder observar se os rótulos condizem com o que a literatura aponta. Em uma situação em que não há conhecimento da quantidade de grupos, nossa metodologia continuará sendo capaz de encontrar rótulos que distinguem os grupos, independente de quantos são esses grupos.

4.2 Transformar Saídas Baseadas em Distância para Grau de Pertinência

A primeira etapa do método transforma saídas de um algoritmo de agrupamento baseada em distância para grau de pertinência. Algoritmos baseados em distâncias estabelecem que quanto menor for a distância de um elemento para um centroide do grupo, mais próximo esse elemento vai estar desse grupo. Com relação ao grau de pertinência, quanto mais próximo de 1 for o valor atribuído ao grupo, mais próximo o elemento vai estar deste grupo.

Para encontrar o grau de pertinência a partir da distância são executados alguns passos. Primeiro é necessário calcular o inverso da distância de cada elemento para cada grupo. A partir desse resultado é possível obter o valor inverso total do elemento em relação às suas distâncias para cada grupo.

Neste trabalho o inverso da distância 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 1 a k. Após calcular os inversos das distâncias do elemento para cada grupo, deve-se somar os resultados de I para encontrar o valor total invertido do elemento em relação às suas distâncias, definido pela fórmula $D = \frac{1}{d_{j,i}}$. Sendo o número de elementos representado por j, com j variando de 1 a n, o somatório dos inversos das distâncias I_j é dado pela Equação 7,

$$I_j = \sum_{i=1}^{k} \frac{1}{d_{j,i}}, \tag{7}$$

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

A partir do somatório dos inversos das distâncias I_j de cada elemento da base, 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 e de cada elemento j, e divide-se pela distância total I_j correspondente. O procedimento geral para determinar o grau de pertinência pode ser visualizado na Equação 8,

$$P_{j,i} = \frac{D}{I_{j,i}}, P_{j,i} \leq 1, \tag{8}$$

onde $P_{j,i}$ Grau de Pertinência; D valor da distância total invertida; e $I_{j,i}$ somatório dos inversos das distâncias de cada elemento para cada grupo.

Com os passos descritos na Subseção 4.2 conclui-se a primeira etapa da metodologia de rotulação, transformação da saída de um algoritmo de agrupamento baseado em distância para grau de pertinência.

4.3 Formular Faixas de Valores, Definir e Exibir Rótulos

A segunda etapa da metodologia proposta está subdividida em:

1. formular faixas de valores para cada atributo em cada grupo formado;
2. verificar a existência de interseções entre faixas de valores;
3. definir e exibir rótulos de cada grupo com faixas de valores que não possuem interseção.

A primeira fase seleciona elementos que possuem um GP maior que o parâmetro GS. Com isto, em cada grupo selecionado são extraídos os valores máximo e mínimo de cada atributo. Esses valores correspondem às faixas de valores de cada atributo em cada grupo.

A segunda fase verifica se existem interseções entre as faixas de valores pertencentes a um mesmo atributo. Caso exista interseção entre as faixas de valores, 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. Se nenhum atributo possuir pelo menos uma faixa de valor capaz de representar cada um dos grupos, o GS é incrementado pelo parâmetro IGS e o processo de seleção de elementos é refeito utilizando um novo valor para GS e, desta forma, são geradas novas faixas de valores a serem analisadas.

A terceira fase define o(s) rótulo(s) verificando se existe pelo menos uma faixa de valores de um atributo sem

interseção em cada grupo (condição de parada). Nessa circunstância, o processo é encerrado e estas faixas de valores servem como rótulo para seus grupos.

5. Resultados Experimentais

As bases de dados *Iris Data Set*[1], *Breast Cancer Wisconsin (Diagnostic) Data Set*[2] e *Parkinson's Disease Classification Data Set*[3], disponíveis no repositório digital UCI Machine Learning[4], foram usadas para realizar experimentos e validar o método apresentado neste trabalho.

O modelo idealizado neste trabalho foi executado 10 vezes sobre cada base testada com o objetivo de obter o grau de dispersão (desvio padrão) e expressar o quanto os resultados se desviaram da média de acertos.

5.1 *Iris Data Set*

A base de dados *Iris* refere-se a amostras de plantas e contém 150 elementos, onde cada um deles possui quatro atributos com valores definidos no conjunto dos números reais. Os atributos com seus respectivos domínios podem ser visualizados na Tabela 2. A base de dados possui informações coletados de três classes de plantas da família Iris (*setosa*, *versicolor* e *virginica*) [36].

Tabela 2. Atributos com os respectivos domínios da base de dados *Iris*.

	Atributos	Significado	Domínio
1	CS	Comprimento da Sépala	Reais
2	LS	Largura da Sépala	Reais
3	CP	Comprimento da Pétala	Reais
4	LP	Largura da Pétala	Reais
5	Tipo	Atributo Classe	1: Setosa 2: Versicolor 3: Virginica

Para a formulação dos rótulos para a base de dados *Iris*, foi usada a métrica quantidade de parâmetros com seus valores descritos na Tabela 3.

Tabela 3. Métrica quantidade de parâmetros com seus respectivos valores aplicados à base de dados *Iris*.

Parâmetros	Valores
Número de Grupos (k)	3
Grau de Seleção (GS) inicial	0,5
Incremento do Grau de Seleção (IGS)	0,0001

[1]Iris Data Set. Disponível em: http://archive.ics.uci.edu/ml/datasets/Iris. Acesso em: 16 set. 2019.

[2]Breast Cancer Wisconsin (Diagnostic) Data Set. Disponível em: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. Acesso em: 16 set. 2019.

[3]Parkinson's Disease Classification Data Set. Disponível em: https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification. Acesso em: 16 set. 2019.

[4]UCI Machine Learning Repository. Disponível em: http://archive.ics.uci.edu/ml/index.php. Acesso em: 16 set. 2019.

Como conhecido na literatura, a base de dados *Iris* possui 3 classes de plantas bem definidas. Assim sendo, ao parâmetro k foi atribuído o valor 3. O GS foi 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. Já o IGS, foi definido como 0,0001. As justificativas para formulação desse último valor são: a) tanto as distâncias geradas pelo algoritmo de agrupamento 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.

Como a metodologia utiliza um algoritmo de agrupamento de dados, o atributo classe que representa os tipos de plantas foi ignorado. Vale ressaltar que a cada execução o algoritmo pode compor grupos com tamanhos diferentes, da mesma forma que os elementos podem ser inseridos em grupos distintos dependendo da posição dos centroides iniciais. Esta colocação é válida para todas as bases de dados avaliadas por esta proposta.

Fundamentando-se nas informações anteriores, a primeira iteração do método sobre a base de dados *Iris* pode ser visualizada na Tabela 4. Nesta tabela são exibidos os primeiros rótulos (faixas de valores) tendo como foco os elementos selecionados em cada grupo, partindo do princípio que o grau de pertinência dos elementos selecionados é maior que o valor inicial do parâmetro grau de seleção.

Tabela 4. *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 4, o modelo detectou interseções entre faixas de valores de um mesmo atributo em grupos distintos (células destacadas em negrito). Por essa razão, estas faixas são descartadas e a metodologia parte para formação de outro conjunto de faixas de valores.

A Tabela 5 exibe o resultado dos rótulos depois da execução da metodologia sobre base de dados *Iris* após 568 iterações, com $GS = 0,5568$. Destaca-se que o IGS incrementa o GS em $0,0001$ a cada iteração do modelo proposto. Este procedimento facilita a formação de novos intervalos e, consequentemente, permite a análise de novas faixas de valores.

Tabela 5. *Iris Data Set* - Rótulos após iteração #568, com $GS = 0,5568$.

	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**

Observando a Tabela 5, verifica-se que ainda existem interseções entre as faixas de valores destacadas em negrito, especificamente os atributos CS, LS e LP. Entretanto, ela também realça que não há interseções entre as faixas de valores dos atributos CP (todos os grupos) e LP (Grupo 1). Desta forma, os rótulos visualizados na Tabela 6, após descartadas as interseções entre faixas de valores dos atributos CS, LS e LP, apresentam faixas de valores únicas em relação aos grupos e atributos da base de dados *Iris*. Esse conjunto de faixas de valores compõe uma identificação para os grupos e pode vir a representar a maioria dos elementos neles contidos. É importante frisar que podem existir várias faixas de valores relacionadas a um grupo, porém não deve existir interseção entre faixas de valores de um mesmo atributo.

Tabela 6. *Iris Data Set* - Rótulos finais e faixas de valores únicas.

	Grupo 1	Grupo 2	Grupo 3
CP	1,0 ∼ 1,9	3,5 ∼ 5,0	5,1 ∼ 6,9
LP	0,1 ∼ 0,6	-	-

Em consonância com as informações alcançadas depois da condição de parada do método, a Tabela 7 apresenta os grupos associados aos seus respectivos rótulos, enfatizando a quantidade de elementos que obedecem aos rótulos formados (TA), a porcentagem de acertos (MTA) e a quantidade de valores de atributos que não pertencem aos intervalos de faixas de valores dos rótulos finais (TE).

Tabela 7. *Iris Data Set* - Grupos e elementos associados aos respectivos rótulos.

Grupos	Rótulos		Análise		
	Atributos	Intervalos	TA	MTA (%)	TE
1	CP LP	1,0 ∼1,9 0,1 ∼0,6	50	100	0
2	CP	3,5 ∼5,0	52	85,25	9
3	CP	5,1 ∼6,9	36	92,31	3

Constata-se na Tabela 7 que o atributo Comprimento da Pétala (CP) está presente em todos os grupos. Essa condição demonstra que os três grupos se diferem pelas faixas de valores desse atributo. Um especialista que eventualmente quiser associar um novo elemento a um grupo teria no comprimento da pétala o principal diferencial para a compreensão do novo elemento. Os demais atributos e suas faixas de valores representam características secundárias e 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 (TE). Os intervalos com seus respectivos atributos que não obedeceram aos rótulos gerados, podem ser visualizados na Tabela 8.

Os intervalos dispostos na Tabela 8 possuem valores nos atributos CP e LP que não existem em nenhuma das faixas de valores formuladas pelo método de rotulação proposto (ver Tabela 6). Isso ocorre porque no momento da formulação dos rótulos foram detectadas interseções entre faixas de valores dos atributos CP e LP em grupos distintos, e desta forma foram ignorados pelo modelo de rotulação. Assim, dos 150 elementos o método conseguiu rotular corretamente 138 (92,52%) baseando-se nas características existentes de cada grupo. A média da taxa de acertos, o total de erros nos três grupos e o desvio padrão após 10 execuções do modelo para a base de dados *Iris Data Set*, podem ser observados na Tabela 9.

Tabela 8. *Iris Data Set* - Elementos não rotulados.

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

Tabela 9. *Iris Data Set* - Resultados da média da taxa de acertos, total de erros e desvio padrão.

Métrica	Valores
Média da taxa de acertos (%)	92,52
Total de erros	12
Desvio padrão	0,005

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

5.2 *Breast Cancer Wisconsin (Diagnostic) Data Set*

A base de dados *Breast Cancer* foi compilada nos Hospitais da Universidade de *Wisconsin, Madison, USA*. Ela contém amostras de diagnósticos de Câncer de Mama (CM) totalizando 699 elementos, onde cada elemento possui 10 atributos [37].

Neste trabalho dois atributos foram desconsiderados. O primeiro por não possuir relevância para formulação dos rótulos, pois representa um identificador sequencial para as amostras. O segundo por caracterizar o atributo classe da base de dados. Este tem por objetivo definir o resultado do diagnóstico como benigno ou maligno para incidência do CM em cada amostra. Por utilizarmos na metodologia proposta neste estudo um algoritmo de aprendizagem de máquina não supervisionado, o atributo diagnóstico foi ignorado. Diante destas características e conforme descrito na Tabela 10, para efeito de análise somente 8 atributos foram considerados relevantes.

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

A Tabela 11 apresenta a métrica quantidade de parâmetros utilizados pelo método para a formulação de rótulos para a

Tabela 10. *Breast Cancer* - Atributos com os seus possíveis domínios.

	Atributos	Significado	Domínio (N*)
1	EA	Espessura do Aglomerado	1 - 10
2	UTC	Uniformidade do Tamanho da Célula	1 - 10
3	UFC	Uniformidade da Forma da Célula	1 - 10
4	AM	Aderência Marginal	1 - 10
5	TUCE	Tamanho Único da Célula Epitelial	1 - 10
6	ND	Núcleo Descoberto	1 - 10
7	CS	Cromatina Suave	1 - 10
8	NN	Nucleulus Normal	1 - 10

base de dados *Breast Cancer*.

Tabela 11. Métrica quantidade de parâmetros com seus respectivos valores aplicados à base de dados *Breast Cancer*.

Parâmetros	Valores
Número de Grupos (k)	2
Grau de Seleção (GS) inicial	0,5
Incremento do Grau de Seleção (IGS)	0,0001

Com base no parâmetro k, o algoritmo agrupamento formulou dois grupos. O primeiro com 232 (33,97%) elementos representando amostras de diagnósticos de câncer de mama na condição maligna e o segundo com 451 (66,03%) elementos caracterizando situação benigna, totalizando 683 amostras analisadas pelo modelo proposto.

Conforme descrito na Tabela 12, a primeira iteração da metodologia proposta sobre a base de dados *Breast Cancer* apresentou interseções entre faixas de valores de todos os atributos em todos os grupos (células destacadas em negrito). Por conseguinte, essas faixas são descartadas e o modelo avança incrementando o GS, levando em consideração o parâmetro IGS.

Tabela 12. *Breast Cancer* - Rótulos gerados tendo como base os elementos selecionados em cada grupo: Iteração #1.

	Grupo 1	Grupo 2
EA	**1 ∼10**	**1 ∼10**
UTC	**1 ∼10**	**1 ∼5**
UFC	**1 ∼10**	**1 ∼8**
AM	**1 ∼10**	**1 ∼10**
TUCE	**2 ∼10**	**1 ∼10**
ND	**1 ∼10**	**1 ∼10**
CS	**1 ∼10**	**1 ∼7**
NN	**1 ∼10**	**1 ∼9**

Para formulação dos rótulos finais foram necessárias 2164 iterações, atualizando o valor do GS para $0,7164$. Os resultados podem ser observados na Tabela 13.

Mesmo ainda existindo interseções entre faixas de valores de muitos atributos em relação aos grupos (células destacadas em negrito), há pelo menos um atributo que não possui interseção entre faixas de valores para os grupos formados. Esta situação indica a condição de parada da proposta defendida neste trabalho. A Tabela 14 exibe os rótulos finais para a base de dados *Breast Cancer*.

Tabela 13. *Breast Cancer* - Rótulos após iteração #2164.

	Grupo 1	Grupo 2
EA	**3 ∼10**	**1 ∼7**
UTC	4 ∼10	1 ∼3
UFC	**3 ∼10**	**1 ∼4**
AM	**2 ∼10**	**1 ∼4**
TUCE	**3 ∼10**	**1 ∼5**
ND	**5 ∼10**	**1 ∼5**
CS	**3 ∼10**	**1 ∼7**
NN	**2 ∼10**	**1 ∼6**

Tabela 14. *Breast Cancer* - Rótulos finais e faixas de valores únicas.

	Grupo 1	Grupo 2
UTC	4 ∼10	1 ∼3

As informações visualizadas na Tabela 15, apresentam os grupos associados aos seus respectivos rótulos, destacando a quantidade de elementos que obedecem aos rótulos formados (TA), a porcentagem de acertos (MTA) e a quantidade de erros (TE) encontrados dentre as amostras analisadas.

Tabela 15. *Breast Cancer* - Grupos e elementos associados aos respectivos rótulos.

Grupos	Rótulos		Análise		
	Atributos	Intervalos	TA	MTA (%)	TE
1	UTC	4 ∼ 10	205	88,36	27
2	UTC	1 ∼ 3	443	98,23	8

De acordo com a metodologia proposta, o atributo Uniformidade do Tamanho da Célula (UTC) é o mais significativo para identificação de amostras acometidas com câncer de mama. O valor para o atributo UTC, 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 [37].

O procedimento de observação física da área afetada é chamado de Aspiração por Agulha Fina (AAF). O domínio para o atributo UTC 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 a entender a formação dos grupos e identificar quais atributos com seus respectivos valores são mais relevantes para detectar a doença em amostras de câncer de mama.

Verificando a Tabela 15, observa-se que 35 elementos não foram rotulados corretamente (TE). Os intervalos com as respectivas faixas de valores de atributos que não obedeceram os rótulos gerados, podem ser conferidos na Tabela 16.

A métrica da TE representa o total de valores de atributos que não existem nos intervalos das faixas de valores geradas

Tabela 16. *Breast Cancer* - Elementos não rotulados.

Atributos	Intervalo	TE	Grupos
UTC	$1 \sim 3$	27	1
UTC	$4 \sim 5$	8	2

pelos experimentos realizados. A Tabela 17 exibe os resultados alcançados referentes à média da taxa de acertos, total de erros verificados nos dois grupos e desvio padrão após 10 execuções da metodologia sobre a base *Breast Cancer*.

Tabela 17. *Breast Cancer* - Resultados da média da taxa de acertos, total de erros e desvio padrão.

Métricas	Valores
Média da taxa de acertos (%)	93,30
Total de erros	35
Desvio padrão	0,02

Finalizando a análise sobre a base *Breast Cancer*, do universo de 683 amostras o método proposto rotulou corretamente 648, representando uma média de acertos de $93,30\%$. Os dados comprovam que os rótulos constantes na Tabela 15 podem auxiliar o especialista a entender os grupos e os respectivos valores de atributos para detectar enfermidades em amostras referentes a diagnósticos de pacientes com câncer de mama.

5.3 *Parkinson's Disease Classification Data Set*

O conjunto de dados *Parkinson's Disease* (PD) foi coletado de 188 pacientes com Doença de Parkison (DP) com idades variando de 33 a 87 anos no Departamento de Neurologia da Faculdade de Medicina de Cerrahpaşa, Universidade de Istambul. O grupo de controle foi composto por 64 Indivíduos Saudáveis (IS), com idades variando entre 41 e 82 anos. Para extrair informações clinicamente úteis para avaliação da PD o microfone utilizado para coleta de dados de sinais da fala foi definido em 44,1 KHz e, após exame médico, a fonação sustentada da vogal /a/ foi coletada de cada sujeito com três repetições [38].

A base PD contém amostras de expressões gênicas divididas em duas classes, DP e IS. A distribuição das amostras é denotada da seguinte forma: 188 com a DP [(107 homens e 81 mulheres) x 3 repetições] e 64 saudáveis [(23 homes e 41 mulheres) * 3 repetições)], totalizando 756 amostras, onde cada uma possui 754 atributos com valores no domínio dos inteiros e reais. Devido a quantidade de atributos algumas tabelas foram suprimidas para facilitar a visualização dos resultados.

Os valores da métrica quantidade de parâmetros disposta Tabela 11 também foi utilizada para a formulação de rótulos para a base de dados PD.

Após 230 iterações com $GS = 0,5230$, a Tabela 18 apresenta os rótulos finais e as respectivas faixas de valores únicas para a base PD. De acordo com a modelo proposto o atributo Entropia de Densidade do Período de Recorrência (EDPR) é o mais relevante para identificar pacientes com a DP. Esse atributo é considerado um dos recursos da fala mais popular usado em estudos de DP [38]. Na Tabela 18 constam três incidências do atributo EDPR devido as três coletas de amostras a qual cada paciente foi submetido. A mesma situação acontece nas Tabelas 19 e 20.

Tabela 18. *Parkinson's Disease* - Rótulos finais e faixas de valores únicas.

	Grupo 1	Grupo 2
EDPR	-1,4400 $\sim$ -0,2133	-0,2119 $\sim$ -0,0324
EDPR	-2,9919 $\sim$ -0,4451	-0,4422 $\sim$ -0,0679
EDPR	-6,2062 $\sim$ -0,9272	-0,9211 $\sim$ -0,1399

Os grupos, atributo, total de acertos (TA), média da taxa de acerto (MTA) e total de erros (TE) associados aos seus respectivos rótulos são exibidos na Tabela 19.

Tabela 19. *Parkinson's Disease* - Grupos e elementos associados aos respectivos rótulos.

Grupos	Rótulos		Análise		
	Atributos	Intervalos	TA	**MTA (%)**	TE
1	EDPR	-1,4400 $\sim$ -0,2133 -2,9919 $\sim$ -0,4451 -6,2062 $\sim$ -0,9272	271	97,48	7
2	EDPR	-0,2119 $\sim$-0,0324 -0,4422 $\sim$-0,0679 -0,9211 $\sim$-0,1399	475	99,37	3

A Tabela 20 apresenta os intervalos com os respectivos atributos que não foram rotulados corretamente (TE). Dos 756 elementos que compõem a base PD 746 ($98,68\%$) foram rotulados corretamente pelo modelo proposto.

Tabela 20. *Parkinson's Disease* - Elementos não rotulados.

Atributos	Intervalos	TE	Grupos
EDPR	-0,2134 $\sim$-0,2069 -0,4452 $\sim$-0,4318 0,9273 $\sim$-0,8994	7	1
EDPR	-0,2243 $\sim$0,2120 -0,4679 $\sim$-0,4423 -0,9746 $\sim$-0,9212	3	2

Os resultados referentes a média da taxa de acertos, total de erros nos dois grupos e desvio padrão sobre a base PD podem ser conferidos na Tabela 21. As informações inseridas nesta tabela atestam que a metodologia defendida neste trabalho foi capaz de formular rótulos e identificar amostras de pacientes com doença de *Parkinson*.

6. Considerações Finais

Como observado nas Tabelas 9, 17 e 21, a formulação dos rótulos nas bases de dados *Iris*, *Breast Cancer* e *Parkinson Disease* alcançaram média de taxa de acertos acima 90%, destacando-se a terceira que atingiu $98,68\%$ de acertos. Com 57 erros encontrados, com média de taxa de erros de $3,59\%$, de um total de 1589 amostras analisadas nas três bases de dados é possível ponderar que a proposta permite uma fácil interpretação das definições geradas.

Tabela 21. *Parkinson's Disease* - Resultados da média da taxa de acertos, total de erros e desvio padrão.

Métricas	Valores
Média da taxa de acertos (%)	98,68
Total de erros	10
Desvio padrão	0,314

Os resultados 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 auxiliar o processo de tomada de decisão.

Portanto, este trabalho idealizou uma metodologia que utiliza um algoritmo de aprendizagem de máquina não supervisionada baseado em distância capaz de elaborar rótulos e, desta forma, consegue 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 no processo de tomada de decisão.

Um ponto a ser mencionado é que caso todos os atributos do conjunto de dados possuam interseções entre os grupos para todos os valores de GS, nenhum atributo será suficiente para distinguir os elementos de forma única. Nessa condição nosso modelo não retornará faixas de valores capazes de representar os grupos. Situações como a relatada sugerem que os grupos possuem altíssimo grau de similaridade entre todos os atributos ou possuem alta taxa de interseções entre os grupos.

Com os resultados descritos acima, atesta-se que a metodologia defendida neste trabalho conseguiu atingir os objetivos inerentes à sua proposta. Constata-se que ela tem capacidade de formular rótulos baseando-se nas diferenças de cada grupo, facilitando a análise sob o ponto de vista de um especialista.

Por fim, como melhorias, sugere-se submeter ao modelo proposto 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; submeter a metodologia a outras bases de dados que contenham atributos do tipo não numérico e/ou não possuam atributos classe; utilizar o algoritmo *K-Means* implementando outra medida de distância e comparar os resultados com a Distância Euclidiana utilizada neste trabalho; e utilizar um segundo algoritmo de aprendizagem de máquina não supervisionada baseado em distância para enriquecer a proposta apresentada nesta investigação.

Contribuições dos autores

Francisco Imperes e Vinicius Machado propuseram a metodologia, desenvolveram o modelo proposto, realizaram os experimentos e analisaram os resultados obtidos. Rodrigo Veras e Kelson Silva também analisaram os resultados dos experimentos, revisaram o documento e propuseram melhorias na abordagem geral deste estudo. Aline Silva viabilizou e estruturou os recursos computacionais necessários para realizar os experimentos.

Referências

[1] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: ⟨https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230⟩.

[2] 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.

[3] RIZKIN, B. A.; HARTMAN, R. L. Supervised machine learning for prediction of zirconocene-catalyzed α-olefin polymerization. *Chemical Engineering Science*, v. 210, p. 115224, 2019. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S000925091930716X⟩.

[4] LIMA, I.; PINHEIRO, C.; SANTOS, F. *Inteligência Artificial*. Elsevier Editora Ltda., 2016. Disponível em: ⟨https://books.google.com.br/books?id=qjJeBgAAQBAJ⟩.

[5] SANTOS, L. et al. Medical image segmentation using seeded fuzzy c-means: A semi-supervised clustering algorithm. *Proceedings of the International Joint Conference on Neural Networks*, v. 2018-July, 2018.

[6] MONTEIRO, S. T.; RIBEIRO, C. H. C. Desempenho de algoritmos de aprendizagem por reforço sob condições de ambiguidade sensorial em robótica móvel. *SBA Controle & Automação*, v. 15, n. 3, p. 320–338, Jul 2004.

[7] COPPIN, B. *Inteligência Artificial: tradução e revisão técnica Jorge Duarte Pires Valério*. 1. ed. [S.l.]: Rio de Janeiro: LCT, 2010.

[8] 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.

[9] AGGARWAL, C. C.; REDDY, C. K. *Data Clustering: Algorithms and Applications*. 1. ed. [S.l.]: Chapman and Hall/CRC, 2013.

[10] TAFISH, M. H.; EL-HALEES, A. M. Breast cancer severity degree predication using data mining techniques in the gaza strip. *International Conference on*

Promising Electronic Technologies, ICPET 2018, DOI: 10.1109/ICPET.2018.00029, p. 124–128, 2018.

[11] LOPES, L. A. et al. Automatic Labelling of Clusters of Discrete and Continuous Data with Supervised Machine Learning. *Knowledge-Based Systems*, v. 106, p. 231 – 241, 2016.

[12] LOPES, L. A. *Rotulação Automática de Grupos com Aprendizagem de Máquina Supervisionada*. 73 p. Dissertação (Mestrado) — Universidade Federal do Piauí, Teresina, 2014.

[13] MACHADO, V. P.; RIBEIRO, V. P.; RABÊLO, R. de A. L. Rotulação de grupos utilizando conjuntos fuzzy. *XII Simpósio Brasileiro de Automação Inteligente - SBAI*, n. 12, p. 355–360, 2015.

[14] ARAÚJO, F. N. C. de et al. Automatic cluster labeling based on phylogram analysis. *2018 International Joint Conference on Neural Networks (IJCNN)*, p. 1–8, 2018.

[15] FACELI, K. et al. *Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina*. [S.l.]: Rio de Janeiro: LCT, 2011.

[16] RUSSEL, S. J.; NORVIG, P. *Inteligência Artificial*. 3. ed. [S.l.]: Rio de Janeiro: Elsevier Editora Ltda, 2013.

[17] VIEIRA, F. do A. et al. Paraconsistent Extractor of Mammographic Images Applied in the Process of Diagnosis of Breast Cancer Assisted by Computer. *IEEE Conferences: Innovations in Intelligent Systems and Applications (INISTA), DOI:10.1109/INISTA.2018.8466280*, p. 1 – 6, 2018.

[18] BUCZAK, A. L.; 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] DETONI, D. et al. Learning to identify at-risk students in distance education using interaction counts. *Revista Informática Teórica Aplicada (Online)*, v. 23, n. 2, p. 124–140, 2016.

[24] 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.

[25] 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.

[26] RIBALDO, R.; CARDOSO, P. C. F.; PARDO, T. A. S. Explorando mapas de relacionamento com base em subtópicos para sumarização multidocumento. *Revista Informática Teórica Aplicada (Online)*, v. 23, n. 1, p. 183–211, 2016.

[27] 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⟩.

[28] LINDER, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, n. 4, p. 18–36, 2009.

[29] KUMAR, A.; KUMAR, S. Density based initialization method for k-means clustering algorithm. *I.J. Intelligent Systems and Applications*, v. 9, n. 10, p. 40–48, 2017.

[30] NEZHAD, A. S.; SALAJEGHEH, M.; NIA, E. T. Clustering scientific articles based on the k-means algorithm case study: Iranian research institute for information science and technology (irandoc). *Iranian Journal of Information Processing Management*, v. 34, p. 871–896, 2019.

[31] CHERRAT, E.; ALAOUI, R.; BOUZAHIR, H. Improving of fingerprint segmentation images based on k-means and dbscan clustering. *International Journal of Electrical and Computer Engineering*, v. 9, n. 4, p. 2425–2432, 2019.

[32] PRAETYO, S. Y. J. et al. Mitigation & identification for local aridity, based of vegetation indices combined with spatial statistics & clustering k means. In: . [S.l.: s.n.], 2019. v. 1235, n. 1.

[33] MULYAWAN, B.; CHRISTANTI, M. V.; WENAS, R. Recommendation product based on customer categorization with k-means clustering method. In: . [S.l.: s.n.], 2019. v. 508, n. 1.

[34] SHABARI, S.; SHETTY, S.; SIDDAPPA, M. Implementation and comparison of k-means and fuzzy c-means algorithms for agricultural data. *International Conference on Inventive Communication and Computational Technologies, ICICCT 2017, DOI: 10.1109/ICICCT.2017.7975168*, p. 105–108, 2017.

[35] LOPES, L. A.; MACHADO, V. P.; RABÊLO, R. de A. L. Automatic cluster labeling through artificial neural networks.

IEEE International Joint Conference on Neural Networks (IJCNN), p. 762–769, 2014.

[36] FISHER, D. Improving inference through conceptual clustering. *In: Proceedings of the Sixth National Conference on Artificial Intelligence. AAAI Press.*, 1987.

[37] MANGASARIAN, O. L.; WOLBERG, W. H. Cancer diagnosis via linear programming. *SIAM News*, 1990.

[38] SAKAR, C. O. et al. A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. *Applied Soft Computing*, v. 74, p. 255 – 263, 2019. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S1568494618305799⟩.