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

Rotulação Automática de Grupos Através do Uso de Filtros de Ganho de Informação de Atributos

Marina dos Reis Barros Alencar

Teresina-PI, 31 de Março de 2022

Marina dos Reis Barros Alencar

Rotulação Automática de Grupos Através do Uso de Filtros de Ganho de Informação de Atributos

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: Vinicius Ponte Machado

Teresina-PI

31 de Março de 2022

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Sistema de Bibliotecas da UFPI – SIBi/UFPI
Biblioteca Setorial do CCN

A368r Alencar, Marina dos Reis Barros.
Rotulação automática de grupos através do uso de filtros de ganho de informação de atributos / Marina dos Reis Barros Alencar. – 2022.
82 f.: il.

Dissertação (Mestrado) – Universidade Federal do Piauí, Centro de Ciências da Natureza, Pós-Graduação em Ciência da Computação, Teresina, 2022.
"Orientador: Prof. Dr. Vinicius Ponte Machado".

1. Sistemas Operacionais (Computação). 2. Agrupamento de Dados. 3. Rotulaçao de Dados. I. Machado, Vinicius Ponte. II.Título.

CDD 005.43

Bibliotecária: Caryne Maria da Silva Gomes. CRB/3-1461

“Rotulação Automática de Grupos Através do Uso de Filtros de Ganho de Informação de Atributos”

MARINA DOS REIS BARROS ALENCAR

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação do Centro de Ciências da Natureza da Universidade Federal do Piauí, como parte integrante dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Aprovada por:

gov.br Documento assinado digitalmente
VINICIUS PONTE MACHADO
Data: 23/04/2022 10:52:23-0300
Verifique em https://verificador.iti.br

_____

Prof. Vinicius Ponte Machado
(Presidente da banca examinadora)

gov.br Documento assinado digitalmente
TICIANA LINHARES COELHO DA SILVA
Data: 19/04/2022 17:48:43-0300
Verifique em https://verificador.iti.br

_____

Profa. Ticiana Linhares Coelho da Silva
(Examinadora externa à instituição)

gov.br Documento assinado digitalmente
Ricardo de Andrade Lira Rabelo
Data: 22/04/2022 10:35:13-0300
Verifique em https://verificador.iti.br

_____

Prof. Ricardo de Andrade Lira Rabêlo
(Examinador interno)

gov.br Documento assinado digitalmente
Rodrigo de Melo Souza Veras
Data: 20/04/2022 09:33:26-0300
Verifique em https://verificador.iti.br

_____

Prof. Rodrigo de Melo Souza Veras
(Examinador interno)

Teresina, 12 de abril de 2022.

Agradecimentos

Agradeço a Deus por me permitir vivenciar e conseguir superar os desafios dessa jornada. Glórias a Deus por esse dia. Tenho em meu coração imensa gratidão pelo Seu cuidado e amor!

Agradeço aos meus pais, Francisco de Sales Barros (in memory) e Irene Alves dos Reis Barros, por todo amor, dedicação e confiança depositadas em mim. Meu pai não está mais presente entre nós, mas ele sabe que essa vitória é também para ele. Minha eterna gratidão também a minha mãe que sempre esteve comigo nessa caminhada, me dando suporte e apoio, nunca mediu esforços para que eu me dedicasse a vencer esta etapa.

Aos meus irmãos, Tathyane e Anderson, por estarem sempre presente, me incentivando e apoiando, por me darem suporte sempre em que precisei e nunca soltarem minha mão.

Ao meu esposo Guilherme, por me apoiar, ajudar e incentivar em todos os momentos.

À minha filha Laura, por me mostrar quanto amor, força e fé eu tinha guardado comigo, que nunca imaginei ter até sua chegada.

Ao meu orientador Vinicius por ser uma referência para mim como profissional, professor e como pessoa. Sou imensamente grata pela paciência, pelos conselhos, por toda orientação dada até aqui. Sou sua fã de carteirinha e todos com quem eu convivo sabem disso. Parabéns pelo profissional que você é.

Aos meus amigos, que estiveram comigo por toda essa caminhada direta ou indiretamente, que de alguma forma contribuíram.

Aos professores de mestrado por todo conhecimento transmitido.

À CAPES pelo apoio financeiro para realização deste trabalho de pesquisa.

"A vida é feita de começos e recomeços.
Somos todos eternos aprendizes."
(Ivan F. Calori)

Resumo

Identificar semelhanças nos dados que não foram rotulados, classificados ou categorizados é uma das funções do aprendizado não supervisionado. O agrupamento (do inglês clustering) é uma técnica que permite dividir automaticamente o conjunto de dados de acordo com uma similaridade. A grande vantagem do uso das técnicas de agrupamento é que, ao agrupar dados similares, pode-se descrever de forma mais eficiente e eficaz as características peculiares de cada um dos grupos identificados. Dessa forma, este trabalho tem como objetivo a interpretação desses grupos através de rótulos. O rótulo é um conjunto de valores relevantes que representam uma definição para um grupo. Esta abordagem utilizou técnicas com aprendizagem de máquina não supervisionada, aplicação dos filtros de ganho de informação através da seleção de atributos e um modelo de discretização. Na metodologia proposta foi aplicado o algoritmo não supervisionado para formação dos grupos e diferentes filtros de seleção de atributos para expor a relevância dos atributos e comparar o funcionamento deles. Também, para contribuir no processo de rotulação foi utilizado um método de discretização auxiliando no cálculo da variação de valores dos dados. O modelo proposto foi aplicado na rotulação das bases de dados disponíveis no repositório UCI, sendo elas, *Íris*, *Seeds*, *Wine* e *Glass*. Obtendo-se uma taxa de acerto média de 83.66% com desvio padrão médio de 4.98.

Palavras-chaves: Aprendizagem de Máquina. Agrupamento. Rotulação. Variação.

Abstract

Identifying similarities in data that has not been labeled, classified, or categorized is one of the functions of unsupervised learning. Clustering is a technique that allows you to automatically divide the data set according to similarity. The great advantage of using clustering techniques is that, by grouping similar data, it is possible to describe in a more efficient and effective way the peculiar characteristics of each of the identified groups. Thus, this work aims to interpret these groups through labels. The label is a set of relevant values that represent a definition for a group. This approach used techniques with unsupervised machine learning, application of information gain filters through the selection of attributes and a discretization model. In the proposed methodology, the unsupervised algorithm was applied to form the groups and different attribute selection filters to expose the relevance of the attributes and compare their functioning. Also, to contribute to the labeling process, a discretization method was used, helping to calculate the variation of data values. The proposed model was applied in the labeling of the databases available in the UCI repository, namely, Iris, Seeds, Wine and Glass. Obtaining an average hit rate of 83.66% with an average standard deviation of 4.98.

Keywords: Machine Learning. grouping. lettering. Variation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

#Elem	Número de Elementos
AM	Aprendizagem de Máquina
ACL	Alcalinidade das Cinzas
ASH	Cinza
BDM	Base de Dados Modelo
BDMD	Base de Dados Modelo Discretizada
CAIBAL	*Cluster-Attibute Interdependency Based Automatic Labeler*
EWD	*Discretização por Larguras Iguais*
EFD	*Discretização por Frequências Iguais*
GP	Grau de Pertinência
GS	Grau de Seleção
IA	Inteligência Artificial
$ID3$	Indução de Arvore de Decisão
IG	Ganho de Informação
MG	Magnésio
MRA	Modelo de Rotulação Automática
PL	Comprimento da Pétala
PW	*Largura da Pétala*
RNA	Redes Neurais Artificiais
RBF	*Radial Basis Function*
RG	*Razão de Ganho*
SA	Seleção de Atributos
SL	*Comprimento da Sépala*
SW	*Largura da Sépala*

Lista de símbolos

$\vec{u}$	Centro de um grupo
$\sum$	Somatório de um conjunto
$\log_2$	Função de logaritmo na base 2
t	Iterações
K	Número de grupos
N	Número de amostras
$\vec{x}$	Vetor de um grupo
$H(Y)$	Entropia de Y
$P(Y)$	Função de densidade
P_i	Probabilidade de uma Amostra
C_i	Classe de uma amostra

Sumário

1 Introdução

Clusterização também conhecida como agrupamento é um método para a análise exploratória de dados utilizada para auxiliar na resolução de problemas de classificação (BACKER, 1995). É uma técnica de mineração de dados multivariados que através de métodos numéricos e das informações das variáveis de cada caso, tem como objetivo agrupar automaticamente por aprendizado não supervisionado os n casos da base de dados em k grupos, geralmente denominados *clusters* ou grupos.

O agrupamento não tem classes predefinidas, e também, não possui exemplos de treinamento de classes rotuladas. O indutor investiga os exemplos fornecidos e procura determinar se alguns deles pode ser agrupado de alguma maneira, formando agrupamentos ou grupos (CHEESEMAN;STUTZ, 1990). A representação escolhida para os dados, é um importante fator a influenciar um agrupamento, ou seja, a seleção dos atributos envolvidos no agrupamento; de modo que uma boa representação dos dados resulta em grupos isolados e compactos (JAIN, 2010).

A compacidade e a separabilidade são dois aspectos utilizados pelos algoritmos de validação de agrupamento através de duas métricas: a similaridade dos dados dos grupos quão compactos e a dissimilaridade de dados inter-grupos quão isolados com base em medidas de distância e/ou estatísticas (SILVA; PERES; BOSCARIOLI, 2017). Estes aspectos são eficientes ao mensurar a qualidade do agrupamento, mas não conseguem avaliar o conteúdo dos grupos e suas características predominantes, que são fatores diretamente relacionados com a seleção de atributos destacada na literatura. Sendo assim, negligencia-se a compreensão dos grupos formados, ou seja, a interpretabilidade dos grupos e usabilidade dos conhecimentos obtidos, o que é de fato o objetivo ao desempenhar um agrupamento (HAN; PEI; KAMBER, 2011).

Um problema relacionado ao agrupamento é a interpretação dos grupos formados que pode ser dificultada pela quantidade de instâncias e atributos presentes na base. Para resolução deste problema tem-se os modelos de rotulação que buscam gerar rótulos que identifiquem de forma única cada grupo a fim de facilitar sua interpretação. Os modelos de rotulação, de modo geral, dão-se pela combinação de técnicas de Aprendizado de Máquina já existentes (ARAUJO; VERAS; MACHADO, 2019).

Atualmente muitas pesquisas tem sido desenvolvidas com o intuito de analisar o agrupamento com foco nas particularidade dos dados (LIMA; MACHADO; LOPES, 2015; MACHADO; RIBEIRO; RABÊLO, 2015; IMPERES Filho et al., 2020; MOURA, 2020), uma tarefa denominada Rotulação de Grupos (SILVA, 2021).

A rotulação sumariza as características comuns dos elementos dos grupos em

um rótulo, facilitando o entendimento e interpretação sobre dados por meio dos valores apresentados pelas características mais importantes de seus elementos (LOPES et al., 2016), podendo ser apresentado como uma ferramenta de auxílio ao especialista (SILVA, 2021).

Para a formação desses rótulos é necessária a aplicação de diferentes técnicas, que preparam, manipulam e processam os grupos com objetivo de identificar o relacionamento entre eles, selecionando as informações destacadas por esses relacionamentos como mais relevantes para formação dos grupos. A técnica utilizada neste trabalho é o filtro ganho de informação baseado na seleção de atributos.

A seleção de atributos é uma técnica que seleciona o melhor subconjunto do grupo original de dados, de acordo com a estratégia de busca e a medida de avaliação utilizada. É um método presente na maioria dos algoritmos de aprendizagem de máquina aplicados na mineração de dados porque determinam os atributos que mais influenciam na tomada de decisão (LLERENA, 2017). Uma maneira de mensurar a qualidade de um atributo para a classificação é avaliar seu grau de associação com a classe. Sendo assim, utiliza-se uma ou mais métricas para a avaliação e ordena-se os atributos de forma decrescente em um *ranking*. A escolha do subconjunto dá-se por meio da comparação dos valores do *ranking* de cada atributo com um limiar *(threshold)*, ou ainda, por um número fixo de melhores atributos, em termos absolutos ou percentuais.

No caso da abordagem filtro, a seleção dos atributos é executada sem considerar o algoritmo de mineração de dados, analisando cada atributo (ou um subconjunto de atributos) determinando o nível de correlação entre os atributos e as classes ou analisando os subconjuntos procurando uma forma heurística o melhor, confiando em características gerais dos dados para validar e selecionar subconjuntos (ALMEIDA et al., 2018a).

Para determinar o grau de associação do atributo com a classe, foram propostas diversas métricas. Essas métricas são empregadas em uma base de dados $D(A_1, A_2, ..., A_n, C), n \geq 1$, com $n + 1$ atributos, onde C é o atributo classe e o seu domínio é $c_1, c_2, c_m, m \geq 2$ (PEREIRA, 2009).

1.1 Motivação e Objetivos

A literatura apresenta alguns modelos de rotulação de grupos, cada um com suas finalidades e limitações. Dentre os trabalhos relacionados ressalta-se o modelo proposto por Machado, Ribeiro e Rabêlo (2013) onde para cada grupo gerado são criadas redes neurais artificiais, com uma certa quantidade de características que descreve um elemento do problema. Cada RNA (Rede Neural Artificial) criada é responsável por avaliar a relevância de um atributo como um potencial candidato ao rótulo de seu grupo. Assim cada rede neural artificial, por sua vez, apresenta uma taxa de acerto em relação ao seu aprendizado,

realizado apenas com os elementos de seus respectivos grupos. Para dar uma maior precisão e confiança às taxas de acerto, foi aplicado o método de *holdout (random subsampling)*, onde o processo é repetido de modo iterativo. Assim, para cada RNA é calculada a média de suas taxas de acerto obtidas em M iterações.

O trabalho de Araujo, Veras e Machado (2019) propõe um modelo de classificação baseado em lógica*Fuzzy* onde utiliza os rótulos gerados pelos modelos de rotulação automática para formação das regras e funções de pertinência. Este trabalho também utiliza Redes Neurais para selecionar os atributos mais relevantes de cada grupo.

Mais uma abordagem que desenvolve pesquisa como método para resolução do problema de rotulação é Imperes Filho et al. (2020), onde usa algoritmo de agrupamento não supervisionado baseado em distância em Grau de Pertinência (GP), onde cada elemento da base de dados analisada recebe um grau de pertinência em relação a cada grupo formado. Com o objetivo formular as faixas de valores para cada atributo em cada grupo formado verifica a existência de interseções entre faixas de valores e, por fim, monta e exibe os rótulos de cada grupo com faixas de valores que não possuem interseção.

Já o trabalho apresentado por Silva (2021) utiliza em sua abordagem regressões para estimar a função que descreve o erro de predição dos atributos em seus respectivos domínios por grupo, permitindo delimitar intervalo em que o erro de predição é o menor possível para cada atributo de cada grupo. Sendo assim, são selecionados os pares atributo-intervalo que representem a maioria dos elementos dos grupos, bem como sejam capazes de distingui-los, para composição dos rótulos.

Com base nesses aspectos, este trabalho tem como objetivo propor um modelo de rotulação de grupos cuja seleção dos atributos é feita pelos filtros de ganho de informação e faixas de valores associados aos rótulos que é baseada em um modelo de discretização de valores. Consequentemente temos como objetivos específicos:

- Aplicação dos filtros de ganho de informação com intuito da descoberta da relevância dos dados;
- A comparação da eficiência desses filtros de seleção para cada base;
- Etapa de discretização de dados.

1.2 Estrutura do Trabalho

Este trabalho está organizado da seguinte forma: Capítulo 1 apresenta a introdução, o Capítulo 2 abrange conceitos e métodos de aprendizagem de máquina utilizados no modelo proposto e traz uma explanação sobre o problema do agrupamento e a rotulação de grupos. O Capítulo 3 aborda os modelos de rotulação de grupos da literatura. O Capítulo

4 apresenta o modelo proposto, seguido dos resultados experimentais para a rotulação de bases de dados da literatura no Capítulo 5. E para finalizar, o Capítulo 6 discute as conclusões resultantes dos experimentos e as próximas etapas da pesquisa.

1.3 Contribuições da Dissertação

Este trabalho propõe a utilização de um método de seleção de atributos para auxiliar no processo de rotulação de dados. Além de contribuir também para o área de pesquisa com uma revisão de outros modelos de rotulação da literatura.

2 Referencial Teórico

Neste capítulo é apresentado o fundamento teórico para contribuir com a compreensão do presente trabalho. Estas explanações mostram como cada técnica atua no método proposto. Na primeira seção apresenta-se uma introdução à área de aprendizagem de máquina, mais focado em um de seus principais paradigmas de aprendizagem, o aprendizado não supervisionado. Neste capítulo também é exposto como cada filtros de seleção de atributos funciona. E os modelos de discretização que são utilizados.

2.1 Aprendizagem de Máquina

Aprendizagem de máquina (AM) é o ramo da inteligência artificial que estuda o desenvolvimento de sistemas que são capazes de aprender com a experiência. Isto é, um sistema que realiza determinada tarefa e que possa obter informações que permitam execuções futuras desta tarefa com um melhor desempenho (MITCHELL, 1997). Existem três principais paradigmas de aprendizagem de máquina.

O aprendizado supervisionado implica necessariamente na existência de dados de entrada e na indicação de uma saída que possa ser aprendida para ocorrer o processo de aprendizagem (BRAGA; CARVALHO; LUDEMIR, 2007).

Já o aprendizado não supervisionado envolve a aprendizagem de padrões na entrada, quando não são fornecidos valores de saídas específicas (NORVIG; RUSSELL, 2004).

Temos ainda o aprendizado por reforço que se dá por meio de um agente que toma decisões, em um ambiente e é recompensado ou punido por suas ações, tendo por objetivo aprender a melhor sequência de decisões a tomar para acumular o maior número de recompensas (SUTTON; BARTO et al., 1998).

A capacidade de produção de dados gerados por meio dos mais variados meios, tem dificultado o processo de interpretação para muitos especialistas que dispõe na análise das informações seu maior recurso para tomada de decisão. Nesse contexto, prover mecanismos que possibilitem a correta interpretação e uso racional dos dados tem sido motivo de estudos para muitos pesquisadores. Uma possível forma de tratar os dados produzidos em demasia é por meio do agrupamento de dados, uma sub-área da Aprendizagem de Máquina Não Supervisionada (IMPERES Filho et al., 2020).

2.1.1 Aprendizagem Não Supervisionada

O aprendizado não supervisionado é um ramo do aprendizado de máquina que aprende com dados de treinamento que não foram rotulados, classificados ou categorizados previamente. Em vez de responder à programação de um operador, o aprendizado não supervisionado identifica semelhanças nos dados e reage com base na presença ou ausência de tais semelhanças em cada novo dado.

Esta técnica tem como objetivo selecionar um conjunto de dados e agrupá-los de acordo com alguma similaridade. Os algoritmos de agrupamento de dados foram desenvolvidos como uma ferramenta para relacionar grande quantidade de dados gerados por diferentes sistemas (ALGULIYEV et al., 2016).

O agrupamento como eixo de pesquisa da aprendizagem de máquina é comumente considerado o problema mais relevante em aprendizado não supervisionado aplicado à organização de informações não rotuladas (POPAT; EMMANUEL, 2014).

A compreensão dos grupos deve-se principalmente, aos valores apresentados pelas características mais importantes de seus objetos. Assim, este conjunto de valores relevantes representa uma definição para um grupo qualquer.

2.1.2 K-means

Este trabalho utiliza o algoritmo *k-means* para realizar o agrupamento dos dados, para assim, aplicar os filtros de seleção de atributos e formar rótulos, contribuindo na compreensão e auxiliando especialistas no processo de tomada de decisão.

O *k-means* é um algoritmo de agrupamento de dados não hierárquicos que utiliza uma técnica iterativa para particionar um conjunto de dados. Proposto num trabalho pioneiro de Stuart Lloyd em 1957, mas, só foi publicado no ano de 1982 (LLOYD, 1982). Esse algoritmo busca minimizar a distância dos elementos de um conjunto de dados com k centros de forma iterativa.

A desvantagem é a sua dependência dos valores iniciais de k, da ordem em que as amostras são processadas, da escolha dos primeiros centros de agrupamento e da geometria das amostras disponíveis para análise. Em alguns casos sua utilização requer experimentação com vários valores de k e diferentes escolhas dos parâmetros iniciais (HART; STORK; DUDA, 2000).

A ideia por trás deste algoritmo é escolher K objetos aleatoriamente ou com alguma heurística que serão à base de cada grupo (denominados centroides), os demais objetos são associados ao centroide mais próximo. A cada passo os centroides são recalculados dentre os objetos de seu próprio grupo e os objetos são realocados para o centroide mais próximo, este procedimento é repetido até que o nível de convergência seja satisfatório de

acordo com alguma heurística estabelecida.

O objetivo do *k-means* é minimizar a média quadrática da distância euclidiana entre os centros de cada grupo e seus respectivos elementos, onde cada centro ou centroide $\vec{u}$ de um dado grupo c_i é definido como a média de seus elementos, representados individualmente por $\vec{x}$, conforme Equação 2.1 (MANNING; RAGHAVAN; SCHÜTZE, 2009).

$$\vec{u}\ (ci) = \frac{1}{|ci|} \cdot \sum_{x \in c_i} \vec{x} \tag{2.1}$$

A complexidade computacional em tempo polinomial do *k-means* é $\theta(t \cdot K \cdot N)$, em que t é o número de iterações, K é o número de *clusters* a serem gerados e N é o número de amostras, de modo que geralmente tem-se t, $K \ll N$.

Devido a sua complexidade, com o algoritmo *k-means* é viável processar amostras muito grandes de dados. Algumas possíveis aplicações que podem utilizar a eficiência do algoritmo não supervisionado *k-means* incluem métodos para agrupamento de similaridades, previsão não-linear, aproximação de distribuições multivariadas, testes não paramétricos para independência entre várias variáveis e classificação de árvores baseadas em distância (MACQUEEN et al., 1967). Devido à sua estrutura compacta e eficiência, o *k-means* foi escolhido para lidar com o problema de agrupamento encontrado neste trabalho.

2.2 Seleção de Atributos

A seleção de atributo é um tema de pesquisa e desenvolvimento constante desde os anos 70 nas áreas de reconhecimento de padrões, aprendizado de maquina e mineração de dados (LIU et al., 2009).

A seleção de atributos é uma técnica utilizada para a redução da dimensionalidade da base de dados, tendo como objetivo reconhecer os atributos relevantes aumentando assim o poder preditivo do classificador. Exitem três abordagens para a seleção de atributos: Embutida, Filtro e *Wrapper* (KOHAVI; JOHN, 1997).

Na abordagem Embutida é identificada quando a seleção de atributo é realizada internamente pelo próprio indutor durante seu treinamento. Dado um conjunto de exemplos representado no formato atributo-valor, o próprio algoritmo de aprendizado de máquina é capaz de decidir quais são os atributos relevantes para representar o conhecimento extraído.

Já a abordagem *Wrapper*, o algoritmo de classificação é executado para cada subconjunto e a avaliação geralmente é feita em termos da acurácia preditiva retornada pelo algoritmo.

E a abordagem Filtro não necessita do auxílio de um indutor para a avaliação dos subconjuntos de atributos (YU; LIU, 2003). No contexto de filtrar atributos relevantes,

segundo algum critério, este trabalho propõe a utilização do método de seleção de atributos baseado na abordagem Filtro para auxiliar no processo de rotulação de dados.

Um dos problemas enfrentados na área de mineração de dados é que as bases possuem grande volume de dados. Sendo assim, a seleção de atributos se faz necessária, pois tem como objetivo remover atributos irrelevantes e/ou redundantes, reduzindo assim a dimensionalidade dessas bases (GUYON; ELISSEEFF, 2006).

As abordagens de seleção de atributos do tipo filtro são independentes do algoritmo de classificação. Esses métodos podem avaliar cada atributo independente dos outros, determinando o grau de correlação entre cada atributo e a classe (YANG; PEDERSEN, 1997).

O termo filtro deriva da ideia de que os atributos irrelevantes são filtrados da base de dados antes da aplicação do algoritmo de classificação (BLUM; LANGLEY, 1997). Os filtros usam as informações da própria base de treinamento para escolher os atributos a serem utilizados posteriormente (ALMEIDA et al., 2018b). Seleção de atributos é usada como método de pesquisa para gerar uma lista de classificação da qual gera-se um *ranker* e descarta um determinado número de atributos, podendo ser apenas ruídos ou atributos que não tenham relevância para o processo de aprendizagem.

Diversas métricas de classificação e seleção de atributos foram propostas na literatura de mineração de dados. O objetivo dessas métricas é descartar irrelevantes ou recursos redundantes de um determinado vetor de recursos. Neste trabalho foram utilizadas quatro filtros de seleção de atributos: O Ganho de Informação *(do inglês Information Gain)*; Razão de Ganho *(do inglês Gain Ratio)*; Correlação *(do inglês Correlation)* e *ReliefF*.

Essas métricas têm como objetivo criar um *ranking* com os atributos das bases de dados, sendo assim, permite verificar a relevância de cada atributo da base e auxilia na composição dos rótulos de dados.

2.2.1 Ganho de Informação

O ganho de informação é um filtro de seleção utilizado para medir a semelhança entre o atributo de uma classe, com todas as outras classes, diminuindo o grau de incerteza (LACERDA et al., 2019). Neste caso, porém, a abordagem do filtro é a de *feature ranking*, que avalia o mérito de cada atributo individualmente. Como o nome sugere, ele considera quanta informação é obtida para a classificação, dado que o atributo seja considerado.

A entropia é comumente usada como medida na teoria da informação, que caracteriza a pureza de uma arbitrária coleção de exemplos. Ela faz parte dos métodos de classificação de atributos do ganho de informação. A medida de entropia é considerada como uma medida da imprevisibilidade do sistema. A entropia de Y é calculada conforme Equação 2.2.

$$H(Y) = \sum_{y \in Y} p(y) log_2(p(y)) \tag{2.2}$$

Onde $p(y)$ é a função de densidade de probabilidade marginal para a variável aleatória Y. Se os valores observados de Y no conjunto de dados de treinamento S são particionados de acordo com os valores de X, e a entropia de Y em relação as partições induzidas por X são menores que a entropia de Y antes do particionamento, então há uma relação entre $Y e X$. Então a entropia de Y após observar X é apresentada na Equação 2.3.

$$H(y/x) = \sum_{x \in X} p(x) \sum_{x \in X} p(y/x) log_2(p(y/x)) \tag{2.3}$$

Onde $p(y/x)$ é a probabilidade condicional de y dado x . Dada a entropia como critério de impureza em um treinamento de conjunto S, pode definir uma medida refletindo informações sobre Y fornecidas por X que representam a quantidade pela qual a entropia de Y diminui. Esta medida é conhecida como ganho de informação, conforme Equação 2.4.

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \tag{2.4}$$

O ganho de informação é uma medida simétrica como apresentado na Equação 2.4. A informação obtida sobre Y após observar X é igual as informações obtidas sobre X após observar Y. A fraqueza do critério de ganho de informação é que ele é tendencioso a favor de recursos com mais valores, ou seja, uma maior quantidade de dados, mesmo quando não são mais informativos.

2.2.2 Razão de Ganho

O filtro razão de ganho (RG) é caracterizado pelo uso de uma métrica para ranquear todos os atributos de uma base de dados, calculada por meio do ganho de informações de um atributo contra o número de saídas testes, com a maior possibilidade de valores (NETTO, 2013).

Razão de ganho é uma medida que avalia a utilidade de um atributo ao medir a taxa de ganho que o mesmo proporciona à discriminação das classes a serem aprendidas. Trata-se de uma modificação da métrica ganho de informação (WITTEN; FRANK, 2002) que visa reduzir o viés provocado por ganho de informação de preferir atributos que apresentam a maior faixa de valores possível. A fim de determinar os melhores atributos, experimentos de seleção de atributos foram então conduzidos, norteados pelos métodos existentes na literatura.

O cálculo do ganho de informação favorece atributos com um grande número de valores. Uma forma de reduzir esse problema é utilizar a razão de ganho, que aplica

uma espécie de normalização do ganho (CASTRO; FERRARI, 2016). Essa normalização tem intenção de reduzir os valores de RG, decrescendo no intervalo $[0, 1]$, onde 1 aponta que a informação advinda da variável independente (preditor) prediz completamente a variável resposta; para GR = 0 indica que não há relação entre o preditor e a variável de resposta, assim, quanto maior o valor de RG, maior a relevância da variável (ASDAGHI; SOLEIMANI, 2019; NOVAKOVIĆ, 2016).

O ganho de informação medida é usado para selecionar o atributo de teste em cada nó de árvore de decisão. Uma árvore de decisão é uma estrutura simples onde nós não terminais representam testes em um ou mais atributos e nós terminais refletem os resultados da decisão. A medida de ganho de informação prefere selecionar atributos com um grande número de valores.

O algoritmo de indução de árvore de decisão ID3 foi aprimorado por C4.5. Ele é um sucessor de ID3, o qual usa uma extensão de ganho de informação conhecida como razão de ganho, onde gera árvores mais precisas e menos complexas e apresenta um método de pós-poda das árvores geradas.

Seja S definido consistindo em s amostras de dados com m distintos atributos. As informações esperadas necessárias para classificar uma determinada amostra como apresentada pela Equação 2.5.

$$H(\frac{Y}{X}) = \sum_{i=1}^{m} P_i log_2(Pi) \tag{2.5}$$

Onde P_i é a probabilidade de que uma amostra arbitrária pertença à classe C_i e é estimado por s_i/s.

O atributo A tem v valores distintos. Onde o número de amostras da classe C_i em um subconjunto $S_i \cdot S_j$ contendo aquelas amostras em S que têm valor de a_j de A. A entropia, ou esperada informações baseadas no particionamento em subconjuntos por A, são apresentadas pela Equação 2.6.

$$E(A) = -\sum_{i=1}^{m} I(S) \frac{S_{1i} + S_{2i} + ... + S_{mi}}{s} \tag{2.6}$$

As informações de codificação que seriam obtidas por ramificação em A é apresentada pela Equação 2.7.

$$Gain(A) = I(S) - E(A) \tag{2.7}$$

2.2.3 Correlação de Atributos

O coeficiente de correlação de Pearson remonta o trabalho conjunto de Karl Pearson e Francis Galton (STANTON, 2001). É um teste que mede a relação estatística entre duas variáveis contínuas. Se a associação entre os elementos não for linear, o coeficiente não será representado adequadamente. O coeficiente pode ter um intervalo de valores de +1 a -1. Um valor de 0 indica que não há associação entre as duas variáveis.

Um valor maior que 0 indica uma associação positiva. Isto é, à medida que o valor de uma variável aumenta, o mesmo acontece com o valor da outra variável. Um valor menor que 0 indica uma associação negativa. Isto é, à medida que o valor de uma variável aumenta o valor da outra diminui.

Para Cohen (1988), valores entre $0,10$ e $0,29$ podem ser considerados pequenos; escores entre $0,30$ e $0,49$ podem ser considerados como médios; e valores entre $0,50$ e 1 podem ser interpretados como grandes.

Dancey e Reidy (2006) apontam para uma classificação ligeiramente diferente: $r = 0,10$ até $0,30$ (fraco); $r = 0,40$ até $0,6$ (moderado); $r = 0,70$ até 1 (forte). Seja como for, o certo é que quanto mais perto de 1 (independente do sinal) maior é o grau de dependência estatística linear entre as variáveis. No outro oposto, quanto mais próximo de zero, menor é a força dessa relação.

O coeficiente pode ser usado para resumir a força da relação linear entre duas amostras de dados. Ele calculado como a covariância das duas variáveis dividida pelo produto do desvio padrão de cada amostra de dados. É a normalização da covariância entre as duas variáveis para dar uma pontuação interpretável, como apresentado na Equação 2.8.

$$CorrelationCoef = covariance(X, Y)/(stdv(X) * stdv(Y)) \tag{2.8}$$

O uso da média e do desvio padrão no cálculo sugere a necessidade de duas amostras de dados terem uma distribuição gaussiana ou semelhante à gaussiana.

Baseado em Moore e McCabe (1989), detaca-se as propriedades do coeficiente e as condições que precisam ser satisfeitas para realizar a análise de correlação de Pearson. Portanto, as observações são as seguintes:

- O coeficiente de correlação de Pearson não diferencia entre variáveis independentes e variáveis dependentes. Dessa forma, o valor da correlação entre X e Y é o mesmo entre Y e X. Para Schield (1995) a correlação não se aplica a distinção de causalidades simples ou recursiva.

- O valor da correlação não muda ao se alterar a unidade de mensuração das variáveis. Por ser tratar de uma medida padronizada, o valor da correlação entre quilos e litros será o mesmo caso o pesquisador utilize toneladas e mililitros. Padronização torna possível a comparação entre diferentes variáveis no que diz respeito a sua magnitude e dispersão.
- O coeficiente tem um caráter adimensional, ou seja, ele é desprovido de unidade física que o defina. Não faz sentido interpretar uma correlação de $0,3$ como sendo 30%, por exemplo. Além disso, ele não se refere à proporção. Logo, uma correlação de $0,4$ não pode ser interpretada como representando o dobro de uma correlação de $0,2$ (CHEN et al., 2002).

Para além das propriedades do coeficiente, algumas condições precisam ser satisfeitas. A correlação exige que as variáveis sejam quantitativas (contínuas ou discretas). Não faz sentido utilizar a correlação de Pearson para dados categóricos. Os valores observados precisam estar normalmente distribuídos.

2.2.4 ReliefF

Dentre as técnicas multivariadas, o algoritmo de seleção de atributos reliefF (KONONENKO, 1994) foi baseado em um algoritmo previamente desenvolvimento chamado Relief (KIRA; RENDELL, 1992).

O algoritmo original relief tem como objetivo avaliar quão bem o atributo é capaz de distinguir entre duas classes, amostrando aleatoriamente uma instância e estimando uma medida de distância entre a instância original e as instâncias mais próximas, uma da mesma classe e a outra de classe diferente. Este algoritmo, então, calcula a distância entre os vizinhos mais próximos para ambas as classes e atualiza uma variável de peso W_i, conforme a Equação 2.9 (KIRA; RENDELL, 1992).

$$W_i = W_i - (x_i - x_i^+)^2 + (x_i - x_i^-)^2 \tag{2.9}$$

O subíndice i refere-se ao atributo analisado, x_i é o atributo i da instância x analisada, enquanto x_i^+ e x_i^- são os atributos das instâncias mais próximas de x para instâncias da mesma classe e de classes diferentes.

Assim, este algoritmo tem como princípio que um bom atributo deve possuir valores similares para instâncias de uma mesma classe e também deve possuir valores diferentes para instâncias de classes diferentes. Ou seja, quanto maior for o peso W_i , mais relevante é o atributo.

Posteriormente, este algoritmo foi modificado de modo a aumentar a confiança na estimativa de quão bem cada atributo consegue separar duas classes. Para isso, foi

aumentado o número de instâncias mais próximas calculadas. Ou seja, no algoritmo relief eram calculadas as distâncias apenas dos vizinhos mais próximos, da mesma classe e da classe oposta. Na modificação proposta no algoritmo reliefF, pode-se calcular esta distância para os m vizinhos mais próximos. Assim, o número de vizinhos mais próximos calculados é um parâmetro que pode ser variado conforme desejado, sendo que quanto maior for este valor, maior é a confiança na estimativa e mais resistente e robusto a ruídos torna-se o algoritmo. Entretanto, nota-se também que quanto maior for este valor, maior será também o tempo necessário para compilação do algoritmo (KONONENKO, 1994).

Outra mudança entre os algoritmos, é que no primeiro, a distância entre os atributos é a euclidiana, que é a raiz quadrada da diferença entre os valores quadráticos dos atributos, ou seja, quando o valor de r na Equação 2.10 é 2. Já no algoritmo reliefF, a distância é calculada como a soma das diferenças absolutas entre os valores dos atributos, ou seja, quando o valor de r na Equação 2.10 é 1 (THEODORIDIS; KOUTROUMBAS, 2008).

$$dist\hat{a}ncia = (\sum_{k=1}^{m} |q - q_k|^r)^{1/r} \tag{2.10}$$

Nesta equação, m é o número de vizinhos mais próximos calculados, q e q_k são os valores do atributo q e os valores de seus k vizinhos. Para apenas um vizinho mais próximo calculado, percebe-se que ambas as distâncias são equivalentes. Entretanto, quando calcula-se para um número maior que 1, as distâncias euclidiana e de Manhattan admitem valores diferentes (ALBONICO et al., 2017).

ReliefF é um algoritmo que tende a ser bastante eficiente em identificar quais os melhores e piores atributos. Além disso, pode ser utilizado apenas para calcular a relevância de um atributo por vez, ou seja, realizar uma análise univariada, ou então calcular o peso de um conjunto de atributos por vez, realizando uma análise multivariada. Entretanto, sua principal desvantagem é não ser capaz de detectar se existe redundância entre os atributos selecionados. Assim, caso existam alguns atributos muito bons, ainda que possuam o mesmo tipo de informação, estes serão selecionados.

Porém, com este tipo de abordagem, torna-se possível identificar quais atributos são bons ou ruins para realizar a classificação, tendo normalmente pesos bastante altos para atributos bons e pesos bastante baixos para atributos ruins. Tornando essa distinção evidente, pode-se ainda utilizar este tipo de abordagem como uma etapa de pré-processamento à etapa de seleção de atributos, reduzindo o conjunto total de atributos analisados por uma abordagem mais complexa, como, por exemplo, a abordagem Wrapper (ALBONICO et al., 2017).

2.3 Discretização

O método de discretização faz a conversão de valores contínuos em valores discretos. A partir de um atributo com valores contínuos, a discretização cria um ponto inicial e final definindo um intervalo e designando uma faixa para cada intervalo. Assim, ao invés de valores contínuos, teremos valores discretos representando as faixas de valores (JAIME, 2020).

No contexto da rotulação a discretização permite a inferência de uma faixa de valor, ou seja, os atributos que podem assumir diferentes valores dentre um domínio contínuo são estabelecidos novos valores discretos. Assim, a discretização auxilia no agrupamento, identificando com menor complexidade uma possível relação entre os atributos e expondo melhores resultados ao enfrentar o problema de classificação envolvendo tais atributos.

A discretização em si pode também ser encarada como uma forma de descoberta de conhecimento, onde valores críticos em um domínio contínuo podem ser revelados. Os intervalos de discretização, portanto, não devem esconder padrões de relacionamento entre as variáveis. Eles devem ser escolhidos com cuidado ou descobertas potenciais podem ser perdidas (YONEYAMA, 2003).

O objetivo de um método de discretização consiste em encontrar um conjunto de pontos de corte de modo à particionar uma faixa de valores contínuos em um conjunto de pequenos intervalos. Assim, um valor discreto será associado a cada intervalo diferente de valores contínuos (KOTSIANTIS; KANELLOPOULOS, 2006).

Apesar da perda de informação, a discretização tem uma função importante na rotulação e conforme os tipos de dados, método e faixas de valores, os rótulos podem sofrer alterações. Pode haver registros que estejam representados em uma determinada faixa e, ao alterar o método de discretização, pode também alterar o tamanho (limites) da faixa, fazendo esse registro mudar de faixa, consequentemente alterando o rótulo, e gerando uma nova visão ao analista do grupo.

Existem vários métodos de discretização, grande parte deles foi desenvolvido para sistemas de classificação, outros têm uma utilização mais genérica.

De acordo com Kotsiantis e Kanellopoulos (2006), Dougherty, Kohavi e Sahami (1995), os métodos de discretização mais comumente utilizados, no âmbito dos métodos não supervisionados, são os de Discretização por Larguras Iguais (do inglês: Equal Width Discretization - EWD) e Discretização por Frequências Iguais (do inglês: Equal Frequency Discretization - EFD).

Esses métodos de discretização não consideram qualquer informação sobre possíveis relacionamentos entre as variáveis. Sendo assim, podem fazer com que os padrões de relacionamentos sejam perdidos. Contudo sua simplicidade de implementação em relação

ao desempenho com outros métodos de discretização os torna muito populares. A maioria dos métodos de discretização, como EWD e EFD produzem intervalos disjuntos, onde cada valor só pode pertencer a um único intervalo.

2.3.1 Discretização por Larguras Iguais - EWD

O método de discretização EWD faz a discretização de um intervalo, entre valores contínuos, dividindo através de um ponto de corte as faixas de tamanhos iguais. (BARON, 2016). No entanto, esta distribuição de categorias pode ser muito desbalanceada e às vezes pode dificultar a capacidade de o atributo ajudar na construção do modelo de decisão.

Sendo a quantidade de valores R a serem gerados na discretização e um determinado intervalo $[a, b]$ representado por valores contínuos, o método EWD divide o intervalo de valores dos atributos em tamanhos iguais R. Assim, para gerar R intervalos serão necessários $R - 1$ pontos de corte. É importante destacar que os elementos pertencentes ao intervalo devem estar ordenados de forma crescente. A largura de cada faixa de valor $(r_1, ..., r_R)$ é representada por ω e pode ser calculada pela diferença entre os limites superiores e inferiores do intervalo dividido pela quantidade R de valores a serem gerados, conforme Equação 2.11.

$$\omega = \frac{b - a}{R} \tag{2.11}$$

Após descobrir o valor de ω, pode-se determinar os pontos de cortes $(C_1, ..., C_{R-1})$ estes pontos determinam as faixas de valores. O primeiro ponto de corte,C_1, é dado pela soma do limite inferior com a distância ω. Os outros pontos de corte $(C_1, .., C_R)$ podem ser calculados pela soma do ponto de corte anterior com ω, conforme Equação 2.12.

$$c_i = \begin{cases} a + \omega, & \text{se } i = 1 \\ c_{i-1} + \omega, & \text{caso contrário} \end{cases} \tag{2.12}$$

Os valores contínuos passarão a ser representados por i, onde i é o índice que indica a qual faixa r_i o valor se encontra. Por exemplo, para se dividir o intervalo $[a, b]$ em R faixas de valores distintas, precisar-se-á de $R - 1$ pontos de corte (LOPES et al., 2016). A Figura 1 mostra esse modelo de discretização. Qualquer valor pertencente ao intervalo $[a, c_1]$ terá um valor discreto associado igual ao índice de sua faixa r_1 . Isto é, um valor que se encontra na faixa r_1 passará a ser representado pelo valor 1. De maneira análoga um valor que se encontra na faixa $r_2 =]c_1, c2]$ por 2 e, finalmente, um valor que se encontra em uma faixa qualquer r_i será representado por i.

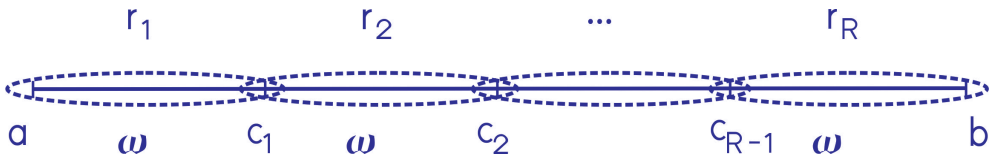

Figura 1 – Discretização por EWD

2.3.2 Discretização por Frequências Iguais - EFD

Ao contrário do EWD, o EFD desloca as margens dos valores para que todas as categorias tenham o mesmo número de instâncias. Em outras palavras, os limites das categorias são definidos pela distribuição, sendo que em cada categoria há o mesmo número de instâncias.

Dado um determinado intervalo $[a, b]$, a quantidade de valores R a serem gerados por discretização, e a quantidade de elementos com valores distintos, $\varepsilon(\varepsilon \geq R)$, ao longo deste intervalo a discretização por EFD irá dividi-lo em R faixas de valores que tenham a mesma quantidade de elementos distintos para gerar R faixas de valores, $(r_1, ..., r_R)$, será necessário um total de $R - 1$ pontos de corte, de modo que cada faixa contenha a mesma quantidade de elementos distintos que pode ser calculada pelo valor inteiro da divisão entre a quantidade de elementos distintos e a quantidade de faixas de valores, como apresentado na Equação 2.13 .

$$\lambda = \frac{\varepsilon}{R} \tag{2.13}$$

Neste método observa-se que há uma má distribuição de valores entre as faixas, quando possui vários valores repetidos de um atributo, isso ocasiona um desequilíbrio na distribuição dos elementos dentro da faixa. Essa situação reflete em faixas com muitos valores e outras com poucos. Quando há um vetor de números distintos ordenado, e calculado o λ através do número de faixas, obtém-se os pontos de corte $(c_1, ..., c_{R-1})$ como delimitadores das faixas do $i.\lambda$-ésimo elemento conforme os valores do intervalo $[a, b]$, apresentados na Equação 2.14.

$$c_i = v_{i\lambda} \tag{2.14}$$

Diferente do método EWD, as faixas de valores podem assumir tamanhos diferentes, sendo assim, os valores contínuos passarão a ser representados por i onde i é o índice que indica a qual r_i o valor se encontra. Desse modo, é necessário $R - 1$ pontos de corte, apresentados na Figura 2, para se dividir o intervalo $[a, b]$ em R faixas de valores diferentes. O valor encontrado na faixa r_1 será representado pelo valor 1 e de maneira análoga, o valor encontrado na faixa $r_2 =]c_1, c_2]$ por 2. Para finalizar, um valor que se encontra em

uma faixa qualquer r_i será representado por i. Ou seja, para qualquer valor pertencente ao intervalo $[a, c_1]$ terá um valor discreto associado igual ao índice de sua faixa r_1.

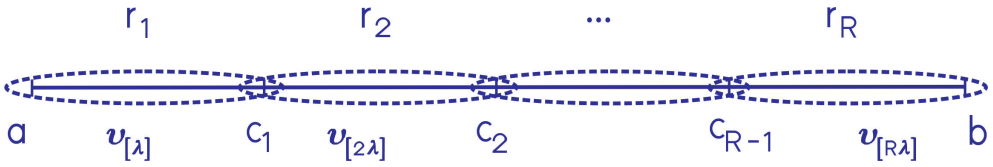

Figura 2 – Discretização por EFD

2.4 Rotulação de Grupos

Existem diversas pesquisas acerca do problema de clusterização, entretanto, poucas são as que focam em rotular os grupos resultantes. A rotulação de um grupo busca resumir sua definição, ou seja, descrevê-lo em função de seus atributos mais relevantes – ou seja, aqueles que são determinantes para o agrupamento – e suas respectivas faixas de valores, a fim de melhor compreendê-lo. Assim, esse conjunto de valores representa uma definição para um grupo qualquer – isto é, um rótulo – capaz de fornecer ao especialista um melhor entendimento sobre os dados (ARAÚJO, 2018).

Segundo Tzerpos (2001), em um esforço para maximizar o desempenho e precisão de algoritmos que lidam com esse problema, muitos pesquisadores desviaram-se do fato de que o principal objetivo era, a princípio, a compreensão dos grupos formados e não a satisfação de um critério abstrato, como por exemplo, a maximização dos graus de similaridade e dissimilaridade entre elementos intra e extra-grupos respectivamente.

O agrupamento consiste em definir um conjunto de grupos que são coerentes internamente, porém claramente diferentes entre si. Isto é, os elementos de um grupo devem ser tão similares quanto possível entre si e, ao mesmo tempo, tão diferentes quanto possível em relação aos elementos de outros grupos (MANNING; RAGHAVAN; SCHÜTZE, 2009).

O trabalho de Yeganova et al. (2009) utiliza a definição de dados rotulados em uma perspectiva que faz detecção e identificação de abreviações na literatura biomédica utilizando aprendizado de máquina supervisionado. É feito uma extração de estruturas textuais (formas curtas, formas longas, formas curtas potenciais e formas longas potenciais) por meio dos textos, que ao serem extraídas naturalmente em pares *i.g.* (forma curta - forma longa), (formas curtas potenciais - formas longas potenciais) são tratadas como exemplos positivos.

Treeratpituk e Callan (2006) trabalham com grupos hierárquicos, de forma que cada grupo pode ser subdivido em subgrupos de forma recursiva. Além disso, os rótulos são restritos a informações textuais. O trabalho proposto por Machado, Ribeiro e Rabêlo

(2013) é uma ferramenta que realiza a extração automática de características dos grupos, onde é fornecido ao especialista um rótulo contendo as características mais relevantes dos elementos de cada grupo. Essas características são compostas por intervalos de valores dos atributos dos dados, de modo que o problema da rotulação é definido como:

> ***Problema da Rotulação:*** *Seja X um conjunto vetores de atributos definidos no $\mathbb{R}^n$ e expressos por $\vec{x} = (x_1, ...x_n)$ tal que $X = \{x_{i,j}\}|_{i,j=1}^{i=m,j=n}$, particionados em grupos $G = \{g_l\}|_{l=1}^{k}$ tal que $g_l \in X$, $g_l \cap g_{l'} = \emptyset \ \forall \ 1 \leq l \leq k$ e $l \neq l'$; o objetivo consiste em apresentar um conjunto de rótulos $R = \{r_{g_l}\}|_{l=1}^{k}$ no qual cada rótulo específico é dado por um conjunto de pares formados por atributos e seus respectivos intervalos de valores, tal que $r_{g_l} = \{(atr_j, [p_j, q_j])\}|_{j=1}^{n^{(g_l)}}$ capaz de melhor expressar o grupo g_l associado.*

onde:

- k é o número de grupos;
- g_l é um grupo qualquer;
- n é a dimensão do problema;
- r_{g_l} é o rótulo do grupo g_l;
- atr_j é um atributo qualquer do problema;
- $[p_j, q_j]$ é o intervalo de valores do atributo atr_j, em que p_j é o limite inferior e q_j o limite superior;
- $n^{(g_l)}$ é o número de atributos no rótulo de g_l.

Para demostrar a rotulação de um agrupamento a Figura 3 apresenta um rótulo com três grupos g_1, g_2 e g_3. O rótulo do grupo g_1 é formado pelos atributos $atr_1 e atr_3$ e seus respectivos intervalos de valor. De mesmo modo, os grupos g_2 e g_3 têm seus rótulos compostos por três e um pares de atributos e intervalos associados: atr_1, atr_2 e atr_3; e atr_2, respectivamente. É importante destacar que os intervalos dos atributos podem ou não ser os mesmos nos diferentes grupos. Por exemplo, $[p_1, q_1]$ podem ter valores diferentes nos grupos g_1 e g_2, indicando apenas sua relação com o atributo atr_1. Entretanto, caso apresentem o mesmo valor, revelando a importância desse intervalo na definição de ambos os grupos, a diferenciação entre os rótulos é determinada pelos demais componentes, de modo que, para grupos diferentes os rótulos devem apresentar pelo menos um par atributo–intervalo diferente, tornando os rótulos específicos para cada grupo.

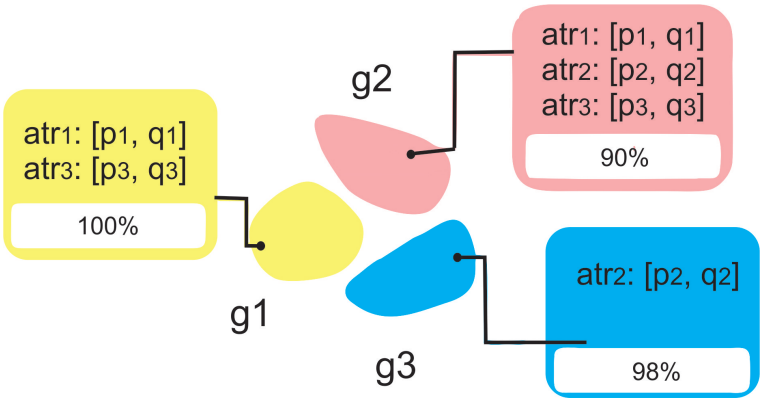

Figura 3 – Estrutura do rótulo

A Figura 3 apresenta o modelo de rotulação que este trabalho utiliza como base, com intenção de analisar se a seleção de atributos utilizada nesta abordagem será melhor que os demais. Ao final do processo de rotulação, tem-se a identificação dos elementos dos grupos de forma resumida, de modo que é possível formar uma sentença descritiva sobre os grupos com base em suas principais características (SILVA, 2021).

3 Trabalhos Relacionados

Este capítulo aborda os principais modelos de rotulação automática de grupos da literatura, apresentando suas metodologias, resultados obtidos e limitações, servindo como complemento teórico para os pesquisadores da área.

3.1 Filtros para Seleção de Atributos

O trabalho de Castro et al. (2004) aborda a seleção de subconjuntos de atributos por meio das abordagens *wrapper* e filtro. As técnicas para seleção de subconjuntos de atributos no contexto de aprendizado de máquina visam identificar os atributos que efetivamente auxiliam para a caracterização da classe de uma instância. Este trabalho têm como objetivo investigar os métodos e técnicas que visam reduzir o conjunto de atributos no contexto de aprendizado indutivo de máquina utilizando as abordagens de seleção de atributos do tipo *wrapper* e filtro. Para a abordagem filtro Castro et al. (2004) utiliza os algoritmos ReliefF e Focus, onde é possível verificar o baixo custo computacional do método RelifF e a complexidade da busca completa executada pelo método Focus.

3.2 Abordagens de Rotulação de Grupos

O trabalho Lopes et al. (2016) utiliza o algoritmo não supervisionado k-means para formar grupos e os métodos de discretização, EWD (Discretização por larguras iguais) e o EFD (Discretização por frequências iguais) para discretizar os dados em faixas de valores. Para formar os rótulos utiliza-se redes neurais artificiais do tipo PMC (Perceptron de múltiplas camadas) para selecionar os atributos mais relevantes, dado todos os valores dos outros atributos. Para cada grupo gerado são criadas m RNAs, onde m é a quantidade de características que descreve um elemento do problema. Cada RNA criado é responsável por avaliar a relevância de um atributo como um potencial candidato ao rótulo de seu grupo. Assim cada RNA, por sua vez, apresenta uma taxa de acerto em relação ao seu aprendizado, realizado apenas com os elementos de seus respectivos grupos (LOPES et al., 2016). Para dar uma maior precisão e confiança às taxas de acerto, foi aplicado o método de *holdout (random subsampling)*, onde o processo é repetido de modo iterativo por M vezes. Assim, para cada RNA é calculado a média de suas taxas de acerto obtidas em M iterações.

Outra abordagem relacionada à rotulação, agora por meio do algoritmo *Fuzzy C-means* é o trabalho de Machado, Ribeiro e Rabêlo (2015) é um modelo capaz de analisar os grupos e produzir definições conhecidas como rótulos. Na formação dos rótulos esse

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 (GP) em cada grupo. O modelo utiliza essa matriz para selecionar os elementos relevantes em cada grupo, esses elementos são utilizados para formular as faixas de valores dos rótulos. Cada faixa de valor é formada por meio de seleção dos valores máximo e mínimo de cada atributo, utilizando para isso os elementos selecionados como relevantes.

Já o trabalho de Araujo, Veras e Machado (2019) propõe um modelo de classificação baseado em lógica*Fuzzy* onde utiliza os rótulos gerados pelos modelos de rotulação automática para formação das regras e funções de pertinência. Este trabalho também utiliza Redes Neurais para selecionar os atributos mais relevantes de cada grupo. Para cada atributo presente nos rótulos uma função de pertinência é formada, empregando cada intervalo de valores para compor uma partição *Fuzzy*

Outra pesquisa que aborda a rotulação é a de Imperes Filho et al. (2020), utiliza o algoritmo *Fuzzy* como método para resolução do problema de rotulação, dividindo sua abordagem em duas etapas. Na primeira etapa 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 analisada recebe um GP em relação a cada grupo formado. Na segunda etapa Imperes Filho et al. (2020) tem como objetivo formular as faixas de valores para cada atributo em cada grupo formado verifica a existências de interseções entre faixas de valores e, por fim, monta e exibe os rótulos de cada grupo com faixas de valores que não possuem interseção. O modelo de rotulação proposto seleciona os elementos que possuem um grau de pertinência maior que o parâmetro grau de seleção. Ele define o valor inicial de 0.5 para o grau de seleção, sendo assim, em cada grupo selecionado são extraídos valores máximos de cada atributo. Esses valores correspondem às faixas de valores de cada grupo.

Em Moura (2020) é apresentado um método não supervisionado de rotulação de grupos que emprega o algoritmo de discretização CAIM *(Class-Attribute Interdependency Maximization)* a fim encontrar faixas de valores representativas nos atributos que serão relevantes para interpretação dos grupos. Do processo de discretização resultarão faixas de valores que serão analisadas e comparadas com os valores ocorrentes dos atributos em cada grupo, para determinar os atributos e faixas de valores representativos. O algoritmo de discretização CAIM é aplicado em aprendizagem supervisionada e utiliza a informação de interdependência entre classe e atributo como critério para a definição de uma quantidade ótima de pontos de corte.

O trabalho apresentado por Silva (2021) é dividido em duas etapas, a primeira etapa utiliza em sua abordagem para seleção de atributos a regressão para estimar a função que descreve o erro de predição dos atributos em seus respectivos domínios por grupo, permitindo delimitar intervalo em que o erro de predição é o menor possível

para cada atributo de cada grupo. È utilizado um conjunto de n modelos de regressão (MR), onde n é o número de características do problema. Cada modelo é treinado para predição de um atributo, utilizando os demais atributos como entrada. Já na segunda etapa são selecionados os pares atributo-intervalo obtidos na primeira etapa que representem a maioria dos elementos dos grupos, bem como sejam capazes de distingui-los, para composição dos rótulos. A Tabela 1 mostra a acurácia média dos métodos sobre rotulação de grupos citadas acima.

Tabela 1 – Modelos de rotulação da literatura.

Métodos	**Seleção de Atributos**	**Acurácia Média**
Lopes, 2016	Rede Neural Artificial	93%
Ribeiro, 2015	Fuzzy C-means com Grau de Pertinência	96.63%
Imperes, 2020	Fuzzy com Grau de Pertinência	94.28%
Moura, 2020	Rede Neural Artificial	97.39%
Silva, 2021	Regressões	98.01%

4 Modelo Proposto

Neste capítulo apresenta-se a proposta de modelo de rotulação automática de grupos baseada nos filtros de ganho de informação.

4.1 Modelo

Para contextualizar com um exemplo, inicialmente tem-se uma base de dados como entrada. Um algoritmo de aprendizagem não supervisionada foi introduzido em nossa metodologia e aplicado com o objetivo de formar vários grupos a partir dos elementos inicialmente fornecidos na base de dados.

Neste trabalho após a formação dos grupos, são utilizados quatro filtros diferentes de seleção de atributos para criar um *ranking* dos atributos mais relevantes da base de dados e logo após, somente os dados mais relevantes são discretizados para converter valores numéricos em valores discretos, assim, formando os intervalos que serão utilizados nos rótulos. Por fim, é apresentado como saída um rótulo específico que melhor define o grupo formado. A Figura 4 mostra as 4 etapas principais desta abordagem.

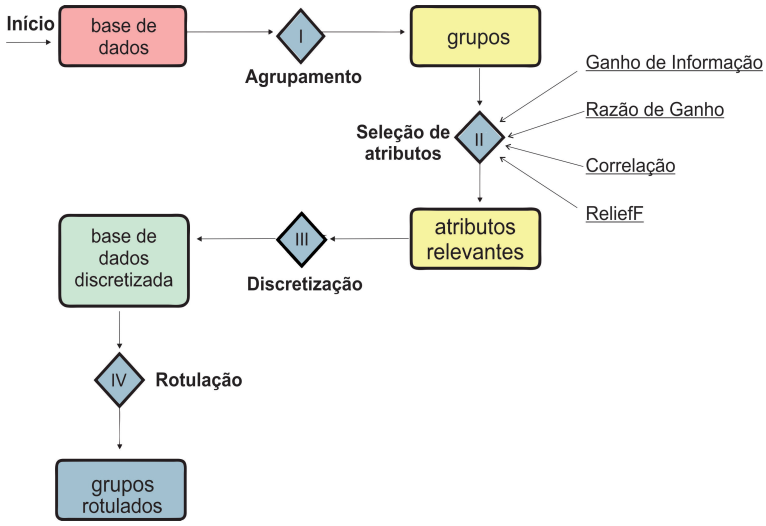

Figura 4 – Modelo Proposto

Para iniciar, essa abordagem recebe como parâmetro de entrada uma base de dados, que contém tipos de dados contínuos, havendo a necessidade de aplicação de um método de discretização. O passo (I) realiza o agrupamento dos dados com o algoritmo não supervisionado. O passo (II) é onde é aplicado os filtros de seleção de atributos para selecionar os dados mais relevantes. O passo (III) faz a discretização dos dados, auxiliando no processo de rotulação. E o último passo (IV) é uma estratégia que seleciona um intervalo de valor para cada atributo relevante selecionado é aplicada de forma a gerar rótulos.

4.1.1 Agrupamento(I)

Para facilitar a compreensão do leitor, a descrição do funcionamento do modelo de rotulação é feita com a sua aplicação em uma base de dados fictícia.

Foi criada uma base denominada Base de Dados Modelo (BDM), para facilitar o entendimento dessa abordagem. A primeira coluna contém um índice responsável por identificar cada elemento de forma única. As demais colunas representam uma característica – ou atributo – do respectivo elemento. Dessa forma, um elemento qualquer pode ser expresso como um vetor de dimensão m, onde m é a quantidade de características existentes na base de dados que o descrevem. Portanto, cada linha da tabela representa um elemento e cada coluna uma característica – exceto a primeira, que representa o índice. O elemento 3, por exemplo, pode ser descrito por $atr_1 = 2.00$, $atr_2 = 108.36$ e $atr_3 = 22.68$; ou seja, pelo vetor $\vec{e_3} = (2.00, 108.36, 22.68)$

Para realizar o agrupamento foi utilizado o algoritmo de aprendizado não supervisionado, *k-means*. O processo de agrupamento foi feito com o valor de $K = 3$, ou seja, o algoritmo dividiu a base em três grupos. Assim, a Tabela 2 mostra a base de dados modelo (BDM) após o processo de agrupamento como saída à associação de cada elemento a um respectivo grupo criado, a coluna grupo indica o grupo a qual um determinado elemento foi associado.

Como forma de auxiliar na compreensão, a Tabela 2 mostra que todos os elementos de *id* entre o 1 e 15 pertencem ao grupo 3, já os elementos entre 16 e 32 pertencem ao grupo 2 e o restante dos elementos fazem parte do grupo 1.

Após aplicar o algoritmo *k-means* e agrupar os elementos em seus respectivos grupos, esses elementos passarão pelo processo de seleção de atributos, onde ao aplicar os filtros de ganho de informação resultará em uma lista de classificação ordenada pela relevância dos elementos.

Tabela 2 – Base de dados modelo após agrupamento

id	atr_1	atr_2	atr_3	**grupo**
1	2.08	92.11	22.07	3
2	1.26	85.03	20.45	3
3	2.0	108.36	22.68	3
4	1.82	100.2	23.09	3
5	1.43	77.59	21.8	3
6	1.14	107.77	18.99	3
7	1.97	98.0	22.32	3
8	1.33	82.01	19.82	3
9	1.66	103.93	21.1	3
10	1.87	88.36	22.45	3
11	1.11	107.82	19.32	3
12	1.85	82.65	20.35	3
13	1.04	102.62	19.46	3
14	1.97	100.37	21.94	3
15	1.52	93.13	20.61	3
16	1.74	43.78	18.72	2
17	1.53	44.01	20.98	2
18	1.5	39.67	21.78	2
19	1.74	55.86	20.31	2
20	1.8	65.72	19.62	2
21	1.42	66.14	21.61	2
22	2.08	67.66	20.74	2
23	1.95	45.7	22.1	2
24	1.77	50.04	20.16	2
25	1.42	53.51	19.64	2

id	atr_1	atr_2	atr_3	**grupo**
26	1.12	62.71	19.07	2
27	2.09	60.58	20.2	2
28	1.95	69.23	19.68	2
29	1.03	47.81	19.47	2
30	1.72	42.35	22.89	2
31	1.53	41.16	22.67	2
32	1.14	61.59	19.9	2
33	1.08	91.93	20.81	1
34	1.62	79.21	18.43	1
35	1.68	80.87	18.42	1
36	1.81	98.24	22.13	1
37	1.3	69.27	18.83	1
38	1.8	101.21	21.61	1
39	1.79	72.02	22.02	1
40	1.56	81.71	22.1	1
41	1.98	77.16	21.71	1
42	1.86	89.12	22.84	1
43	1.55	76.01	19.74	1
44	1.97	81.57	19.83	1
45	1.75	90.92	21.39	1
46	1.47	101.77	19.2	1
47	1.44	93.61	21.03	1
48	1.51	98.65	19.24	1
49	1.06	68.82	21.68	1
50	1.48	80.4	21.43	1

4.1.2 Seleção de Atributos (II)

A seleção de atributos permite, por exemplo, a ordenação de atributos segundo algum critério de importância, a redução da dimensionalidade do espaço de busca de atributos e a remoção de dados contendo ruídos, entre outros. A seleção de atributos pode ser relevante em casos nos quais a medição de certos atributos é custosa, pois pode permitir que um subconjunto, representativo e menor que original, seja selecionado. Como resultado da realização de seleção de atributos, será possível compreender os grupos, por meio dos rótulos, com um menor número de atributos possível.

A etapa II possui como entrada um conjunto de grupos e apresenta como saída um conjunto de atributos relevantes para cada grupo gerado que será utilizado em sua rotulação. Nesta etapa também ocorre a detecção dos atributos mais relevantes para a definição de um grupo qualquer. Portanto, para isso, serão utilizados os filtros ganho de informação que tem a capacidade de detectar relação entre variáveis. Neste contexto do

processo de rotulação a seleção de atributos vai detectar os atributos relevantes de cada grupo.

Os quatro filtros de ganho de informação utilizam o atributo *grupo* como parâmetro. Neste caso o grupo formado é usado como atributo classe, apesar de cada filtro ter seus parâmetros, a alteração deles não gerou classificação diferente por conta da base de dados modelo ser uma base com poucos atributos e elementos. Foi aplicado na base de dados modelo (BDM) os filtros de ganho de informação com a intenção de gerar um *ranking* dos atributos de acordo com sua relevância. A Tabela 3 mostra essa classificação, devido a BDM ter poucos elementos essa classificação não mudou de um filtro para outro, apenas o valor do *ranking*.

Alguns filtros, se não todos, necessitam de um atributo classe para realizar o *ranking*. Neste caso, o atributo classe utilizado foi o atributo resultante da aplicação do *k-means*, o atributo responsável por dividir os elementos em k grupos, ele é um atributo nominal, que é uma das exigências para ser um atributo classe.

Tabela 3 – Ranking dos atributos da Base de Dados Modelo.

Ganho de Informação		**Razão de Ganho**		**Correlação**		**ReliefF**	
Atr	**Relev**	**Atr**	**Relev**	**Atr**	**Relev**	**Atr**	**Relev**
atr_2	0.882	atr_2	0.831	atr_2	0.5583	atr_2	0.22566
atr_3	0	atr_3	0	atr_3	0.1032	atr_3	-0.00919
atr_1	0	atr_1	0	atr_1	0.0218	atr_1	-0.01487

4.1.3 Discretização (III)

A terceira etapa corresponde ao processo de discretização dos dados, a seleção de atributo cria um *rankig*, permitindo assim, discretizar apenas os dados dos atributos mais relevantes de cada base de dados. As bases de dados para a aplicação da discretização necessariamente tem de possuir dados não categóricos, ou seja, dados numéricos. Todas as bases testadas neste trabalho são bases com dados não categóricos.

A discretização dos dados é onde converte valores numéricos em valores discretos podendo apresentar uma variedade de valores dentro de um determinado domínio. Além disso, a discretização forma os intervalos que são utilizados no rótulo. Para demonstrar o processo, será utilizada a técnica de discretização por larguras iguais (EWD), selecionando os atributos atr_1, atr_2, atr_3, considerando a faixa de valor $R = 3$ para a discretização.

Os valores iniciais são convertidos em valores discretos. Os elementos situados entre o valor mínimo e o primeiro ponto de corte passarão a ser representados pelo valor discreto 1; os valores situados entre os dois pontos de cortes pelo valor 2; e os valores entre o segundo ponto de corte e o valor máximo, pelo valor 3.

Na técnica de discretização pode haver perda de informação, mas por outro lado, ela permite o algoritmo lidar com faixa de valores caracterizados com valores discretos. As faixas de valores discretos auxiliam na compreensão dos grupos e tem o intuito de apresentar melhores resultados em relação à velocidade e precisão, além de auxiliar na formação dos rótulos.

A Figura 5 mostra essa conversão dos valores numéricos em valores discretos aplicando o método de discretização por larguras iguais na base de dados modelo.

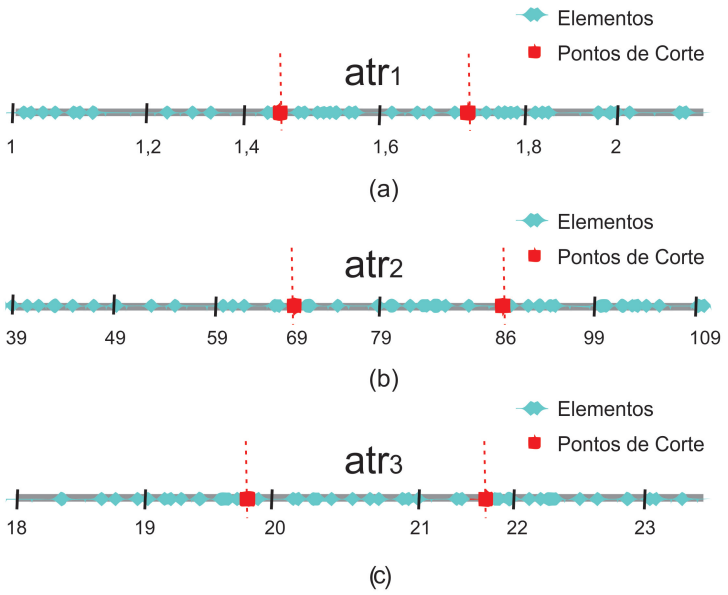

Figura 5 – Discretização dos atributos utilizando o método EWD com R=3

Como resultado obtém-se uma nova base de dados, mesma denominação utilizada por Lopes et al. (2016), Base de Dados Modelo Discretizada (BDMD), análoga apresentada na Tabela 4. Esses novos valores obtidos com a discretização serão utilizados para auxiliar a rotulação dos dados.

4.1.3.1 Cálculo da Variação V

A variação V é o cálculo feito por meio da porcentagem (%) em cima dos valores de *ranking* gerados pelos filtros de seleção de atributos. Esse cálculo é utilizado para eliminar uma possível ambiguidade entre grupos selecionando todos os atributos que possuem até uma diferença V em relação ao atributo com maior taxa de acerto e descartar os demais.

A variação V é considerada importante porque ela atua como limitador para evitar que todos os atributos apareçam nos rótulos. O valor de V pode variar em cada problema

Tabela 4 – Base de Dados Modelo Discretizada

id	atr_1	atr_2	atr_3	**grupo**
1	3	3	2	3
2	2	2	2	3
3	3	3	3	3
4	3	3	3	3
5	2	2	2	3
6	1	3	1	3
7	3	3	2	3
8	1	2	1	3
9	2	3	2	3
10	3	3	2	3
11	1	3	1	3
12	3	2	2	3
13	1	3	3	3
14	3	3	2	3
15	2	3	2	3
16	2	1	1	2
17	2	1	2	2
18	2	1	2	2
19	2	1	2	2
20	3	2	1	2
21	2	2	2	2
22	3	2	2	2
23	3	1	2	2
24	3	1	2	2
25	2	1	1	2

id	atr_1	atr_2	atr_3	**grupo**
26	1	2	1	2
27	3	2	2	2
28	3	2	1	2
29	1	1	1	2
30	2	1	3	2
31	2	1	3	2
32	1	1	1	2
33	1	3	3	1
34	2	2	1	1
35	2	2	1	1
36	3	3	2	1
37	1	2	1	1
38	3	3	2	1
39	3	2	2	1
40	2	2	2	1
41	3	2	2	1
42	3	3	1	1
43	2	2	1	1
44	3	2	1	1
45	3	3	2	1
46	2	3	1	1
47	2	3	2	1
48	2	3	1	1
49	1	2	2	1
50	2	2	2	1

e deve ser ajustado visando sempre a assumir o menor valor possível de modo a distinguir os grupos apresentados.

Nesta abordagem utilizamos o valor da variação V variando entre 5% e 82%. Esses valores foram escolhidos de acordo com os testes realizados em conformidade com as bases de dados e os métodos de seleção utilizados. Para demonstrar o cálculo da variação V utilizamos o valor de *ranking* do atributo mais relevante de cada filtro de seleção da Tabela 3.

Uma porcentagem é escolhida de acordo com a quantidade de atributos que deseja selecionar, essa porcentagem é multiplicada pelo valor do *ranking* do atributo mais relevante, ou seja, do primeiro colocado na lista do filtro de seleção. O valor final da variação é a diferença do valor de *ranking* pela multiplicação citada anteriormente. Para demostrar esse cálculo utilizamos os filtros de correlação e reliefF conforme a Tabela 5.

Tabela 5 – Cálculo da Variação V.

Correlação			**ReliefF**		
Porc(%)	**Resultado**	**Atributo**	**Porc(%)**	**Resultado**	**Atributo**
50%	0.2915	atr_2	50%	0.1128	atr_2
85%	0.0874	atr_2, atr_3	80%	0.0451	atr_2
98%	0.0116	atr_2, atr_3, atr_1	98%	0.0225	atr_2

Os atributos selecionados na Tabela 5 vão compor os rótulos com suas respectivas faixa de valores selecionadas na discretização.

Para auxiliar nos resultados da aplicação dos filtros de seleção de atributos, é feito uma média aritmética para cada teste variando o valor de V. Esta varição permite adicionar ou remover atributos para compor os rótulos, sendo assim, a média aritmética é feita a partir da variação de V, auxiliando na comparação da eficiência do método de seleção para cada base de dados.

4.1.4 Rotulação (IV)

A rotulação é a última etapa dessa abordagem, ela sintetiza as características comuns dos elementos dos grupos em um rótulo, facilitando a compreensão e a interpretação sobre os dados por meio dos valores apresentados. Para auxiliar na compreensão utiliza-se a base de dados modelo discretizada apresentada na Tabela 4.

Analisando grupo 1 em relação ao atributo atr_1 o valor de maior ocorrência é 2, para o atributo atr_2 é 2 e para o atributo atr_3 o valor de maior ocorrência também é 2. Comparando esses valores com a Figura 5. Os intervalos seriam definidos como $atr_1 = [1.38 - 1.74]$, $atr_2 =]62.56 - 85.46]$ e $atr_3 =]19.8 - 21.6]$. Portanto como exemplo, podemos considerar os atributos selecionados do rótulo r_{c1} . Assim, os limites dos intervalos dos atributos atr_1, atr_2 e atr_3 são respectivamente $p_1 = 1.38$, $q_1 = 1.74]$, $]p_2 = 62.56$, $q_2 = 85.46]$ e $]p_3 = 19.8$, $q_3 = 21.6]$ pois representam os limites das faixas de valores obtidas na discretização para os valores de maior frequência. Caso não houvesse discretização em um dos atributos selecionados o valor de maior ocorrência representaria o atributo.

No caso da base de dados modelo em relação ao cálculo da variação V apresentada na Tabela 5, nota-se que ao utilizar o filtro correlação de atributos e aplicar diferentes porcentagens de V resultou em grupos com diferentes atributos. Já ao utilizar o filtro reliefF mesmo alterando o valor de V, o único atributo que compõe os grupos é o atr_2. Portanto utilizando o filtro reliefF resultou no mesmo atributo variando as faixas de valores para cada rótulo.

- $r_{c1} = (atr_2, 62.56 \backsim 85.46)$;
- $r_{c2} = (atr_2, 39.67 \backsim 62.56)$;

- $r_{c3} = (atr_2, 85.46 \backsim 108.36)$.

Com o uso do filtro de correlação essa rotulação foi diferente, os três atributos foram classificados para compor os rótulos. Observa-se a repetição de algumas faixas para o atributo atr_3 nos três rótulos e para o atributo atr_1 para os rótulos 1 e 2.

- $r_{c1} = (atr_2, 62.56 \backsim 85.46), (atr_3, 19.98 \backsim 22.53), (atr_1, 1.38 \backsim 1.74)$;
- $r_{c2} = (atr_2, 39.67 \backsim 62.56), (atr_3, 19.98 \backsim 22.53), (atr_1, 1.38 \backsim 1.74)$;
- $r_{c3} = (atr_2, 85.46 \backsim 108.36), (atr_3, 19.98 \backsim 22.53), (atr_1, 1.74 \backsim 2.09)$.

Com a intenção de simplificar, os rótulos, eles podem ser escritos da seguinte maneira: $r_{c1} = [1.38 \backsim 1.74]$ e $atr_2[62.56 \backsim 85.46]$. Sendo assim, a compreensão é que os elementos do grupo c_1 possuem seu atributo atr_1 variando de 1.38 a 1.74 e seu atributo atr_2 variando de 62.56 a 85.46.

Algoritmo 1 Modelo Proposto

```
1: Carrega a base de dados;
2: Executar o algoritmo não supervisionado (clustering);
3: para cluster ← 1 até k faça
4:     para iteração ← 1 até M faça
5:         para atributo ← 1 até N faça
6:             utilizar seleção de atributos;
7:             calcular a relevância;
8:         fim para
9:     fim para
10:    Selecionar os atributos mais relevantes;
11:    Aplicar variação V;
12:    Associar os valores aos intervalos;
13:    Calcular média das taxas de acerto;
14: fim para
15: Exibir os rótulos;
```

Para resumir, o Algoritmo 1 apresenta essa proposta de rotulação em pseudocódigo, onde é resumido os passos para realizar a rotulação dos dados, iniciando com a formação dos grupos e depois com a aplicação da seleção de atributos.

5 Resultados e Discussão

Este capítulo descreve os experimentos realizados para avaliação e autenticação do modelo proposto. Nos experimentos, quatro bases de dados da literatura foram utilizadas: Íris (FISHER, 1987), Sementes (KULCZYCKI; CHARYTANOWICZ, 2011), Vinhos (AEBERHARD; COOMANS; VEL, 1994) e Vidros (EVETT; SPIEHLER, 1989), disponíveis no repositório UCI (BACHE; LICHMAN, 2013). A escolha das bases de dados deu-se pela presença em pelo menos um dos trabalhos relacionados, permitindo a comparação do desempenho do modelo proposto com a literatura. E também com o propósito de permitir a comparação entre os modelos de rotulação, as bases de dados foram agrupadas utilizando o algoritmo k-means, com k igual ao número de classes de cada problema, seguindo a abordagem utilizada nos modelos de Lopes et al. (2016), Imperes Filho et al. (2020), Moura (2020) e Silva (2021).

Apesar de algumas bases de dados já possuírem o atributo classe, este fator não influenciou ao realizar o agrupamento, porque ele independente do atributo classe.

Para cada base de dados foram aplicados quatro filtros de seleção de atributos, são eles, ganho de informação, razão de ganho, correlação e reliefF. Cada filtro cria um *ranking* de relevância dos atributos. O cálculo da variação V implica na quantidade de atributos selecionados para compor os rótulos. Após seleção dos atributos mais relevantes é realizado a discretização dos dados, onde converte os valores numéricos em valores discretos, auxiliando nas faixas de intervalo dos rótulos. Neste capítulo estão contidos somente os testes em que resultaram a melhor acurácia, os outros testes encontram-se no Apêndice A deste trabalho.

5.1 Íris

A base de dados refere-se à identificação de plantas, também encontrada no repositório de dados *UCI Machine Learning* (BACHE; LICHMAN, 2013) e apresentada em (FISHER, 1987). Esse conjunto de dados contempla três diferentes tipos de plantas com base nas características Comprimento da Sépala (SL), Comprimento da Pétala (PL), Largura da Pétala (PW) e Largura da Sépala (SW). onde cada amostra se refere a um tipo específico de planta, sendo no total 150 instâncias:

- 50 elementos do tipo Iris-setosa;
- 50 elementos do tipo Iris-versicolour;
- 50 elementos do tipo Iris-virginica.

A base foi agrupada pelo k-means utilizando $k = 3$, o método de discretização utilizado foi o EWD com a faixa $R = 3$. A Tabela 6 mostra o *ranking* dos atributos mais relevantes gerados após aplicação dos quatro métodos de seleção de atributo.

Tabela 6 – Ranking dos atributos da base Iris.

Ganho de Informação		**Razão de Ganho**		**Correlação**		**ReliefF**	
Atr	**Rank**	**Atr**	**Rank**	**Atr**	**Rank**	**Atr**	**Rank**
PL	1.625	PW	0.872	PL	0.534	PW	0.362
PW	1.379	PL	0.486	PW	0.513	PL	0.361
SL	0.879	SL	0.672	SW	0.405	SL	0.175
SW	0.4	SW	0.257	SL	0.39	SW	0.113

O atributo classe utilizado para gerar a relevância dos atributos foi o atributo grupo, ele foi gerado ao aplicar o *k-means*, ou seja, em todos os teste não foi utilizado o atributo classe da própria base. Os filtros ganho de informação e correlação foram os que apresentaram os mesmos atributos mais relevantes do *ranking*, comprimento da pétala (PL) e largura da pétala (PW). Onde o cálculo do ganho de informação é baseado na entropia e o cálculo de correlação é baseado na associação das variáveis em relação ao seus scores.

Já ao aplicar os filtros razão de ganho e reliefF, eles selecionam no *ranking* os mesmos dois atributos mais relevante, mas de forma invertida. O filtro razão de ganho calcula através do algoritmo C4.5, essa métrica cria um viés que favorece atributos caracterizados por um número menor de valores distintos. Já o reliefF seleciona as variáveis de forma a separar as instâncias de diferentes classes.

O filtro correlação de atributo junto com a variação V=5% foi o filtro que teve a maior taxa média de acerto 98.63% dentre os quatro testados, onde foi selecionado somente o atributo (PL) para compor os rótulos, cada um com sua respectiva faixa de valor. A Tabela 7 mostra que todos os elementos dos grupos 1 e 2 obedeceram ao rótulo gerado.

Tabela 7 – Filtro correlação aplicado a base Iris.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	50	PL	[1,0 - 2,9]	0	100
2	51	PL	[2,9 - 4,9]	0	100
3	49	PL	[4.9 - 6,9]	3	95.9

Abaixo é apresentado o rótulo gerado para a base íris com a aplicação do filtro correlação e a variação V=5%.

- r_{c1} = {(PL, 1.0 - 2.9)};

- r_{c2} = {(PL, 2.9 - 4.9)};
- r_{c3} = {(PL, 4.9 - 5.9)}.

5.2 Wine

A base de dados vinhos resulta da análise química de vinhos cultivados na mesma região da Itália, mas derivados de três cultivos diferentes. A análise baseia-se nas quantidades de 13 componentes encontrados em cada um dos três tipos de vinhos:

- Álcohol;
- Malic.acid;
- Ash;
- Acl;
- Mg;
- Phenols;
- Flavanoids;
- Nonflavanoids.phenols;
- Proanth;
- Color.int;
- Hue;
- Od - OD280/OD315 *of diluted wines*;
- Proline.

Os 173 dados foram agrupados pelo k-means utilizando $k = 4$, o método de discretização utilizado foi o EWD com a faixa $R = 3$. A Tabela 8 mostra o *ranking* da relevância dos atributos gerados após aplicação dos quatro métodos de seleção de atributo.

Já no caso da base wine o atributo *proline* é o mais relevante para todos os filtros de seleção, seguido do atributo *alcohol* com exceção para o filtro ganho de informação que apresenta como segundo mais relevante o atributo *flavanoids*.

Tabela 8 – Ranking dos atributos da base Wine.

Ganho de Informação		**Razão de Ganho**		**Correlação**		**ReliefF**	
Atr	Rank	Atr	Rank	Atr	Rank	Atr	Rank
Proline	2.208	Proline	1	Proline	0.435	Proline	0.2462
Flavanoids	0.504	Alcohol	0.363	Alcohol	0.311	Alcohol	0.0691
Alcohol	0.358	Phenols	0.345	Flavanoids	0.261	Flavanoids	0.0636
Phenols	0.345	Mg	0.34	Phenols	0.249	Phenols	0.0502
Mg	0.284	Flavanoids	0.325	Mg	0.246	Od	0.0493
Malic.acid	0.28	Acl	0.271	Acl	0.21	Color.int	0.0426
Od	0.255	Od	0.264	Od	0.196	Hue	0.0335
Acl	0.228	Color.int	0.241	Color.int	0.176	Acl	0.0297
Color.int	0.206	Hue	0.22	Proanth	0.172	Malic.acid	0.0296
Hue	0.195	Malic.acid	0.19	Nonflavanoid phenols	0.172	Nonflavanoid phenols	0.0267
Nonflavanoid phenols	0.155	Proanth	0.185	Hue	0.138	Mg	0.0264
Proanth	0.149	Nonflavanoid phenols	0.155	Ash	0.111	Ash	0.0233
Ash	0	Ash	0	Malic.acid	0.106	Proanth	0.0133

O filtro ganho de informação junto com a variação V=60% resultou em uma taxa média de acerto de 91.66%, o melhor resultado para esta base em comparação com os outros filtros testados. Nesse teste foi selecionado somente o atributo *proline* variando suas faixas de valores por grupo. O atributo *proline* teve todos os elementos obedecendo ao rótulo em todos os grupos com exceção do grupo 3, onde somente 38 elementos obedeceram ao rótulo gerado e resultou 66.66% de taxa de acerto para o grupo 3, como mostra a Tabela 9.

Tabela 9 – Filtro ganho de informação aplicado a base Wine.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	66	Proline	[278 - 721]	0	100
2	23	Proline	[1164 - 1608]	0	100
3	57	Proline	[278 - 721]	19	66.66
4	32	Proline	[721 - 1164]	0	100

O rótulo formado com a aplicação do filtro ganho de informação junto com a variação V=60% na base wine é:

- r_{c1} = {(Proline, 278 - 721) };
- r_{c2} = {(Proline, 1164 - 1608) };
- r_{c3} = {(Proline, 278 - 721) };

- r_{c4} = {(Proline, 721 - 1164) }.

5.3 Seeds

A base de dados seeds se refere a identificação de sementes de trigo e descreve 210 amostras de 3 tipos de sementes de trigo:

- 70 do tipo *Kama*;
- 70 do tipo *Rosa*;
- 70 do tipo *Canadian*.

As amostras são descritas por 7 características geométricas: Área (A), Perímetro (P), Densidade (D), Comprimento da Semente (CSM), Largura da Semente (LS), Coeficiente de Assimetria (CA) e Comprimento do Sulco da Semente (CSS). Conforme sugere o número de classes, a base foi agrupada em 4 grupos e discretizada pelo método EWD em R=3.

Ao aplicar os filtros de seleção de atributos na base seeds é notado uma semelhança de *ranking* entre os filtros ganho de informação e reliefF. A Tabela 10 mostra o *ranking* dos atributos da base seeds para cada filtro de seleção. O atributo (A) é o mais relevante ao aplicar os filtros ganho de informação e reliefF seguindo do atributo (P). Já para o filtro razão de ganho, o atributo (P) é considerado o mais relevante e para o filtro correlação o atributo mais relevante é o (LKG).

Tabela 10 – Ranking dos atributos da base Seeds.

Ganho de Informação		**Razão de Ganho**		**Correlação**		**ReliefF**	
Atr	**Rank**	**Atr**	**Rank**	**Atr**	**Rank**	**Atr**	**Rank**
A	1.3657	P	0.783	LKG	0.539	A	0.2473
P	1.2355	A	0.734	A	0.537	P	0.2436
WK	1.1667	LKG	0.626	P	0.537	WK	0.2176
LK	1.0991	WK	0.509	LK	0.528	LK	0.2023
LKG	0.7608	LKG	0.504	WK	0.518	LKG	0.1717
C	0.2957	C	0.327	C	0.42	C	0.0908
AC	0.1936	AC	0.209	AC	0.359	AC	0.0892

Os atributos (C) e (AC) são considerados os de menor relevância da base, ou seja, apesar de cada filtro realizar diferentes cálculos para encontrar a relevância dos atributos, todos eles tiveram os mesmos resultados para esses dois atributos que são os menos relevantes da lista.

A Tabela 11 mostra o resultado ao aplicar o filtro reliefF com a variação V=15%, foram selecionados os atributos (A), (P) e (WK). Este filtro apresentou melhor taxa média de acerto 87.59% em relação aos outros para a base seeds.

Tabela 11 – Filtro reliefF aplicado a base Seeds.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	69	A	[10.59 - 14.12]	1	98.55
		P	[12.41 - 14.02]	1	98.55
		WK	[2.63 - 3.1]	4	94.20
2	48	A	[17.65 - 21.18]	0	100
		P	[15.63 - 17.25]	0	100
		WK	[3.57 - 4.03]	3	93.75
3	62	A	[10.59 - 14.12]	30	48.38
		P	[14.02 - 15.63]	15	75.80
		WK	[3.1 - 3.57]	10	83.87
4	31	A	[14.12 - 17.65]	0	100
		P	[14.02 - 15.63]	7	77.41
		WK	[3.1 - 3.57]	6	80.64

O rótulo formado com a aplicação do filtro reliefF e a variação V=15% na base seeds é:

- r_{c1} = {(A, 10.59 - 14.12), (P, 12.41 - 14.02), (WK, 2.63 - 3.1) };
- r_{c2} = {(A, 17.59 - 21.18),(P, 15.63 - 17.25), (WK, 3.57 - 4.03) };
- r_{c3} = {(A, 10.59 - 14.12),(P, 14.02 - 15.63), (WK, 3.1 - 3.57) };
- r_{c4} = {(A, 14.12 - 17.65),(P, 14.02 - 15.63), (WK, 3.1 - 3.57) }.

5.4 Glass

A base de dados vidros descreve 214 amostras de 6 tipos de vidros com base no nível de 10 componentes químicos : RI, Na, Mg, Al, Si, K, Ca, Ba e Fe. Os tipos de vidros podem ser agrupados em diferentes grupos:

- 70 elementos de janelas de construção – vidro processado;
- 76 elementos de janelas de construção – vidro não processado;
- 17 elementos de janelas de veículos – vidro processado;
- 13 elementos de recipientes;
- 9 elementos de utensílios de cozinha;
- 29 elementos de faróis.

O agrupamento da base foi feito com *k-means* onde $K = 6$ e o método de discretização utilizado foi o EWD com $R = 4$ para todos os atributos; a variação V variou para cada filtro de seleção de atributo.

Ao aplicar a discretização R=4, foi gerado somente quatro faixas de valores. Sendo assim, a quantidade de faixas é menor que a de grupos e com isso ocorre repetições de faixas nos grupos. Um dos objetivos desta abordagem é comparar os resultados com outros trabalhos, isso justifica a escolha do valor de R=4.

A Tabela 12 mostra os atributos em ordem de relevância para cada filtro de seleção de atributo. O atributo mais relevante da base é o (Mg) para todos os filtros, seguindo do atributo (Ca) para os filtros ganho de informação e reliefF, e o atributo (Ba) para os filtros razão de ganho e correlação. Já o atributo (Fe) é o que tem menor relevância para todos os filtros de seleção.

Tabela 12 – Ranking dos atributos da base Glass.

Ganho de Informação		**Razão de Ganho**		**Correlação**		**ReliefF**	
Atr	Rank	Atr	Rank	Atr	Rank	Atr	Rank
Mg	0.777	Mg	0.902	Mg	0.4511	Mg	0.2709
Ca	0.748	Ba	0.636	Ba	0.2793	Ca	0.0885
K	0.632	Ca	0.472	Al	0.255	Al	0.0641
Ri	0.519	Na	0.459	Na	0.2269	Ri	0.0614
Na	0.499	K	0.386	Ri	0.2083	Na	0.057
Al	0.495	Ri	0.364	Ca	0.1955	Ba	0.0426
Ba	0.364	Al	0.346	K	0.1077	Si	0.041
Si	0.255	Si	0.303	Si	0.0913	K	0.04
Fe	0	Fe	0	Fe	0.0785	Fe	0.0154

Ao aplicar o filtro ganho de informação com a variação V=20% resultou uma taxa média de acerto 81.73%. A Tabela 13 mostra esse resultado. Foram selecionados os atributos (Mg), (Ca) e (K) para compor os grupos. O atributo (K) apresentou em todos os grupos a mesma faixa de valor e com todos os elementos obedecendo ao rótulo, com exceção somente do grupo 2, em que apresentou outra faixa com 3 elementos de um total de 5 que não obedeceram ao rótulo.

Tabela 13 – Filtro ganho de informação aplicado a base Glass.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	75	Mg	[3.36 - 4.49]	12	84
		Ca	[8.12 - 10.81]	7	90.66
		K	[0 - 1.55]	0	100
2	5	Mg	[2.24 - 3.36]	2	60
		Ca	[5.43 - 8.12]	0	100
		K	[4.65 - 6.21]	3	40
3	33	Mg	[0 - 1.12]	11	66.66
		Ca	[8.12 - 10.81]	4	87.87
		K	[0 - 1.55]	0	100
4	16	Mg	[0 - 1.12]	5	68.75
		Ca	[10.81 - 13.5]	5	68.75
		K	[0 - 1.55]	0	100
5	7	Mg	[0 - 1.12]	0	100
		Ca	[10.81 - 13.5]	3	57.14
		K	[0 - 1.55]	0	100
6	78	Mg	[3.36 - 4.49]	13	83.33
		Ca	[5.43 - 8.12]	28	64.10
		K	[0 - 1.55]	0	100

O rótulo formado com a aplicação do filtro ganho de informação e a variação V=20% na base glass é:

- r_{c1} = {(Mg, 3.36 - 4.49), (Ca, 8.12 - 10.81), (K, 0 - 1.55) };
- r_{c2} = {(Mg, 2.24 - 3.36), (Ca, 5.43 - 8.12), (K, 4.65 - 6.21)};
- r_{c3} = {(Mg, 0 - 1.12), (Ca, 8.12 - 10.81), (K, 0 - 1.55) };
- r_{c4} = {(Mg, 0 - 1.12), (Ca, 10.81 - 13.5), (K, 0 - 1.55)};
- r_{c5} = {(Mg, 0 - 1.12), (Ca, 10.81 - 13.5), (K, 0 - 1.55)};
- r_{c6} = {(Mg, 3.36 - 4.49), (Ca, 5.43 - 8.12), (K, 0 - 1.55)}.

O modelo idealizado neste trabalho foi testado com outros valores de variação V e os resultados podem ser verificados no Apêndice A deste trabalho.

5.5 Avaliação de Performance do Rotulador

O objetivo é realizar comparações da rotulação de dados deste trabalho com outras abordagens, a Tabela 14 apresenta acurácia média dos métodos de Lopes et al. (2016), Machado, Ribeiro e Rabêlo (2015), Araújo (2018), Moura (2020) e Silva (2021).

A Tabela 14 mostra a acurácia média de algumas abordagens. Nota-se que para a base íris este modelo de rotulação proposto tem o melhor resultado de acurácia com 98.63%, esse resultado é da aplicação do filtro correlação junto com a variação V=5%, onde é selecionado somente o atributo PL para compor os rótulos.

Tabela 14 – Comparativo da Acurácia entre as Abordagens de Rotulação.

Modelo	**Técnicas utilizadas**	**Base Íris**	**Base Wine**	**Base Seeds**	**Base Glass**
Lopes, 2016	Discretização e Classificação com RNA	94.14%	X	89%	95.54%
Ribeiro, 2015	Lógica Fuzzy com Grau de Pertinência	98%	X	89%	98.43%
Moura, 2020	Discretização com CAIM	89.47%	100%	97.19%	97.93
Imperes, 2020	Distância Euclidiana com Grau de Seleção	92.52%	X	92.86%	99.08%
Silva, 2021	Regressão com RBF	94.33%	100%	89.33	88.5%
Método Proposto	Filtro de Seleção de Atributos	98.63%	91.66%	87.59%	81.73%

Já para a base wine esta rotulação de dados teve uma acurácia um pouco inferior em relação as abordagens de Silva (2021) e Moura (2020). No caso do modelo proposto, a base wine foi agrupada com k=4, já no trabalho feito por Silva (2021) com a utilização de regressão, esse agrupamento foi de k=6 e Moura (2020) utilizando *CAIM* agrupou com k=3.

Dentre os testes feitos com a base seeds para esta abordagem, o que obteve melhor acurácia 87.59% foi utilizando o filtro reliefF com a variação V=15%. A base seeds foi agrupada com k=4 baseando-se nos teste feito por Lopes et al. (2016). Todos os outros modelos de rotulação tiveram a acurácia superior a deste modelo de rotulação.

A base glass foi a última base a ser aplicada a rotulação, o agrupamento foi k=6 e o teste onde obteve-se uma melhor acurácia 81.73% foi com o filtro ganho de informação e uma variação v=20%. Dentre os modelos de rotulação comparados, este foi o que obteve menor acurácia.

Além dos modelos de rotulação formarem os rótulos, o agrupamento e a variação V tem influência na rotulação desses dados, para cada filtro de seleção de atributos foi selecionado um valor de k e de variação V na tentativa de selecionar melhor os dados que obedeçam aos rótulos.

A acurácia inferior em relação aos outros métodos, também se dá pela relação da faixa de valores, sendo que essa relação não é o objetivo deste método proposto. Métodos como o de Silva (2021) e Moura (2020) fazem a relação dessas faixas automaticamente, o que pode favorecer diretamente esses modelos.

6 Conclusões e Trabalhos Futuros

Neste capítulo, são apresentadas as conclusões do método de rotulação com filtro de seleção de atributos. E algumas sugestões para trabalhos futuros.

6.1 Conclusões

A interpretação dos dados de um agrupamento, usualmente, depende da análise de um especialista a fim de identificar padrões nos dados que expliquem a formação dos grupos, um processo que despende tempo e recursos. Em vista disso, os Modelos de Rotulação de Grupos são ferramentas que fornecem auxílio ao especialista na interpretação do agrupamento por meio de uma definição resumida das principais características dos grupos (SILVA, 2021).

Neste trabalho, foi proposto um modelo de classificação com a utilização de seleção de atributos baseado no processo de rotulação de grupos. O processo de rotulação é feito de acordo com os grupos fornecidos, sendo assim, o algoritmo de agrupamento tem grande influência nos rótulos gerados. Portanto para a formação dos grupos, este trabalho utiliza o algoritmo não supervisionado.

Inicialmente é aplicado um modelo de rotulação sobre os agrupamentos para extração dos rótulos a fim de auxiliar na interpretação dos grupos. Após a rotulação, os rótulos são utilizados para gerar regras e funções para os grupos, além de facilitar a interpretação do conhecimento adquirido.

O modelo proposto utiliza seleção de atributos para rotulação dos grupos. Inicialmente é usado o algoritmo *k-means* para realizar o agrupamento dos dados, após os dados divididos em grupos foi aplicado o método de discretização *EWD* para transformar os valores contínuos em faixa de valores, assim um valor discreto é associado a cada intervalo diferente de valores contínuos. Por fim, é aplicado os filtros de seleção de atributos junto com a variação V para selecionar os atributos mais relevantes de cada grupo.

O desempenho do modelo foi avaliado em quatro bases de dados da literatura: *Íris, Wine, Seeds e Glass*, também utilizadas em pelo menos um dos modelos de rotulação da literatura,Machado, Ribeiro e Rabêlo (2013), Imperes Filho et al. (2020), Moura (2020), Silva (2021), permitindo assim a comparação dos resultados.

Ao utilizar os filtros de seleção de atributos foi notado que os valores de *ranking* variam para cada base de dados, onde esses valores influenciam na aplicação da variação V. O cálculo da variação V auxilia na composição dos rótulos, esse cálculo gera um intervalo de valor onde são selecionados somente os atributos em que seu valor de *ranking* esteja contido nele.

A seleção de atributos mostrou que também é eficiente na formação de rótulos, cada filtro de seleção de atributos tem suas técnicas de cálculo para selecionar os atributos mais relevantes. As bases íris e seeds apresentaram uma acurácia entre 98.63% e 82.58%. Já as bases wine e glass resultaram em uma acurácia entre 87.94% e 81.73%, ao aplicar os quatro filtros de seleção de atributos.

Os resultados obtidos nos experimentos mostram que o modelo é eficaz em rotular grupos, apresentando uma acurácia entre 81.73% e 98.63% e o desvio padrão entre 1.6 e 7.3 para as bases de dados utilizadas. Quanto ao número de elementos na rotulação das bases de dados, o modelo proposto apresenta o melhor desempenho com a base de dados *Íris* utilizando o filtro correlação e a variação V=5% com apenas 3 elementos classificados incorretamente de um total de 150 elementos.

Alguns modelos de rotulação referenciados fazem o processo de discretização para todos os atributos da base de dados, no caso do modelo proposto somente os atributos mais relevantes e selecionados pela variação V são os que necessitam da aplicação da discretização. Apesar do modelo proposto ter resultado em algumas acurácias inferiores, tem a vantagem de não precisar aplicar a discretização em todos os atributos.

Outro detalhe que atua diretamente na rotulação é o fato deste modelo proposto não utilizar o atributo *tipo* das bases para ser o atributo classe ao realizar os testes, o atributo utilizado como classe é o atributo *grupo* que varia de acordo com o valor de K.

6.2 Trabalhos Futuros

Algumas limitações e alternativas podem ser apontadas no intuito de instigar a continuidade da pesquisa em trabalhos futuros, por exemplo: alterar o valor de k na formação dos grupos, encontrar a quantidade de grupo ideal para a rotulação dos dados, ou até mesmo, aplicar outro algoritmo de agrupamento; Tentar criar uma parâmetro para a variação V, de acordo com as características das base de dados. Aplicar o modelo proposto em base de dados maiores, com outras características.

Referências

AEBERHARD, S.; COOMANS, D.; VEL, O. D. Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition*, Elsevier, v. 27, n. 8, p. 1065–1077, 1994. Citado na página 33.

ALBONICO, G. A. M. et al. Seleção de atributos e classificação automática de lesões mamárias em imagens de ultrassom. Universidade Estadual do Oeste do Paraná, 2017. Citado na página 13.

ALGULIYEV, R. et al. Batch clustering algorithm for big data sets. In: IEEE. *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*. [S.l.], 2016. p. 1–4. Citado na página 6.

ALMEIDA, T. B. et al. *Seleção de atributos usando abordagem Wrapper para classificação hierárquica multirrótulo*. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2018. Citado na página 2.

ALMEIDA, T. B. et al. *Seleção de atributos usando abordagem Wrapper para classificação hierárquica multirrótulo*. Dissertação (Mestrado Universidade Tecnológica Federal do Paraná), 2018. Citado na página 8.

ARAÚJO, F. N. C. d. *Rotulação automática de clusters baseados em análise de filogenias*. Dissertação (Mestrado Universidade Federal do Piauí - UFPI), 2018. Citado 2 vezes nas páginas 17 e 40.

ARAUJO, S.; VERAS, R.; MACHADO, V. P. Modelo de classificação de grupos baseado em rotulação e lógica fuzzy. *Anais do 14º Simpósio Basileiro de Automação Inteligente (SBAI)*, 2019. Citado 3 vezes nas páginas 1, 3 e 22.

ASDAGHI, F.; SOLEIMANI, A. An effective feature selection method for web spam detection. *Knowledge-Based Systems*, Elsevier, v. 166, p. 198–206, 2019. Citado na página 10.

BACHE, K.; LICHMAN, M. *(UCI) Machine Learning Repository*. 2013. Acessado em: 23-08-2021. Disponível em: <http://archive.ics.uci.edu/ml>. Citado na página 33.

BACKER, E. *Computer Assisted Reasoning in Cluster Analysis*. New York: Pretice Hall; Har/Dskt edition, 1995. Citado na página 1.

BARON, G. On influence of representations of discretized data on performance of a decision system. *Procedia Computer Science*, Elsevier, v. 96, p. 1418–1427, 2016. Citado na página 15.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. *Artificial intelligence*, Elsevier, v. 97, n. 1-2, p. 245–271, 1997. Citado na página 8.

BRAGA, A. P.; CARVALHO, A.; LUDEMIR, T. Redes neurais artificiais: teoria e aplicação. *Rio de Janeiro: LTC*, 2007. Citado na página 5.

CASTRO, L. N. d.; FERRARI, D. G. Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações. *São Paulo: Saraiva*, v. 5, 2016. Citado na página 10.

CASTRO, P. A. D. de et al. Improving a pittsburgh learnt fuzzy rule base using feature subset selection. In: IEEE. *Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*. [S.l.], 2004. p. 180–185. Citado na página 21.

CHEESEMAN;STUTZ. *Bayesian classification (Autoclass): Theory and results advances in knowleddge discovery and data mining*. [s.n.], 1990. Disponível em: <http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass-c-program.html>. Acesso em: 10 out 2021. Citado na página 1.

CHEN, P. Y. et al. *Correlation: Parametric and nonparametric measures*. [S.l.]: Sage, 2002. Citado na página 12.

COHEN, J. *Statistical power analysis for the behavioral sciences*. [S.l.]: Erlbaum, 1988. Citado na página 11.

DANCEY, C.; REIDY, J. Estadísticas sin matemáticas a la psicología: Usar spss para windows. porto alegre: Artmed. *Revista Psicología*, v. 8, n. 1, p. 5, 2006. Citado na página 11.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. In: *Machine learning proceedings 1995*. [S.l.]: Elsevier, 1995. p. 194–202. Citado na página 14.

EVETT, I. W.; SPIEHLER, E. J. Rule induction in forensic science. In: *Knowledge Based Systems*. [S.l.: s.n.], 1989. p. 152–160. Citado na página 33.

FISHER, D. H. Improving inference through conceptual clustering. In: *AAAI*. [S.l.: s.n.], 1987. v. 87, p. 461–465. Citado na página 33.

GUYON, I.; ELISSEEFF, A. An introduction to feature extraction. In: *Feature extraction*. [S.l.]: Springer, 2006. p. 1–25. Citado na página 8.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. 3. ed. San Francisco: Elsevier, 2011. Citado na página 1.

HART, P. E.; STORK, D. G.; DUDA, R. O. *Pattern classification*. [S.l.]: Wiley Hoboken, 2000. Citado na página 6.

IMPERES Filho, F. et al. Group labeling methodology using distance-based data grouping algorithms. *Revista de Informática Teórica e Aplicada*, v. 27, n. 1, p. 48–61, 2020. Citado 6 vezes nas páginas 1, 3, 5, 22, 33 e 43.

JAIME, T. F. *Uso de Algoritmos de Aprendizado de Máquina Supervisionado para Rotulação de Dados*. Dissertação (Mestrado Universidade Federal do Piauí - UFPI), 2020. Citado na página 14.

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, Elsevier, v. 31, n. 8, p. 651–666, 2010. Citado na página 1.

KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: *Machine learning proceedings 1992*. [S.l.]: Elsevier, 1992. p. 249–256. Citado na página 12.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial intelligence*, Elsevier, v. 97, n. 1-2, p. 273–324, 1997. Citado na página 7.

KONONENKO, I. Estimating attributes: Analysis and extensions of relief. In: SPRINGER. *European conference on machine learning*. [S.l.], 1994. p. 171–182. Citado 2 vezes nas páginas 12 e 13.

KOTSIANTIS, S.; KANELLOPOULOS, D. Discretization techniques. *A recent survey*, 2006. Citado na página 14.

KULCZYCKI, P.; CHARYTANOWICZ, M. A complete gradient clustering algorithm. In: SPRINGER. *International Conference on Artificial Intelligence and Computational Intelligence*. [S.l.], 2011. p. 497–504. Citado na página 33.

LACERDA, M. G. et al. Influência dos parâmetros de segmentação de imagem em ortofotomosaicos confeccionados a partir de fotografias obtidas por aeronaves remotamente pilotadas de pequeno porte. In: *Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto*. [S.l.: s.n.], 2019. v. 1, p. 4. Citado na página 8.

LIMA, B. V. A. de; MACHADO, V. P.; LOPES, L. A. Automatic labeling of social network users scientia. net through the machine learning supervised application. *Social Network Analysis and Mining*, Springer, v. 5, n. 1, p. 1–10, 2015. Citado na página 1.

LIU, H. et al. Feature selection with dynamic mutual information. *Pattern Recognition*, Elsevier, v. 42, n. 7, p. 1330–1339, 2009. Citado na página 7.

LLERENA, S. E. *Redução dimensional de dados de alta dimensão e poucas amostras usando Projection Pursuit*. Tese (Doutorado) — Universidade de São Paulo, 2017. Citado na página 2.

LLOYD, S. Least squares quantization in pcm. *IEEE transactions on information theory*, IEEE, v. 28, n. 2, p. 129–137, 1982. Citado na página 6.

LOPES, L. et al. Automatic labelling of clusters of discrete and continuous data with supervised machine learning. *Knowledge-Based Systems*, v. 106, 05 2016. Citado 7 vezes nas páginas 2, 15, 21, 29, 33, 40 e 41.

MACHADO, V. P.; RIBEIRO, V. P.; RABÊLO, R. A. L. Automatic labeling of groupings through supervised machine learning. *In: Encontro Nacional de Inteligência Artificial e Computacional - ENIAC*, 2013. Citado 3 vezes nas páginas 2, 18 e 43.

MACHADO, V. P.; RIBEIRO, V. P.; RABÊLO, R. A. L. Rotulaçao de grupos utilizando conjuntos fuzzy. *In: XII Simpósio Brasileiro de Automação Inteligente-SBAI*, n. 12, p. 355–360, 2015. Citado na página 1.

MACHADO, V. P.; RIBEIRO, V. P.; RABêLO, R. A. L. Rotulacao de grupos utilizando conjuntos fuzzy. In: *XII Simposio Brasileiro de Automacao Inteligente-SBAI*. [S.l.: s.n.], 2015. p. 355–360. Citado 2 vezes nas páginas 21 e 40.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967. v. 1, n. 14, p. 281–297. Disponível em: <https://projecteuclid.org/euclid.bsmsp/1200512992>. Citado na página 7.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Probabilistic information retrieval. *Introduction to Information Retrieval*, Cambridge Univ. Press, p. 220–235, 2009. Citado 2 vezes nas páginas 7 e 17.

MITCHELL, T. M. Machine learning. McGraw-Hill Science/Engineering/Math: New York, NY, USA, v. 18, n. 3, 1997. Citado na página 5.

MOORE, D. S.; MCCABE, G. P. *Introduction to the Practice of Statistics*. [S.l.]: WH Freeman/Times Books/Henry Holt & Co, 1989. Citado na página 11.

MOURA, M. R. d. S. *CAIBAL - Cluster-Attribute Interdependency Based Automatic Labeler*. Dissertação (Mestrado Universidade Federal do Piauí - UFPI), 2020. Citado 6 vezes nas páginas 1, 22, 33, 40, 41 e 43.

NETTO, O. P. *Um filtro iterativo utilizando árvores de decisão*. Tese (Doutorado) — Universidade de São Paulo, 2013. Citado na página 9.

NORVIG, P.; RUSSELL, S. Inteligência artificial. *Editora Campus*, v. 20, 2004. Citado na página 5.

NOVAKOVIĆ, J. Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, v. 21, n. 1, 2016. Citado na página 10.

PEREIRA, R. B. *Seleção lazy de atributos para a tarefa de classificação*. Tese (Doutorado) — Master's thesis, UFF-Universidade Federal Fluminense, Brazil, 2009. Citado na página 2.

POPAT, S. K.; EMMANUEL, M. Review and comparative study of clustering techniques. *International journal of computer science and information technologies*, Citeseer, v. 5, n. 1, p. 805–812, 2014. Citado na página 6.

SCHIELD, M. Correlation, determination and causality in introductory statistics. *American Statistical Association, Section on Statistical Education*, 1995. Citado na página 11.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: com aplicações em R*. [S.l.]: Elsevier Brasil, 2017. Citado na página 1.

SILVA, S. E. L. *Automatic Group Labeling Based on Regression Error Analysis*. Dissertação (Mestrado Universidade Federal do Piauí - UFPI), 2021. Citado 9 vezes nas páginas 1, 2, 3, 19, 22, 33, 40, 41 e 43.

STANTON, J. M. Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, Taylor & Francis, v. 9, n. 3, 2001. Citado na página 11.

SUTTON, R. S.; BARTO, A. G. et al. *Introduction to reinforcement learning*. [S.l.]: MIT press Cambridge, 1998. v. 135. Citado na página 5.

THEODORIDIS, S.; KOUTROUMBAS, K. Pattern recognition. In: SAN DIEGO. [S.l.]: Academic Press, 2008. Citado na página 13.

TREERATPITUK, P.; CALLAN, J. Automatically labeling hierarchical clusters. In: *Proceedings of the 2006 international conference on Digital government research*. [S.l.: s.n.], 2006. p. 167–176. Citado na página 17.

TZERPOS, V. *Comprehension-driven software clustering*. Tese (Doutorado), 2001. Citado na página 17.

WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002. Citado na página 9.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: NASHVILLE, TN, USA. *Icml*. [S.l.], 1997. v. 97, n. 412-420, p. 35. Citado na página 8.

YEGANOVA, L. et al. How to interpret pubmed queries and why it matters. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 60, n. 2, p. 264–274, 2009. Citado na página 17.

YONEYAMA, T. *Discretização para Aprendizagem Bayesiana: aplicação no auxílio à validação de dados em proteção ao vôo*. Tese (Doutorado) — Instituto Tecnológico de Aeronáutica, 2003. Disponível em: <http://www.ele.ita.br/~jackson/files/msc.pdf>. Citado na página 14.

YU, L.; LIU, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. [S.l.: s.n.], 2003. p. 856–863. Citado na página 7.

Apêndices

APÊNDICE A – RESULTADOS

A variação V influenciou na formação dos rótulos, devido a isso, foram feitos vários testes alterando somente o valor de V para cada filtro de seleção de cada base de dados, os resultados dos testes são descritos a seguir.

A.1 Base Íris

As Tabelas 15 e 16 mostram os resultados da aplicação do filtro ganho de informação na base íris mudando somente o valor da variação V. Foram selecionados os atributos (PL) e (PW) para compor os grupos no primeiro teste, ao variar o valor de V foi adicionado o atributo (SL) para compor os grupos.

Ao aplicar o filtro ganho de informação junto com o cálculo da variação V=10% na base de dados íris foi gerado uma acurácia de 97.62%, e com a variação V=30% na base de dados íris foi gerado uma acurácia de 89.91%.

Tabela 15 – Filtro ganho de informação aplicado a base Iris com V=10%.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	50	PL	[1,0 - 2,9]	0	100
		PW	[0,1 - 0,9]	0	100
2	51	PL	[2,9 - 4,9]	0	100
		PW	[0,9 - 1.7]	1	98
3	49	PL	[4.9 - 6,9]	3	95.9
		PW	[1.7 - 2,5]	4	91.83

Tabela 16 – Filtro ganho de informação com V=30% aplicado a base Íris.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	50	PL	[1,0 - 2,9]	0	100
		PW	[0,1 - 0,9]	0	100
		SL	[4.3 - 5.5]	5	90
2	51	PL	[2,9 - 4,9]	0	100
		PW	[0,9 - 1.7]	1	98
		SL	[5.5 - 6.7]	11	78.43
3	49	PL	[4.9 - 6,9]	3	95.9
		PW	[1.7 - 2,5]	4	91.83
		SL	[5.5 - 6.7]	22	55.10

Já a Tabela 17 mostra os resultados ao testar o filtro razão de ganho com uma variação V=30% na base íris. Neste caso foi adicionado o atributo (SL) para compor os grupos, onde reduziu o valor da acurácia devido a esse atributo ter tido muito elementos classificados incorretamente no grupo 3, gerando somente 55.10% de taxa de acerto.

Tabela 17 – Filtro razão de ganho com V=30% aplicado a base Íris.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	50	PW	[0,1 - 0,9]	0	100
		PL	[1,0 - 2,9]	0	100
		SL	[4.3 - 5.5]	5	90
2	51	PW	[0,9 - 1.7]	1	98
		PL	[2,9 - 4,9]	0	100
		SL	[5.5 - 6.7]	11	78.43
3	49	PW	[1.7 - 2,5]	4	91.83
		PL	[4.9 - 6,9]	3	95.9
		SL	[5.5 - 6.7]	22	55.10

A Tabela 18 mostra o resultado ao aplicar o filtro correlação com a variação $V = 10\%$, onde o resultado selecionou somente os atributos (PL) e (PW) gerando uma acurácia de 97.62%. A mesma taxa média de acerto ao aplicar o filtro ganho de informação.

Tabela 18 – Filtro correlação com V=10% aplicado a base Íris.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	50	PL	[1,0 - 2,9]	0	100
		PW	[0,1 - 0,9]	0	100
2	51	PL	[2,9 - 4,9]	0	100
		PW	[0,9 - 1.7]	1	98
3	49	PL	[4.9 - 6,9]	3	95.9
		PW	[1.7 - 2,5]	4	91.83

A Tabela 19 mostra o resultado ao aplicar o filtro reliefF junto com V=5% onde é selecionado os atributos (PW) e (PL). Resultando em uma taxa média de acerto de 89.91

Tabela 19 – Filtro reliefF com V=5% aplicado a base Íris.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	50	PW	[0,1 - 0,9]	0	100
		PL	[1,0 - 2,9]	0	100
2	51	PW	[0,9 - 1.7]	1	98
		PL	[2,9 - 4,9]	0	100
3	49	PW	[1.7 - 2,5]	4	91.83
		PL	[4.9 - 6,9]	3	95.9

A.2 Base Wine

A Tabela 20 mostra os grupos formados da base wine após aplicar o filtro ganho de informação junto com a variação V=78%, onde foi selecionados os atributos *Proline* e *Flavanoids* obtendo uma acurácia de 78.48%. Já a Tabela 21 mostra o mesmo filtro só que com o valor de V = 82% onde foi adicionado mais o atributo *Alcohol* para compor os grupos, resultando em uma acurácia menor que a anterior 76.01%.

Tabela 20 – Filtro ganho de informação com V=78% aplicado a base Wine.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	66	Proline	[278 - 721]	0	100
		Flavanoids	[0.34 - 1.92]	30	54.54
2	23	Proline	[1164 - 1608]	0	100
		Flavanoids	[3.5 - 5.08]	11	52.17
3	57	Proline	[278 - 721]	19	66.66
		Flavanoids	[0.34 - 1.92]	17	70.17
4	32	Proline	[721 - 1164]	0	100
		Flavanoids	[1.92 - 3.5]	5	84.37

Tabela 21 – Filtro ganho de informação com V=82% aplicado a base Wine.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	66	Proline	[278 - 721]	0	100
		Flavanoids	[0.34 - 1.92]	30	54.54
		Alcohol	[12.3 - 13.57]	20	69.69
2	23	Proline	[1164 - 1608]	0	100
		Flavanoids	[3.5 - 5.08]	11	52.17
		Alcohol	[13.57 - 14.83]	4	82.60
3	57	Proline	[278 - 721]	19	66.66
		Flavanoids	[0.34 - 1.92]	17	70.17
		Alcohol	[12.3 - 13.57]	16	71.92
4	32	Proline	[721 - 1164]	0	100
		Flavanoids	[1.92 - 3.5]	5	84.37
		Alcohol	[12.3 - 13.57]	16	50

A Tabela 22 mostra a formação dos grupos ao aplicar o filtro razão de ganho junto com a variação V=66% na base wine, o resultado foi uma acurácia de 69.29% onde foram selecionados quatro atributos para compor os grupos. O que foi notado é que os atributos *Phenols* e *Mg* tiveram muitos elementos classificados incorretamente nos grupos, devido a isso a acurácia reduziu muito em relação aos outros testes.

Tabela 22 – Filtro razão de ganho com V=66% aplicado a base Wine.

Grupo	# Elem.	Rótulo		Análise	
		Atr	Faixa de Valores	Erro	Acerto(%)
1	66	Proline	[278 - 721]	0	100
		Alcohol	[12.3 - 13.57]	20	69.69
		Phenols	[1.94 - 2.9]	34	48.48
		Mg	[70 - 100.6]	13	80.30
2	23	Proline	[1164 - 1608]	0	100
		Alcohol	[13.57 - 14.83]	4	82.60
		Phenols	[2.9 - 3.88]	11	52.17
		Mg	[100.6 - 131.2]	7	69.56
3	57	Proline	[278 - 721]	19	66.66
		Alcohol	[12.3 - 13.57]	16	71.92
		Phenols	[0.98 - 1.94]	28	50.87
		Mg	[70 - 100.6]	28	50.87
4	32	Proline	[721 - 1164]	0	100
		Alcohol	[12.3 - 13.57]	16	50
		Phenols	[1.94 - 2.9]	14	56.25
		Mg	[100.6 - 131.2]	13	59.37

A Tabela 23 mostra o resultado ao aplicar o filtro correlação junto com V=30% onde somente dois atributos fazem a composição dos grupos e resulta em uma acurácia de 80.10%. Onde o atributo *Proline* tem seus elementos bem classificados em todos os grupos com exceção somente do grupo 3.

Tabela 23 – Filtro correlação com V=30% aplicado a base Wine.

Grupo	# Elem.	Rótulo		Análise	
		Atr	Faixa de Valores	Erro	Acerto(%)
1	66	Proline	[278 - 721]	0	100
		Alcohol	[12.3 - 13.57]	20	69.69
2	23	Proline	[1164 - 1608]	0	100
		Alcohol	[13.57 - 14.83]	4	82.60
3	57	Proline	[278 - 721]	19	66.66
		Alcohol	[12.3 - 13.57]	16	71.92
4	32	Proline	[721 - 1164]	0	100
		Alcohol	[12.3 - 13.57]	16	50

O último teste foi feito com o filtro reliefF com V=75% resultando em uma acurácia de 76.01%,como mostra a Tabela 24, o que se nota é ue para a base de dados wine, quanto mais atributos são selecionados para compor os grupos menor é o valor da acurácia.

Tabela 24 – Filtro reliefF com V=75% aplicado a base Wine.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	66	Proline	[278 - 721]	0	100
		Alcohol	[12.3 - 13.57]	20	69.69
		Flavanoids	[0.34 - 1.92]	30	54.54
2	23	Proline	[1164 - 1608]	0	100
		Alcohol	[13.57 - 14.83]	4	82.60
		Flavanoids	[3.5 - 5.08]	11	52.17
3	57	Proline	[278 - 721]	19	66.66
		Alcohol	[12.3 - 13.57]	16	71.92
		Flavanoids	[0.34 - 1.92]	17	70.17
4	32	Proline	[721 - 1164]	0	100
		Alcohol	[12.3 - 13.57]	16	50
		Flavanoids	[1.92 - 3.5]	5	84.37

A.3 Base Seeds

A Tabela 25 mostra os grupos após a aplicação do filtro ganho de informação junto com a variação V=15% para a base seeds, onde diferente da base wine, ao selecionar mais atributos o valor da acurácia aumenta, neste caso ao selecionar três atributos resultou em 87.59 de acurácia. Nota-se também que o grupo 3 é onde os três atributos selecionados tem grande parte dos seus elementos classificados incorretamente.

Tabela 25 – Filtro ganho de informação com V=15% aplicado a base Seeds.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	69	A	[10.59 - 14.12]	1	98.55
		P	[12.41 - 14.02]	1	98.55
		WK	[2.63 - 3.1]	4	94.20
2	48	A	[17.65 - 21.18]	0	100
		P	[15.63 - 17.25]	0	100
		WK	[3.57 - 4.03]	3	93.75
3	62	A	[10.59 - 14.12]	30	48.38
		P	[14.02 - 15.63]	15	75.80
		WK	[3.1 - 3.57]	10	83.87
4	31	A	[14.12 - 17.65]	0	100
		P	[14.02 - 15.63]	7	77.41
		WK	[3.1 - 3.57]	6	80.64

A Tabela 26 é o resultado da aplicação do filtro razão de ganho com variação V=20%, onde selecionou três atributos também para compor os grupos, em relação ao teste feito com filtro ganho de informação substitui o terceiro atributo, que neste caso é o (LKG) resultando em uma acurácia 83.70%.

Tabela 26 – Filtro razão de ganho com V=20% aplicado a base Seeds.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	69	P	[12.41 - 14.02]	1	98.55
		A	[10.59 - 14.12]	1	98.55
		LKG	[4.51 - 5.19]	23	66.66
2	48	P	[15.63 - 17.25]	0	100
		A	[17.65 - 21.18]	0	100
		LKG	[5.87 - 6.55]	6	87.5
3	62	P	[14.02 - 15.63]	15	75.80
		A	[10.59 - 14.12]	30	48.38
		LKG	[4.51 - 5.19]	140	77.41
4	31	P	[14.02 - 15.63]	7	77.41
		A	[14.12 - 17.65]	0	100
		LKG	[5.19 - 5.87]	8	74.19

Já para o teste feito com o filtro correlação onde possui um *ranking* diferente dos outros filtros, neste caso ao aplicá-lo junto a uma variação V=3% selecionou quatro atributos sendo o mais relevante o atributo (LKG) apresentou uma acurácia de 82.58%. Como mostra a Tabela 27

Tabela 27 – Filtro correlação com V=3% aplicado a base Seeds.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	69	LKG	[4.1 - 5.19]	23	66.66
		A	[10.59 - 14.12]	1	98.55
		P	[12.41 - 14.02]	1	98.55
		LK	[4.89 - 5.49]	1	98.55
2	48	LKG	[5.87 - 6.55]	6	87.5
		A	[17.65 - 21.18]	0	100
		P	[15.63 - 17.25]	0	100
		LK	[6.08 - 6.67]	9	81.25
3	62	LKG	[4.51 - 5.19]	14	66.66
		A	[10.59 - 14.12]	30	48.38
		P	[14.02 - 15.63]	15	75.80
		LK	[5.49 - 6.08]	31	50
4	31	LKG	[5.19 - 5.87]	8	74.19
		A	[14.12 - 17.65]	0	100
		P	[14.02 - 15.63]	7	77.41
		LK	[5.49 - 6.08]	4	87.09

A Tabela 28 mostra o filtro reliefF junto com a variação V=20% aplicado a base seeds resultando em uma taxa média de acerto 86.66% .

Tabela 28 – Filtro reliefF com V=20% aplicado a base Seeds.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	69	A	[10.59 - 14.12]	1	98.55
		P	[12.41 - 14.02]	1	98.55
		WK	[2.63 - 3.1]	4	94.20
		LK	[4.89 - 5.49]	1	98.55
2	48	A	[17.65 - 21.18]	0	100
		P	[15.63 - 17.25]	0	100
		WK	[3.57 - 4.03]	3	93.75
		LK	[6.08 - 6.67]	0	100
3	62	A	[10.59 - 14.12]	30	48.38
		P	[14.02 - 15.63]	15	75.80
		WK	[3.1 - 3.57]	10	83.87
		LK	[5.49 - 6.08]	31	50
4	31	A	[14.12 - 17.65]	0	100
		P	[14.02 - 15.63]	7	77.41
		WK	[3.1 - 3.57]	6	80.64
		LK	[5.49 - 6.08]	4	87.09

A.4 Base Glass

A Tabela 29 mostra o resultado ao aplicar o filtro ganho de informação com uma variação V=5%, onde foram selecionados dois atributos (Mg), (Ca) para compor os grupos, resultando em uma acurácia de 77.60%.

Tabela 29 – Filtro ganho de informação com V=5% aplicado a base Glass.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	75	Mg	[3.36 - 4.49]	12	84
		Ca	[8.12 - 10.81]	7	90.66
2	5	Mg	[2.24 - 3.36]	2	60
		Ca	[5.43 - 8.12]	0	100
3	33	Mg	[0 - 1.12]	11	66.66
		Ca	[8.12 - 10.81]	4	87.87
4	16	Mg	[0 - 1.12]	5	68.75
		Ca	[10.81 - 13.5]	5	68.75
5	7	Mg	[0 - 1.12]	0	100
		Ca	[10.81 - 13.5]	3	57.14
6	78	Mg	[3.36 - 4.49]	13	83.33
		Ca	[5.43 - 8.12]	28	64.10

A Tabela 30 mostra o resultado ao aplicar o filtro razão de ganho com V=50%, onde selecionou três atributos para compor os grupos e resultou em uma acurácia 78.75,

um pouco maior que a do teste feito com o filtro ganho de informação. Neste caso foi adicionado o terceiro atributo (Ca) onde ele teve seus elementos bem classificados em todos os grupos com exceção somente do grupo 6.

Tabela 30 – Filtro razão de ganho com V=50% aplicado a base Glass.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	75	Mg	[3.36 - 4.49]	12	84
		Ba	[0 - 0.79]	0	100
		Ca	[8.12 - 10.81]	7	90.66
2	5	Mg	[2.24 - 3.36]	2	60
		Ba	[1.57 - 2.36]	3	40
		Ca	[5.43 - 8.12]	0	100
3	33	Mg	[0 - 1.12]	11	66.66
		Ba	[0 - 0.79]	13	60.60
		Ca	[8.12 - 10.81]	4	87.87
4	16	Mg	[0 - 1.12]	5	68.75
		Ba	[0 - 0.79]	0	100
		Ca	[10.81 - 13.5]	5	68.75
5	7	Mg	[0 - 1.12]	0	100
		Ba	[0 - 0.79]	1	85.7
		Ca	[10.81 - 13.5]	3	57.14
6	78	Mg	[3.36 - 4.49]	13	83.33
		Ba	[0 - 0.79]	0	100
		Ca	[5.43 - 8.12]	28	64.10

A Tabela 31 mostra o resultado após aplicar o filtro correlação com a variação V=40%, houve uma mudança no terceiro atributo selecionado em relação ao teste feito com filtro razão de ganho. No teste anterior o terceiro atributo era o Ca, já nesse é o Al. A mudança também resulto em uma acurácia um pouco inferior também 76.35%. O atributo Al teve seus grande parte dos seus elementos classificados incorretamente no grupo 3.

Tabela 31 – Filtro correlação com V=40% aplicado a base Glass.

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
1	75	Mg	[3.36 - 4.49]	12	84
		Ba	[0 - 0.79]	0	100
		Al	[1.09 - 1.89]	18	76
2	5	Mg	[2.24 - 3.36]	2	60
		Ba	[1.57 - 2.36]	3	40
		Al	[2.69 - 3.5]	2	60
3	33	Mg	[0 - 1.12]	11	66.66
		Ba	[0 - 0.79]	13	60.60
		Al	[1.89 - 2.69]	14	57.57

Grupo	**# Elem.**	**Rótulo**		**Análise**	
		Atr	**Faixa de Valores**	**Erro**	**Acerto(%)**
4	16	Mg	[0 - 1.12]	5	68.75
		Ba	[0 - 0.79]	0	100
		Al	[1.09 - 1.89]	2	87.5
5	7	Mg	[0 - 1.12]	0	100
		Ba	[0 - 0.79]	1	85.7
		Al	[0.29 - 1.09]	3	57.14
6	78	Mg	[3.36 - 4.49]	13	83.33
		Ba	[0 - 0.79]	0	100
		Al	[1.09 - 1.89]	10	87.17

Com os teste apresentados no Apêndice nota-se a influência da variação V na quantidade de atributos selecionados para compor os grupos, sendo assim, essa seleção afeta na taxa de acerto de cada grupo consequentemente na acurácia do filtro aplicado. Em alguns casos, como na base seeds, ao aplicar o filtro correlação, quanto mais adiciona-se atributo para compor os grupos, maior é a acurácia, mas para as outras bases e os outros filtro, esse aumento na quantidade de atributos decresce o resultado da acurácia.