Universidade Federal do Rio Grande do Norte
Centro de Ciências Exatas e da Terra
Departamento de Informática e Matemática Aplicada
Programa de Pós-Graduação em Sistemas e Computação
Mestrado Acadêmico em Sistemas e Computação

Seleção de Atributos Baseado em Algoritmos de Agrupamento para Tarefas de Classificação

Carine Azevedo Dantas

Natal-RN
Fevereiro de 2017

Carine Azevedo Dantas

Seleção de Atributos Baseado em Algoritmos de Agrupamento para Tarefas de Classificação

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação de Sistemas e Computação da Universidade Federal do Rio Grande do Norte.

Orientadora

Prof[a]. Dr[a]. Anne Magály de Paula Canuto

UFRN – Universidade Federal do Rio Grande do Norte
DIMAp – Departamento de Informática e Matemática Aplicada

Natal-RN

Fevereiro de 2017

Catalogação da Publicação na Fonte. UFRN / SISBI / Biblioteca Setorial
Centro de Ciências Exatas e da Terra – CCET.

Dantas, Carine Azevedo.
Seleção de atributos baseada em algoritmos de agrupamento para tarefas de classificação / Carine Azevedo Dantas. - Natal, 2017.
71 f. : il.

Orientadora: Profa. Dra. Anne Magály de Paula Canuto.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Norte. Centro de Ciências Exatas e da Terra. Programa de Pós-Graduação em Sistemas e Computação.

1. Inteligência artificial – Dissertação. 2. Aprendizado de máquina – Dissertação. 3. Seleção de atributos – Dissertação. 4. Classificação – Dissertação. 5. Algoritmos de agrupamento – Dissertação. I. Canuto, Anne Magály de Paula. II. Título.

RN/UF/BSE-CCET CDU: 004.8

CARINE AZEVEDO DANTAS

Seleção de atributos baseado em algoritmos de agrupamento para tarefas de classificação

Esta Dissertação foi julgada adequada para a obtenção do título de mestre em Sistemas e Computação e aprovado em sua forma final pelo Programa de Pós-Graduação em Sistemas e Computação do Departamento de Informática e Matemática Aplicada da Universidade Federal do Rio Grande do Norte.

Dra. Anne Magaly de Paula Canuto – UFRN
(Presidente)

Dr. Martin Alejandro Musicante – UFRN
(Vice-Coordenador do Programa)

Banca Examinadora

Dr. Bruno Motta de Carvalho - UFRN
(Examinador)

Dr. Daniel Sabino Amorim de Araújo - UFRN
(Examinador)

Dr. Joao Carlos Xavier Junior – UFRN
(Examinador)

Dra. Adriana Takahashi – UERN
(Examinadora)

Fevereiro, 2017

Seleção de atributos baseado em algoritmos de agrupamento para tarefas de classificação

Autor: Carine Azevedo Dantas
Orientador(a): Profª. Drª. Anne Magály de Paula Canuto

Resumo

Com o aumento do tamanho dos conjuntos de dados utilizados em sistemas de classificação, a seleção dos atributos mais relevantes se tornou uma das principais tarefas da fase de pré-processamento. Em um conjunto de dados é esperado que todos os atributos que o descreve sejam relevantes, porém isso nem sempre acontece. Selecionar o conjunto de atributos mais relevantes ajuda a reduzir a dimensionalidade dos dados sem afetar o desempenho, ou até mesmo melhorá-lo, para que se possa obter melhores resultados quando utilizado na classificação de dados. Os métodos de seleção de características existentes selecionam os melhores atributos para uma base de dados como um todo, sem levar em consideração as particularidades de cada instância. A Seleção de atributos baseada em algoritmos de agrupamento, método proposto deste trabalho, seleciona os atributos mais relevantes para cada grupo de instâncias, utilizando algoritmos de agrupamento para agrupá-las de acordo com as suas semelhanças. Este trabalho efetua uma análise experimental de diferentes técnicas de agrupamento aplicadas a essa nova abordagem de seleção de atributos. Para isso, são utilizados os algoritmos de agrupamento k-Médias, DBscan e *Expectation-Maximization* (EM) como métodos de seleção. São efetuadas análises de desempenho e estatísticas para verificar qual desses algoritmos de agrupamento melhor se adequa a essa nova Seleção de Atributos. Assim, a contribuição deste trabalho é apresentar uma nova abordagem, através de uma versão Semidinâmica e outra Dinâmica, para seleção de atributos baseada em algoritmos de agrupamento e determinar qual dos métodos de agrupamento realiza uma melhor seleção e obtém um melhor desempenho na construção de classificadores mais acurados.

Palavras-chave: Seleção de atributos, Classificação, Algoritmos de agrupamento.

An Unsupervised-based Feature Selection for Classification tasks

Author: Carine Azevedo Dantas
Supervisor: Prof.ª Dr.ª Anne Magály de Paula Canuto

Abstract

With the increase of the size on the data sets used in classification systems, selecting the most relevant attribute has become one of the main tasks in pre-processing phase. In a dataset, it is expected that all attributes are relevant. However, this is not always verified. Selecting a set of attributes of more relevance aids decreasing the size of the data without affecting the performance, or even increase it, this way achieving better results when used in the data classification. The existing features selection methods elect the best attributes in the data base as a whole, without considering the particularities of each instance. The Unsupervised-based Feature Selection, proposed method, selects the relevant attributes for each instance individually, using clustering algorithms to group them accordingly with their similarities. This work performs an experimental analysis of different clustering techniques applied to this new feature selection approach. The clustering algorithms k-Means, DBSCAN and Expectation-Maximization (EM) were used as selection methods. Anaysis are performed to verify which of these clustering algorithms best fits to this new Feature Selection approach. Thus, the contribution of this study is to present a new approach for attribute selection, through a Semidynamic and a Dynamic version, and determine which of the clustering methods performs better selection and get a better performance in the construction of more accurate classifiers.

Keywords: Feature Selection, Classification, Clustering Algorithms.

Lista de Figuras

1 Fluxograma do Algoritmo k-Médias . . . . . p. 19

2 Ilustração da estrutura geral da Versão Semidinâmica do método proposto. p. 33

3 Ilustração da estrutura geral da Versão Dinâmica do método proposto. p. 36

4 Estrutura Metodológica dos Experimentos. . . . . p. 40

Lista de Tabelas

1 Bases de Dados Usadas nos Experimentos. . . . . p. 41

2 Média das ordenações e desvio padrão para diferentes proporções utilizando o k-Médias como seletor de atributos. . . . . p. 48

3 Resultados do Teste de *Friedman* (p-valor) para diferentes proporções utilizando o k-Médias como seletor de atributos. . . . . p. 48

4 Média das ordenações e desvio padrão para diferentes proporções utilizando o DBSCAN como seletor de atributos . . . . . p. 49

5 Resultados do Teste de *Friedman* (p-valor) para diferentes proporções utilizando o DBSCAN como seletor de atributos. . . . . p. 50

6 Média das ordenações e desvio padrão para diferentes proporções utilizando o EM como seletor de atributos. . . . . p. 51

7 Resultados do Teste de *Friedman* (p-valor) para diferentes proporções utilizando o EM como seletor de atributos. . . . . p. 52

8 Média das ordenações e desvio padrão para todos algoritmos de agrupamento, separado por base. . . . . p. 53

9 Resultados do Teste de *Friedman* (p-valor) para todos algoritmos de agrupamento, separado por base. . . . . p. 54

10 Média das ordenações e desvio padrão dos algoritmos de agrupamento para todos os métodos de classificação. . . . . p. 54

11 Resultados do teste de *Friedman* (p-valor) para todos os algoritmos de agrupamento, separado por métodos de classificação . . . . . p. 55

12 Média das ordenações e desvio padrão da comparação das duas versões. p. 56

13 Resultados do teste de *Friedman* (p-valor) para comparação entre as duas versões do método proposto. . . . . p. 57

14 Acurácia média e desvio padrão da comparação da melhor versão da Seleção Semidinâmica com os métodos existentes de redução de dimensionalidade . . . . . p. 59

15 Resultados do teste de *Friedman* (p-valor) da comparação da melhor versão da Seleção Semidinâmica com os métodos existentes de redução de dimensionalidade . . . . . p. 60

16 Acurácia média das proporções do k-Médias . . . . . p. 67

19 Acurácia média dos algoritmos de agrupamento para todos os métodos de classificação . . . . . p. 67

17 Acurácia média das proporções do DBSCAN . . . . . p. 68

18 Acurácia média das proporções do EM . . . . . p. 68

20 Acurácia média das proporções da Versão Semidinâmica (V1) . . . . . p. 69

21 Acurácia média das proporções da Versão Dinâmica (V2) . . . . . p. 69

22 Acurácia média da comparação da melhor versão com alguns métodos existentes de redução de dimensionalidade . . . . . p. 70

Lista de Abreviaturas e Siglas

PCA – Análise de Componentes Principais

LDA – Análise Discriminante Linear

DBSCAN – Density-Based Spatial Clustering

EM – *Expectation Maximization*

k-NN – k-Nearest Neighbors

SVM – Support Vector Machine

SDFS – SemiDynamic Features Selection

Sumário

1 Introdução p. 12

1.1 Motivação . . . . . p. 13

1.2 Objetivos . . . . . p. 14

1.3 Organização do Trabalho . . . . . p. 15

2 Conceitos p. 16

2.1 Seleção de Atributos . . . . . p. 16

2.2 Extração de Características . . . . . p. 16

2.2.1 Análise de Componentes Principais . . . . . p. 17

2.2.2 LDA . . . . . p. 17

2.3 Técnicas de Agrupamento . . . . . p. 18

2.3.1 k-Médias . . . . . p. 18

2.3.2 DBSCAN . . . . . p. 20

2.3.3 *Expectation Maximization* . . . . . p. 22

2.4 Algoritmos de Classificação . . . . . p. 22

2.4.1 k-NN . . . . . p. 22

2.4.2 Árvore de Decisão . . . . . p. 23

2.4.3 SVM . . . . . p. 24

2.4.4 *Naive Bayes* . . . . . p. 25

2.5 Testes Estatísticos . . . . . p. 25

2.5.1 Teste de *Friedman* . . . . . p. 26

3 Trabalhos Relacionados p. 28

3.1 Seleção de Atributos para Classificação . . . . . p. 28

3.2 Seleção de Atributos Utilizando Algoritmos de Agrupamento . . . . . p. 30

4 Nova Abordagem de Seleção de Atributos Baseado em Algoritmos de Agrupamento p. 32

4.1 Arquitetura Geral da Versão Semidinâmica . . . . . p. 32

4.1.1 Funcionamento . . . . . p. 34

4.2 Arquitetura Geral da Versão Dinâmica . . . . . p. 35

4.2.1 Funcionamento . . . . . p. 37

4.3 Considerações Finais do Capítulo . . . . . p. 38

5 Análise Experimental p. 39

5.1 Bases de Dados Utilizadas . . . . . p. 41

5.2 Métodos e Materiais . . . . . p. 44

6 Resultados p. 46

6.1 Considerações Iniciais . . . . . p. 46

6.2 Análise dos Percentuais de Seleção de Atributos . . . . . p. 47

6.2.1 Análise de Proporção do k-Médias . . . . . p. 47

6.2.2 Análise de proporção do DBSCAN . . . . . p. 49

6.2.3 Análise de proporção do EM . . . . . p. 50

6.2.4 Conclusão da Primeira Análise . . . . . p. 51

6.3 Análise entre os Diferentes Algoritmos de Agrupamento . . . . . p. 52

6.3.1 Conclusão da Segunda Análise . . . . . p. 55

6.4 Comparação entre a Versão Semidinâmica e Dinâmica . . . . . p. 56

6.4.1 Conclusão da Terceira Análise . . . . . p. 57

6.5 Comparação da Abordagem Proposta com Outros Métodos de Redução de Dimensionalidade . . . . . p. 58

7 Conclusão p. 61

7.1 Considerações Finais . . . . . p. 61

7.1.1 Principais Contribuições . . . . . p. 62

7.1.2 Trabalhos Futuros . . . . . p. 63

Referências p. 64

Apêndice A – Resultados das Médias das Acurácias p. 67

1 Introdução

Com um crescimento demasiadamente rápido do uso de sistemas informatizados e o crescimento da Internet, o volume de dados gerados e armazenados tem aumentado a cada dia. Estes dados podem ser de diversas áreas, como educação, saúde, comércio, dentre outras (WITTEN; FRANK, 2005).

A medida que o volume de dados aumenta, a dificuldade de compreendê-los aumenta drasticamente. Visando obter informações úteis destes grandes conjuntos de dados, surgiu um novo ramo na computação, o Aprendizado de Máquina (MITCHELL, 1997). Analisar enormes conjuntos de dados e encontrar alguma semelhança que contém algum significado entre eles é uma tarefa bastante complexa para a capacidade humana. Portanto, Aprendizado de Máquina é um processo que explora e analisa volumes de dados, com o intuito de filtrar, entender o que é relevante e fazer um bom uso dessas informações no processo de tomada de decisão.

Conjuntos de dados são formados por objetos que podem representar um determinado elemento do mundo real, seja ele físico, como um carro, ou abstrato, como sintomas de um paciente. Esses objetos também são conhecidos como instâncias ou registros. Cada objeto é descrito por um conjunto fixo de características, também conhecido por atributos, responsável por descrever as suas propriedades. Em cada instância pode ocorrer a presença de um atributo especial, chamado de classe ou rótulo, utilizado para identificar instâncias com características similares (WITTEN; FRANK, 2005).

As tarefas mais comuns no Aprendizado de Máquina são a classificação e o agrupamento (BERRY; LINOFF, 2000). A classificação tem o intuito de construir um modelo que possa ser usado para associar uma instância a uma determinada classe. O modelo analisa o conjunto de instâncias fornecidas, em que cada uma dessas instâncias já está associada a uma classe, com a finalidade de aprender a classificar uma nova instância que venha a surgir. Já o agrupamento visa identificar e alocar as instâncias em grupos, de forma que elas sejam similares entre si, mas diferentes de instâncias pertencentes a

outros grupos. O que difere esta tarefa da classificação é que os registros não precisam estar previamente associados a uma determinada classe (HARRISON, 1998).

Devido às diversas fontes das quais os dados são obtidos, é possível que os dados não estejam adequados ao método de mineração de dados que será utilizado. Dependendo da qualidade dos dados, algumas ações precisam ser tomadas para adequá-los as tarefas que serão realizadas. Esta adequação se dá pela etapa de pré-processamento.

A etapa pré-processamento do processo de Descoberta de Conhecimento em Banco de Dados inicia-se com uma análise inicial dos dados para obter informações definidas, a fim de identificar e tratar dados defeituosos, valores desconhecidos e atributos irrelevantes. Uma das fases importante do pré-processamento é a seleção de atributos (CHANDRASHEKAR; SAHIN, 2014). Pode ocorrer que alguns atributos não sejam importantes ou adicionem pouca informação para o objetivo de classificação (WITTEN; FRANK, 2005). Portanto, esta fase propõe obter um subconjunto de atributos que substituam um conjunto de dados original, visando reduzir a dimensionalidade e a complexidade para extrair algum conhecimento, ajudando a compreender os dados, reduzir a exigência de processamento computacional e melhorar o desempenho do modelo.

Para se obter o melhor conjunto de atributos possíveis é necessário um profundo conhecimento sobre o domínio do problema. Já que isso nem sempre é possível, uma das possibilidades é usar métodos de seleção de atributos automática para definir os atributos mais importantes ou relevantes para criação do modelo que será utilizado durante a classificação dos dados.

1.1 Motivação

A etapa de Seleção de Atributos, está presente apenas no início do processo de Mineração de Dados, porém sua aplicação requer grande cuidado, pois requisita muito processamento e leva uma boa parte do tempo necessário para melhorar os dados que serão utilizados (LIU; MOTODA, 1998). Realizar essa etapa de maneira automática, com o auxílio de um computador é extremamente vantajoso, devido ao grande volume de dados utilizados atualmente nos problemas de classificação. Alguns métodos de classificação não lidam de maneira adequada com atributos irrelevantes, como por exemplo, a árvore de decisão, tendo em vista que no momento da sua construção, um atributo irrelevante pode ser escolhido devido à diminuição dos dados disponíveis para a escolha deste atributo, ocorrendo um aumento da profundidade da árvore e diminuindo a sua precisão. Por isso,

existem diversos métodos que realizam a seleção dos melhores atributos de um conjunto de dados.

Os métodos mais conhecidos na literatura selecionam os melhores atributos para toda uma base de dados, sem levar em consideração as particularidades de cada instância. Uma alternativa para os métodos de seleção tradicional seria um método que escolhe o melhor subconjunto de atributos para cada instância, ou um grupo de instâncias, de forma dinâmica. Neste processo de seleção dinâmica, visa-se selecionar para cada instância os atributos que causam um maior impacto durante a classificação, explorando o potencial de cada instância individualmente, tendo em vista que os atributos selecionados para uma instância qualquer não são necessariamente os mesmos para as demais instâncias.

Sendo assim, a principal motivação deste trabalho é apresentar uma nova abordagem de seleção de atributos, que busca proporcionar um melhor desempenho para tarefas de classificação, simplificando os modelos de predição e diminuindo o custo computacional através da redução dos atributos menos relevantes para cada uma das instâncias, melhorando assim, a performance dos algoritmos de classificação que utilizarem essa seleção de atributos como método de redução de dimensionalidade. Essa nova abordagem será apresentada através de duas versões, uma Semidinâmica e outra Dinâmica.

1.2 Objetivos

O principal objetivo deste trabalho é propor uma nova abordagem de redução de dimensionalidade para conjuntos de dados utilizados em sistemas de classificação. Essa abordagem utiliza algoritmos de agrupamento para agrupar os padrões semelhantes e extrair os melhores atributos para cada instância ou um grupo de instâncias, baseado na estratégia divisão-e-conquista. Esta ideia consiste em dividir a base de dados em grupos, em seguida selecionar os melhores atributos para cada grupo, após isto, o processo de classificação para uma determinada instância é realizado baseado em sua similaridade com os grupos formados. Como objetivos da abordagem proposta, pode-se citar:

- Apresentar uma abordagem eficiente para Seleção de Atributos. Essa nova abordagem será apresentada através de duas versões, uma Semidinâmica e uma Dinâmica.
- Melhorar o desempenho dos métodos de classificação que utilizem a nova abordagem como método de seleção de atributos, através da redução dos atributos irrelevantes e do uso da nova abordagem para utilizar apenas os atributos mais importantes para

cada instância.

- Diminuir o custo computacional dos métodos de classificação utilizados através da redução dos atributos, pois uma menor dimensionalidade implica em um menor tempo de processamento para cada algoritmo de aprendizagem utilizado.

- Melhorar a acurácia de modelos preditivos, tendo em vista que atributos irrelevantes afetam diretamente no aprendizado do classificador. Então, a seleção dos atributos mais importantes e a exclusão daqueles que menos impactam positivamente na classificação, aumenta a chance do classificador realizar uma previsão correta da classe de cada uma das instâncias, pois a complexidade de um problema aumenta exponencialmente com o aumento do número de atributos do conjunto de dados, ficando mais difícil para o algoritmo de aprendizagem encontrar um modelo preciso (BELLMAN, 1956).

- Investigar a melhor estrutura para a abordagem proposta, principalmente no que se refere ao algoritmo de agrupamento a ser utilizado como seletor dos atributos mais relevantes.

1.3 Organização do Trabalho

O restante deste trabalho está organizado em 6 capítulos: No Capítulo 2 serão apresentados os principais conceitos utilizados na realização deste trabalho. O Capítulo 3 apresenta os trabalhos relacionados ao tema do trabalho. No Capítulo 4, a metodologia do trabalho é apresentada, ilustrando o funcionamento das versões desenvolvidas. No Capítulo 5 será apresentado a análise experimental com os principais parâmetros e configurações utilizados na abordagem proposta. O Capítulo 6 apresenta os resultados obtidos com a utilização do método proposto nos experimentos realizados. Por fim, o Capítulo 7 apresenta a conclusão do trabalho, enfatizando os pontos positivos obtidos com o trabalho e levantando ideias para que o trabalho continue futuramente.

2 Conceitos

Este capítulo tem a finalidade de apresentar os principais fundamentos teóricos que foram utilizados durante a concepção deste trabalho.

2.1 Seleção de Atributos

O conjunto de características dos elementos de um problema é denominado de conjunto atributos, onde cada elemento é um atributo que caracteriza o problema definido por um especialista. No entanto, pode ocorrer que alguns atributos não sejam importantes ou adicionem pouca informação para o objetivo de classificação. Um problema clássico deste contexto é a seleção do melhor subconjunto de atributos que maximize uma medida de avaliação (GUYON; ELISSEEFF, 2003).

A seleção de atributos é um passo da fase de pré-processamento que tem a finalidade de selecionar o melhor subconjunto de atributos que pode substituir o conjunto de dados original. A seleção de atributos não só reduz a dimensão do conjunto de dados, mas também ajuda na compreensão de dados, reduzindo a exigência de computação e melhorando o desempenho do modelo (CHANDRASHEKAR; SAHIN, 2014).

2.2 Extração de Características

A extração de características consiste em outra maneira de reduzir a dimensionalidade dos dados. Esse processo ocorre através da criação de novos atributos, normalmente usados para diminuir o tamanho do conjunto de dados.

Na extração de características existem dois principais extratores bastante utilizados na literatura, são eles: A análise de componentes principais (PCA) e a análise discriminante linear (LDA). Nas próximas seções serão apresentadas o funcionamento de cada um algoritmo (FORMAN, 2003).

2.2.1 Análise de Componentes Principais

O PCA é um método de aprendizado de máquina que foi criado em 1901 pelo um dos grandes contribuídores para o desenvolvimento da estatística, *Karl Pearson* (PEARSON, 1901). O PCA é um método de extração de características, que consiste em obter novas variáveis a partir dos atributos iniciais. Com o PCA, obtemos um pequeno número de componentes principais (combinações lineares) de um conjunto de variáveis. O objetivo não é apenas reduzir e sim conseguir preservar o máximo possível das informações contidas nas variáveis originais. O componente principal é o arranjo que melhor representa a distribuição dos dados e o componente secundário é perpendicular ao componente principal (MARDIA; KENT, 1979).

Abaixo serão apresentados os principais passos deste algoritmo:

- Inicialmente é calculada a média de todos os valores dados.
- Todo o conjunto de dados é normalizado.
- Após a normalização, é calculada a matriz de covariância.
- Extraem-se os autovetores e autovalores desta matriz.
- É escolhido os k autovetores com maior quantidade de informação associada, ou seja, o autovetor com o maior autovalor associado, corresponde ao componente principal do conjunto de dados usado. Isso significa que esse é o relacionamento mais significativo entre as dimensões dos dados.
- Monta-se a matriz de transformação baseada nos autovetores selecionados previamente. Os autovetores desta matriz formam uma nova base que segue a variação dos dados.

2.2.2 LDA

A LDA também é um método de aprendizado de máquina que busca extração de características. O LDA gera um novo conjunto de dados de menor dimensionalidade que representa as classes dos dados originais, minimizando a dispersão entre os registros da mesma classe e maximizando a distância entre as classes (FISHER, 1936). O método tenta encontrar a melhor direção, de maneira que quando os dados são projetados em um plano, as classes possam ser separadas.

Com esta técnica é possível reduzir os dados por meio de um mapeamento dos objetos em um espaço de alta dimensionalidade para um espaço de características com menor dimensionalidade.

Abaixo serão apresentados os principais passos deste algoritmo:

- Inicialmente é calculada a média de todos os valores dados.
- Todo o conjunto de dados é normalizado.
- Após a normalização, é calculado a matriz de covariância.
- Em seguida, é calculada a matriz de covariância conjunta.
- É calculada a inversa da matriz de covariância conjunta.
- Aplicar a função discriminante.
- Atribuir um objeto a um grupo que maximize a função discriminante.

2.3 Técnicas de Agrupamento

Agrupamento é uma técnica de aprendizado não supervisionado, ou seja, quando não tem o conhecimento do atributo classe de uma base de dados. O objetivo principal das técnicas de agrupamento é agrupar os objetos de maneira que a similaridade entre os que formam um mesmo grupo (intragrupo) seja a maior possível e a similaridade dos objetos de um grupo em comparação aos componentes dos outros grupos (intergrupo) seja a menor possível (GAN; MA; WU, 2007).

Neste trabalho foram utilizados três técnicas de agrupamento, são elas: k-Médias, DBSCAN e *Expectation Maximization*. Nas próximas subseções serão apresentadas o funcionamento de cada algoritmo.

2.3.1 k-Médias

O método k-Médias (MACQUEEN, 1997) é um algoritmo de agrupamento baseado na distância, de forma que, k instâncias são escolhidas aleatoriamente e são definidos centróides dos seus grupos. O procedimento é feito de maneira relativamente simples e fácil, classificando um determinado conjunto de dados através de um certo número de grupos (assumindo k-grupos), fixado a *priori* (BERRY; LINOFF, 2000).

Para iniciar o algoritmo k-Médias é necessário receber o parâmetro k que indica a quantidade de grupos nos quais se quer formar com os dados da base. A Figura 1 ilustra o fluxograma do k-Médias.

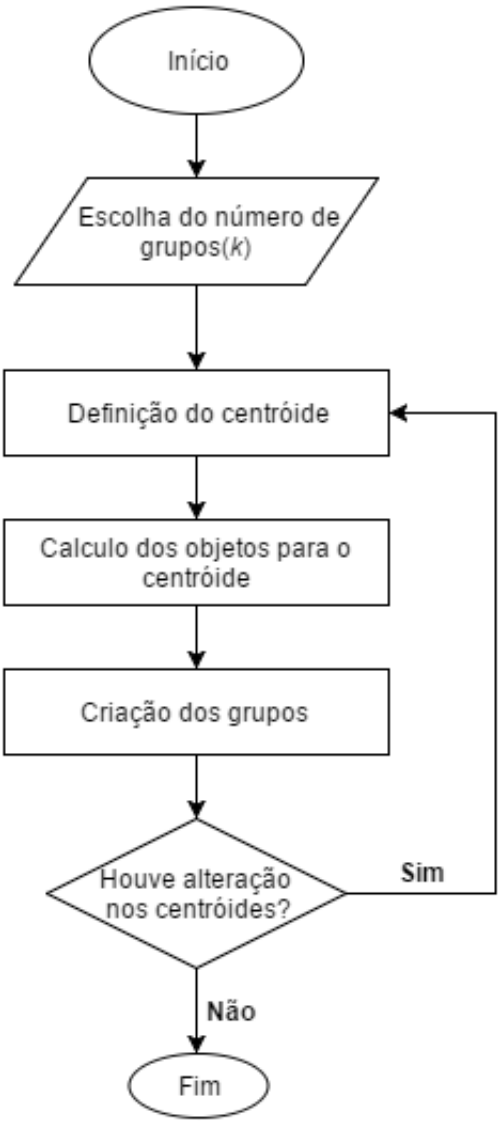


Figura 1: Fluxograma do Algoritmo k-Médias

,

- O primeiro passo propriamente dito é particionar os objetos, de forma aleatória, em k grupos não vazios, onde cada um desses objetos será o centróide, ou seja, o representante de seu grupo, ainda unitário.

- O segundo passo, cada objeto restante será testado em cada grupo, através de uma função de distância, e serão alocados para o grupo que tenham a maior similaridade, ou seja, menor distância.

- O terceiro passo é atualizar os centróides de cada agrupamento. As novas posições dos centróides será a média dos objetos que pertencem ao agrupamento.

- No quarto passo, se elementos mudarem de agrupamentos no Passo 2, repetir os passos 2-3. Se não, o algoritmo para.

O k-Médias utiliza medidas diretas de distância para calcular a similaridade entre as instâncias, esta é usada para calcular a distância entre as instâncias restantes e os centróides de cada grupo. Uma das mais utilizadas para o k-Médias é a Distância Euclidiana (CLIFFORD; STEPHENSON, 1975).

Este algoritmo é simples, intuitivo e também eficiente em tratar grandes conjuntos de dados. Porém, ele é sensível a ruídos, onde as instâncias com valores altos podem influenciar no cálculo das médias do grupos.

2.3.2 DBSCAN

O método do DBSCAN é um algoritmo de agrupamento baseado em densidade, onde seu principal objetivo é buscar por objetos que possuem mais que um limiar de vizinhos dentro de um determinado raio (THOM; KRAMER, 2010).

Para melhor entendimento do algoritmo, é necessário ter conhecimento de definições básicas (ESTER et al., 1996), que serão apresentadas a seguir:

Os dois parâmetros de entrada do DBSCAN, são:

- *Raio de vizinhança de um ponto(ϵ)*: Para cada instância da base de dados, o algoritmo determina ϵ e verifica a quantidade de pontos que estão dentro do raio, se esse valor ultrapassar um certo número, é formado o agrupamento.

- *Número mínimo de pontos(η)*: O parâmetro η tem como objetivo especificar a quantidade mínima de instâncias em um dado raio ϵ.

O algoritmo identifica três tipos de objetos em uma base de dados:

- *Pontos centrais*: São pontos que estão no interior de uma região densa. Sendo assim, é uma região em que todos os pontos têm pelo menos η pontos num raio de ϵ ao seu redor;

- *Pontos de borda*: São pontos que estão na fronteira de uma região densa, isto é, eles estão na vizinhança de um objeto central, mas não são pontos centrais;
- *Pontos outliers*: São pontos que estão em uma região de baixa densidade, ou seja, não pertencem a nenhum agrupamento.

Alcançável por densidade diretamente: Um ponto p é alcançável pela densidade diretamente de um ponto q, com respeito a ϵ e a η, se p está na e-vizinhança de q, e q é um ponto central.

Alcançável por densidade: Um ponto p é alcançável por densidade de um ponto q, com respeito a ϵ e a η em um conjunto D, se existe uma cadeia de objetos $\{p_1, ..., p_n\}$, tais que $p_1 = q$, $p_n = p$ e p_{i+1} é alcançável por densidade diretamente de p_i com respeito a ϵ e a η, para $1 \leq i \leq n$, p_i em D.

Para iniciar o algoritmo DBSCAN é necessário receber os parâmetros ϵ e η. Em seguida, para cada ponto da base de dados, começando por um objeto aleatório p, é verificada a vizinhança para encontrar os agrupamentos. Um novo agrupamento é formado com p sendo o centro, se p for um ponto central. Se nenhum ponto é alcançável por densidade a partir de p, o algoritmo visita o próximo ponto da base de dados, p é chamado de ponto de borda. O DBSCAN coleta objetos alcançáveis por densidade diretamente de pontos centrais iterativamente, que pode ocorrer a junção de alguns grupos alcançáveis por densidade. O algoritmo finaliza quando nenhum ponto pode ser alocado a qualquer grupo (BERRY; LINOFF, 2000).

Dois pontos centrais pertencem ao mesmo grupo se a distância entre eles for menor ou igual a ϵ. Quando o ponto de borda está próximo de um ponto central, este é alocado para o mesmo agrupamento do ponto central. Pontos que não foram alocados a nenhum grupo, são chamados de ruídos (*outliers*).

Uma das vantagens do DBSCAN é determinar automaticamente o número de agrupamentos, pois não precisa que o usuário especifique a quantidade de agrupamentos. Outra vantagem é ser menos sensível a ruídos, ele consegue encontrar pontos de ruídos isolados. Por outro lado, ele é sensível aos parâmetros de entrada (ϵ e η), pois é bastante difícil de determinar. Para encontrar os agrupamentos ele depende que usuário informe os parâmetros adequados, fazendo com que muitas vezes perca muito tempo por executar os passos várias vezes até chegar em um resultado satisfatório.

2.3.3 *Expectation Maximization*

Expectation Maximization é um algoritmo de agrupamento que utiliza uma abordagem baseada em probabilidade, onde cada grupo é representado por uma distribuição paramétrica, como a distribuição normal ou distribuição gaussiana. Desta maneira, o algoritmo utiliza modelos para grupos, o que permite otimizar o ajuste entre os dados e esses modelos. Esse algoritmo visa, na prática, encontrar a mistura de gaussianas, de maneira a otimizar a máxima verossimilhança. O EM possui duas etapas principais, a etapa E (*Expectation*) e a etapa M (*Maximization*) (MLADENOVIC et al., 2011).

O primeiro passo do algoritmo consiste em estimar os valores iniciais dos parâmetros do modelo, que são a média e o desvio padrão dos grupos que serão montados pelo algoritmo. O próximo passo consiste no melhoramento do agrupamentos pelas etapas E e M. A etapa E é responsável calcular a probabilidade de uma instância pertencer a um determinado grupo e associá-la ao grupo que obteve a maior probabilidade. A etapa M refina os parâmetros do modelo de mistura de gaussianas, através da probabilidade estimada na etapa anterior, atualizando os parâmetros para que possa maximizar as probabilidades da distribuição dos dados.

Seu funcionamento é bastante semelhante ao k-Médias, já que o número de grupos é informado a *priori*, e os passos de associação de um padrão de teste a um determinado grupo se repetem até que haja convergência, podendo ou não atingir um ótimo global (WITTEN; FRANK, 2005).

2.4 Algoritmos de Classificação

Classificação de dados é o processo de determinar o valor do atributo classe de uma determinada instância através da criação de um modelo de previsão a partir de um algoritmo de aprendizagem. Após esta etapa, o classificador gerado é aplicado a novos padrões para determinar as suas classes. Neste trabalho serão utilizados quatro algoritmos de classificação, descritos nas próximas subseções.

2.4.1 k-NN

O algoritmo k-NN é uma técnica de aprendizado baseado em instâncias que leva em consideração os k vizinhos mais próximos da instância a ser classificada. No k-NN, dado um espaço definido pelos atributos, em que uma instância representa um ponto

nesse espaço, de modo que possa ser calculado a distância entre dois pontos quaisquer através de uma métrica definida (WITTEN; FRANK, 2005). Os vizinhos mais próximos são definidos através da distância Euclidiana, de acordo com a Equação 2.1.

Para duas instâncias $A = a_1, a_2, ..., a_n$ e $B = b_1, b_2, ..., b_n$, onde todos os atributos são numéricos e n é a quantidade de atributos, a distância euclidiana d é definida como (WITTEN; FRANK, 2005):

$$d_{(A,B)} = \sqrt{(a_1 - b_1)^2 + ... + (a_n - b_n)^2} \tag{2.1}$$

Quando são definidos os k vizinhos mais próximos para cada instância não rotulada, cada um dos vizinhos vota em uma classe, a fim de obter um rótulo para o objeto testado. O voto em si é definido pela classe que caracteriza o vizinho.

Um problema da aplicação da distância euclidiana é que, atributos de diferentes escalas de valores podem influenciar o resultado, sendo assim necessário um pré-processamento das bases de dados, transformando as escalas dos atributos para uma escala comum (WITTEN; FRANK, 2005). Todas as bases de dados utilizadas neste trabalho foram aplicadas um pré-processamento, transformando as escalas dos valores numéricos para o intervalo real de zero a um.

2.4.2 Árvore de Decisão

O algoritmo utiliza a estratégia dividir-para-conquistar, ou seja, dado um problema, este é dividido em partes menores, resolvidos, e em seguida, essas soluções são combinadas para produzir o resultado do problema geral. É um método eficiente de construir classificadores que descobre ou prever as classes baseadas nos valores de atributos de uma base de dados. A classificação de uma instância se inicia pela raiz da árvore, e esta é percorrida até que se chegue um nó folha.

As árvores de decisão são compostas por uma folha que representa uma classe e um nó de decisão, que indica um teste que será realizado no valor de um atributo. Cada resposta possível do teste, é determinado de galho, que induzirá para uma sub-árvore ou uma folha. Cada percurso da árvore, ou seja, da raiz até o nó folha é chamado de regra de classificação.

Primeiramente, a fase de aprendizagem da árvore de decisão é a sua construção. Para isto ocorrer, é necessário escolher um atributo para partilhar as instâncias da base de

dados. Logo em seguida, a árvore é estendida atribuindo um nó para cada valor possível do atributo escolhido. Para cada nó, é verificado se todas as instâncias da base de dados que podem ser alocados a esse nó são da mesma classe, então associa-se esta classe ao nó. Se não tiver instâncias que se adequem no valor desse nó, será necessário associar o nó a uma classe determinada a partir de outra informação. Senão, repetir os passos para os demais exemplos que não foram classificados (QUINLAN, 1986).

Depois da fase de construção, temos a fase de reconhecimento. Para classificar uma instância começa pela raiz da árvore, e esta é percorrida até que chegue a um nó folha. Será realizado um teste em cada nó de decisão que irá conduzir a instância a uma subárvore (QUINLAN, 1986).

Uma das principais vantagens é o seu fácil entendimento e interpretação, uma vez que a classificação é adquirido de forma objetiva, facilitando a interpretação dos resultados. Também, a árvore de decisão consegue lidar bem tanto com dados categóricos como dados nominais. Por fim, é um algoritmo robusto que tem um bom desempenho quando se trata de grandes quantidades de informação. Como desvantagens das árvores de decisões, podemos citar o fato dela não poder reutilizar a árvore sem reconstruí-la novamente, de forma eficiente. Outra desvantagem é a criação de uma quantidade volumosa de ramificações, o que proporciona a geração de árvores com ligações prescindíveis e bastante complexas (HANSEN; DUBAYAH; DEFRIES, 1996).

2.4.3 SVM

O algoritmo SVM foi desenvolvido por *Vapnik* e colaboradores (GUNN et al., 1998). O SVM é uma técnica de aprendizado para problemas linearmente separáveis, capaz de encontrar um hiperplano ótimo que maximiza a margem (espaço) de separação entre duas classes. Para isto, as instâncias da base de dados são rotuladas como positivas e negativas e projetados em um espaço de alta dimensão, utilizando uma função chamada de *kernel*. Esta função tem a finalidade de projetar os vetores de características de entrada em um espaço de alta dimensão, para classificação de problemas que são encontrados em espaços originais não linearmente separáveis. Após a projeção, o hiperplano no espaço é otimizado para maximizar a margem de separação entre os dados positivos e negativos (GUNN et al., 1998).

Uma das vantagens deste classificador é a boa capacidade de generalização e que não depende de ajuste de parâmetros, como por exemplo, as redes neurais. Como limitação, é difícil de escolher um bom *kernel*, e também o tempo de treinamento pode ser bem longo

dependendo do número de exemplos e dimensionalidade dos dados.

2.4.4 *Naive Bayes*

O *Naive Bayes* é um classificador supervisionado probabilístico baseado no teorema de Bayes. Tem como o objetivo determinar a classe de maior probabilidade para cada instância a ser classificada. Este classificador é conhecido como ingênuo, pois assume que os atributos são condicionalmente independentes. Para calcular as probabilidades essenciais para a classificação é utilizada a Equação 2.2, que mostra como alterar as probabilidades a priori tendo em conta novas evidências, com o intuito de obter probabilidades a posterior (HAN; PEI; KAMBER, 2011).

$$P(C_i/A) = \frac{P(C_i) \cdot P(A/C_i)}{P(A)}, \tag{2.2}$$

Onde:

- $P(C_i/A)$ é a probabilidade que o padrão pertença a classe C_i dado um exemplo A.
- $P(A/C_i)$ é a probabilidade de um exemplo da classe C_i ter A como o seu conjunto de atributos.
- $P(C_i)$ é a probabilidade a priori da classe C.
- *P(A)* é a probabilidade a priori de um exemplo A.

É um algoritmo simples, que possui um rápido treinamento e uma rápida classificação, pois consegue calcular todas as probabilidades com uma única passagem pela base de dados. Além disso, o *Naive Bayes* não é sensível a características irrelevantes, devido a pequenas alterações nas probabilidades relativas entre classes. Uma das suas principais desvantagens é o fato de assumir que todos os seus atributos são condicionalmente independentes. Apesar dessa hipótese quase sempre ser violada, o método é bastante competitivo e utilizado na prática.

2.5 Testes Estatísticos

Para validar o desempenho dos resultados obtidos em algum problema é necessário realizar testes estatísticos. No contexto estatístico, um resultado é significante se tiver

pouca probabilidade que tenha acontecido por acaso, portanto, procura-se com o uso de um teste estatístico, determinar se a diferença encontrada(se indica se um problema é superior em relação a outro). É necessário formular duas hipóteses: A hipótese nula que é tida como verdadeira na circunstância que está sendo testada, em que o objetivo dos testes de hipóteses é sempre tentar rejeitar a hipótese nula. A hipótese alternativa representa o que se deseja provar ou estabelecer, ela deve ser oposta a hipótese nula.

Para a execução do teste estatístico deve-se definir com antecedência o nível de significância, o qual representa a probabilidade de que o resultado estatístico esteja correto. Comumente utiliza-se o nível de significância $p = 0{,}05$, ou seja, se o valor de p for menor do que esse valor, a hipótese nula é rejeitada, garantindo um grau de confiança maior do que 95%.

Os testes estatísticos são divididos em paramétricos e não paramétricos. Os testes paramétricos presume que os dados obedeçam a uma distribuição normal. Já os testes não-paramétricos não necessitam de requisitos fortes, como a normalidade.

Neste trabalho, o teste utilizado foi o *Friedman* e seu respectivo teste *post-hoc* com o nível de significância de 5%.

2.5.1 Teste de *Friedman*

O teste de *Friedman* é o teste não-paramétrico utilizado para comparar dados amostrais vinculados, ou seja, quando o mesmo indivíduo é avaliado mais de uma vez. É um teste bastante útil quando pretende afirmar a hipótese de que as k observações relacionadas deriva da mesma população (FRIEDMAN, 1937).

Suponha que todas as observações estejam alinhadas de maneira que, n é a quantidade de linhas e k a quantidade de colunas. São atribuídos postos de *1* a k para cada linha da tabela. Em seguida, calcula-se a estatística de *Friedman*, que é dado pela Equação 2.3.

$$x_r^2 = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1), \tag{2.3}$$

onde, R_i é a soma dos postos na coluna i e quando $n < 9$ e $k = 3$ ou 4, é utilizado a distribuição exata; Caso contrário, utilizar a tabela x^2.

Através das tabelas, obtêm-se o p-valor para identificar se existem diferenças significativas entre as amostras (FRIEDMAN, 1937). Caso exista diferença significativa entre

as amostras, é necessária a aplicação de um teste *post-hoc*, a fim de decidir quais grupos são significativamente diferentes uns dos outros (CONOVER, 1980).

3 Trabalhos Relacionados

Este capítulo apresenta diversos trabalhos que realizaram seleção de atributos para classificação, além de outros que utilizaram algoritmos de agrupamento para a seleção de atributos.

3.1 Seleção de Atributos para Classificação

Na literatura encontramos diferentes trabalhos relacionados a seleção de atributos para classificação (YU; LIU, 2003); (CHU et al., 2012); (OLIVEIRA; DUTRA; RENNó, 2005); (PAPPA, 2002); (PEREIRA, 2009); (PARK; KIM, 2015).

Em Yu e Liu (2003) é proposto um método de filtro que pode identificar os atributos mais relevantes, bem como redundância entre os atributos, buscando selecionar os atributos mais correlacionados as classes e menos correlacionados entre si. Para esse fim, com objetivo de analisar a correlação entre os atributos incluindo de classe, utilizou-se uma medida de correlação denominada Incerteza Simétrica. Em seguida, a classificação é realizada utilizando os algoritmos Árvore de decisão e *Naive Bayes*.

Em Chu et al. (2012) fez uma comparação entre vários métodos tradicionais de seleção de atributos para classificação de padrões que diferenciassem pacientes que tivesse a doença de Alzheimer e pessoas com capacidade cognitiva normal. O trabalho tem o objetivo de mostrar o impacto que uma boa seleção de atributos tem na acurácia de um método de classificação. O classificador utilizado para realização dos testes foi o SVM (*Support Vector Machine*).

No trabalho de Oliveira, Dutra e Rennó (2005), aplica métodos de extração e seleção de atributos para classificação de regiões. Os métodos de seleção de atributos são utilizados para reduzir a dimensionalidade dos atributos de texturas obtidos das regiões utilizadas no trabalho. O objetivo do trabalho é mostrar a potencialidade de junção de métodos de extração com métodos de seleção de atributos. Para isso, três diferentes métodos de

seleção já conhecidos na literatura são utilizados, são eles: SFS (*Sequential Forward Feature Selection*), SBS (*Sequential Backward Feature Selection*) e um algoritmo de escolha fixa de subconjuntos de atributos, que consiste em uma busca exaustiva para se encontrar todas as combinações possíveis do conjunto de atributos que será selecionado.

Outro método bastante utilizado na seleção de atributos é o algoritmo genético. Em Pappa (2002) verifica o comportamento de algoritmos genéticos multiobjetivos na seleção de atributos para problemas de classificação, além disso, propõe uma versão modificada do SFS (*Sequential Forward Feature Selection*). Vários pesquisadores desenvolveram diferentes versões de algoritmos genéticos multiobjetivos, variando a maneira como o valor da função de avaliação é atribuído aos indivíduos da população (ZITZLER; THIELE, 1999),(DAS; PANIGRAHI, 2009),(ZHOU et al., 2011).

A maioria dos trabalhos encontrados na literatura trata a seleção de atributos como uma fase do pré-processamento onde os melhores atributos para um conjunto de dados são escolhidos, e o restante dos atributos são descartados e não mais utilizados durante o processo de classificação. Em Pereira (2009) propõe uma estratégia de seleção de atributos diferente, chamada de seleção de atributos do tipo *lazy*. Esse método visa adiar a seleção de atributos até o ponto em que a mesma é submetida ao classificador. Assim como na seleção dinâmica, método proposto nesse trabalho, o objetivo principal é selecionar os melhores atributos para a classificação correta de uma instância em particular, ou seja, cada instância pode ter um conjunto de atributos diferentes e mais adequado para classificá-la, diferente da maioria dos métodos de seleção de atributos. Para avaliar a qualidade de cada atributo na classificação de uma instância é utilizada a entropia para medir quão bem os valores dos atributos de uma instância influenciam no momento de determinar a classe de um determinado padrão. Para validar o método foi utilizado o classificador k-NN e um conjunto de 40 bases de dados da UCI.

Existem poucos trabalhos que utilizam k-NN como método de seleção de atributos. Por exemplo, em Park e Kim (2015) é apresentado um novo método de seleção de atributos baseado em comitês de classificadores compostos por vários k-NN. O novo método se chama SRKNN e funciona de maneira similar ao *Random Forest*, que é construído através da junção de múltiplas Árvores de Decisão.

3.2 Seleção de Atributos Utilizando Algoritmos de Agrupamento

Poucos trabalhos são encontrados utilizando algoritmos de agrupamento para realizar a seleção dos atributos que serão utilizados no processo de classificação. A maioria dos trabalhos apenas utilizam a seleção para realizar um melhor agrupamento dos dados (LAW; FIGUEIREDO; JAIN, 2004); (BOUTSIDIS; MAGDON-ISMAIL, 2013); (BHONDAVE et al., 2014); (SANTHANAM; PADMAVATHI, 2015); (COVõES, 2010).

Law, Figueiredo e Jain (2004), por exemplo, utiliza a seleção de atributos para prover um conjunto de características que será aplicado no algoritmo de agrupamento *Expectation Maximization*. Por outro lado, Boutsidis e Magdon-Ismail (2013) e Bhondave et al. (2014), apresentam dois métodos de seleção de atributos para melhorar a performance de algoritmos de agrupamento.

O trabalho de Boutsidis e Magdon-Ismail (2013) apresenta um algoritmo determinístico de seleção de atributos para o método k-Médias, que reduz a dimensionalidade, selecionando um conjunto de características que otimiza k-Médias em uma representação com baixa dimensionalidade do conjunto de dados. Já Bhondave et al. (2014) utiliza um método de seleção de atributos para melhorar a desempenho do algoritmo de agrupamento *Expectation-Maximization*. O principal objetivo deste trabalho é analisar o impacto dos métodos de seleção de atributos na tarefa de agrupar os dados e não em relação à acurácia de classificação.

Todas as abordagens mencionadas se concentram na aplicação de seleção de atributos métodos de classificação ou agrupamento de dados. Muito pouco tem sido feito para explorar o uso de algoritmos de agrupamento no processo de seleção de atributos. Em Santhanam e Padmavathi (2015), os autores utilizam o método k-Médias junto com o algoritmo genético para selecionar os atributos mais relevantes da base de dados Pima Indians Diabetes do repositório UCI. O k-Médias é utilizado para remover os ruídos da base de dados, enquanto o algoritmo genético seleciona o conjunto de atributos mais relevantes.

Um dos poucos trabalhos existentes é o de Covões (2010), que propõe variantes de um algoritmo que realiza a seleção de atributos por meio de algoritmos de agrupamento, que são utilizados para identificar a redundância entre os atributos de uma base de dados. Este trabalho tem como objetivo desenvolver variantes do algoritmo Filtro Silhueta Simplificado (COVõES; HRUSCHKA, 2009), que consiste em um algoritmo que utiliza

agrupamento de dados para agrupar atributos por semelhança, para poder analisar quais atributos apresentam informações redundantes, removendo-os do conjunto de dados. Neste trabalho são realizadas diferentes variações, como as medidas de correlação e os critérios de seleção de atributos.

4 Nova Abordagem de Seleção de Atributos Baseado em Algoritmos de Agrupamento

Foram desenvolvidas duas versões da nova abordagem de Seleção de Atributos apresentada neste trabalho. A primeira versão desenvolvida foi a Semidinâmica. No início deste capítulo, será apresentado sua arquitetura geral e o seu funcionamento. Na arquitetura geral será apresentado uma ilustração da sua estrutura e a formalização do método proposto para a versão Semidinâmica. Logo após, será apresentado o algoritmo demonstrando os principais passos a nível de implementação.

A segunda versão, chamada de versão Dinâmica, é apresentada logo após, mostrando as principais diferenças e alterações que foram necessárias para o seu desenvolvimento.

4.1 Arquitetura Geral da Versão Semidinâmica

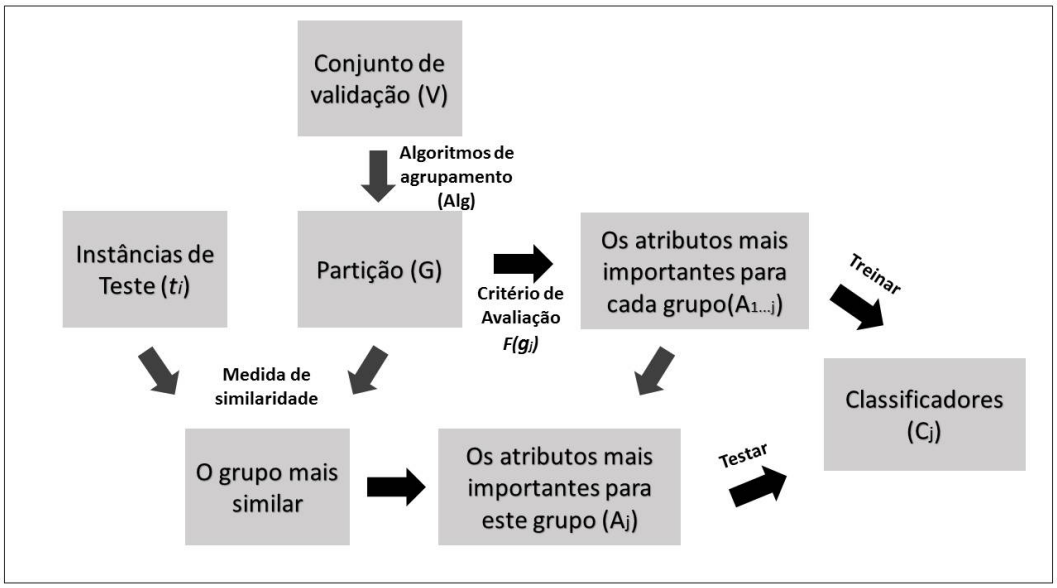
Enquanto a seleção de características tradicional seleciona os melhores atributos para um conjunto de dados como um todo, a seleção semidinâmica, proposta neste trabalho, seleciona os melhores atributos de uma instância individualmente. O principal objetivo do método proposto é ter apenas os atributos mais relevantes para cada instância ou para um grupo de instâncias para ser utilizado na classificação. O funcionamento geral do método proposto pode ser visualizado na Figura 2.

Para entender melhor o método proposto, suponha que seja B uma base de dados, composta por I instâncias e A atributos. As instâncias são divididas em 3 conjuntos, treinamento $TR = \{tr_1, tr_2, ..., tr_{ntr}\}$, validação $V = \{v_1, v_2, ..., v_{nv}\}$ e teste $T = \{t_1, t_2, ..., t_{nt}\}$.

Os passos do método proposto podem ser descritos a seguir:

1. Um algoritmo de agrupamento Alg é aplicado a V, que é responsável pelo agrupa-

Figura 2: Ilustração da estrutura geral da Versão Semidinâmica do método proposto.



mento das instâncias em grupos similares, através da Equação 4.1:

$$G = Alg(V), \tag{4.1}$$

onde, $G = \{g_1, ..., g_j\}$ é criada, em que j é o número de grupos formados;

2. Além disso, cada grupo g_j possui um ponto central p_j, que pode ser definido como o centro de massa do grupo.

   2.1. Para cada grupo g_j é aplicado uma função $F(g_j)$ como critério de avaliação, para definir a importância deste grupo. Esta função vai fornecer E, onde $E = \{e_1, .., e_n\}$. O valor e_i para cada atributo i de A, usando como base as instâncias de g_j, define a importância deste atributo baseado no critério utilizado por F.

   2.2. Baseado neste critério de avaliação, todos os atributos são ordenados através da Equação 4.2:

$$R_j = rank(F(g_j)), \tag{4.2}$$

   onde, os N atributos mais bem posicionados no *ranking* são selecionados para cada grupo g_j.

   2.3. Os N_{g_j} atributos são selecionados para representar g_j, formando o conjunto A_j, como mostra a Equação 4.3:

$$A_j = select(R_j, N), \tag{4.3}$$

onde, $A_j = \{att_1, att_2, ..., att_N\} | A_j \subset A$;

3. Depois de selecionar os atributos mais importantes para cada grupo g_j, foram treinados j classificadores C, um para cada grupo g_j de G, através da Equação 4.4:

$$C_j = train(TR, A_j) \tag{4.4}$$

4. Quando testamos uma nova instância t_i, esta é comparada com a partição G fornecida pelo algoritmo de agrupamento Alg, a fim de definir o grupo ao qual t_i pertence. Esta definição ocorre através da Equação 4.5, que consiste no cálculo da distância entre t_i aos pontos centrais p_y de cada grupo y de G, sendo assim, a instância t_i pertencerá ao grupo g_y, que possua a menor distância para o seu ponto central.

$$d_i = \min_{y=1}^{j}(dist(t_i, p_y)); \tag{4.5}$$

   4.1. Então t_i será classificado pelo classificador C_y e utilizando o sub-conjunto de atributos A_y, através da Equação 4.6:

$$R_{t_i} = test(t_i, C_y, A_y), \tag{4.6}$$

   onde, R é a acurácia obtida quando um conjunto de Testes T é aplicado ao classificador C;

Como podemos observar, os atributos selecionados podem variar drasticamente, dependendo do grupo ao qual a instância pertence.

4.1.1 Funcionamento

O Algoritmo 1 apresenta os principais passos da Seleção de Atributos Semidinâmica.

A entrada do algoritmo consiste no conjunto de dados que serão utilizados e é dividida em 3 subgrupos: V, TR e T (Linha 2). O primeiro passo é a utilização de um algoritmo de agrupamento Alg para o V (Linha 3). Uma vez que a partição G é criada, todos os casos pertencentes a um determinado grupo são analisados por um critério de avaliação $F(g_j)$ (Linhas 6-8).

Algoritmo 1 Seleção Semidinâmica de atributos

```
1:  procedure SELEÇÃO SEMIDINÂMICA DE ATRIBUTOS
2:      Entrada: B =(V, TR, T)
3:      G ← Alg(V)
4:      para cada grupo g_j faça
5:          para cada instância v_i em g_j faça
6:              para cada atributo a_i em v_i faça
7:                  E ← F(g_j)
8:              fim para
9:              R_j = rank(E)
10:             A_j = select(R_j, N)
11:         fim para
12:     fim para
13:     para cada grupo g_j faça
14:         C_j ← Train(TR, A(g_j))
15:     fim para
16:     para cada padrão de teste t_i em T faça
17:         Escolhe grupo g_j através de d_i = dist(t_i, p_y)
18:         acc ← Test(C(g_j), T, A(g_j))
19:     fim para
20: fim procedure
```

Com a aplicação de $F(g_j)$, os melhores atributos são selecionados para cada grupo g_j. Estes atributos formam o conjunto A_j, que consiste nos melhores atributos do conjunto R_j, que possui os atributos ordenados de acordo com sua importância (Linhas 9-10).

Em seguida, j classificadores C são treinados utilizando o conjunto de dados TR, um para cada grupo g_j, levando em consideração apenas os atributos selecionados A_j para cada grupo g_j (Linhas 13-15).

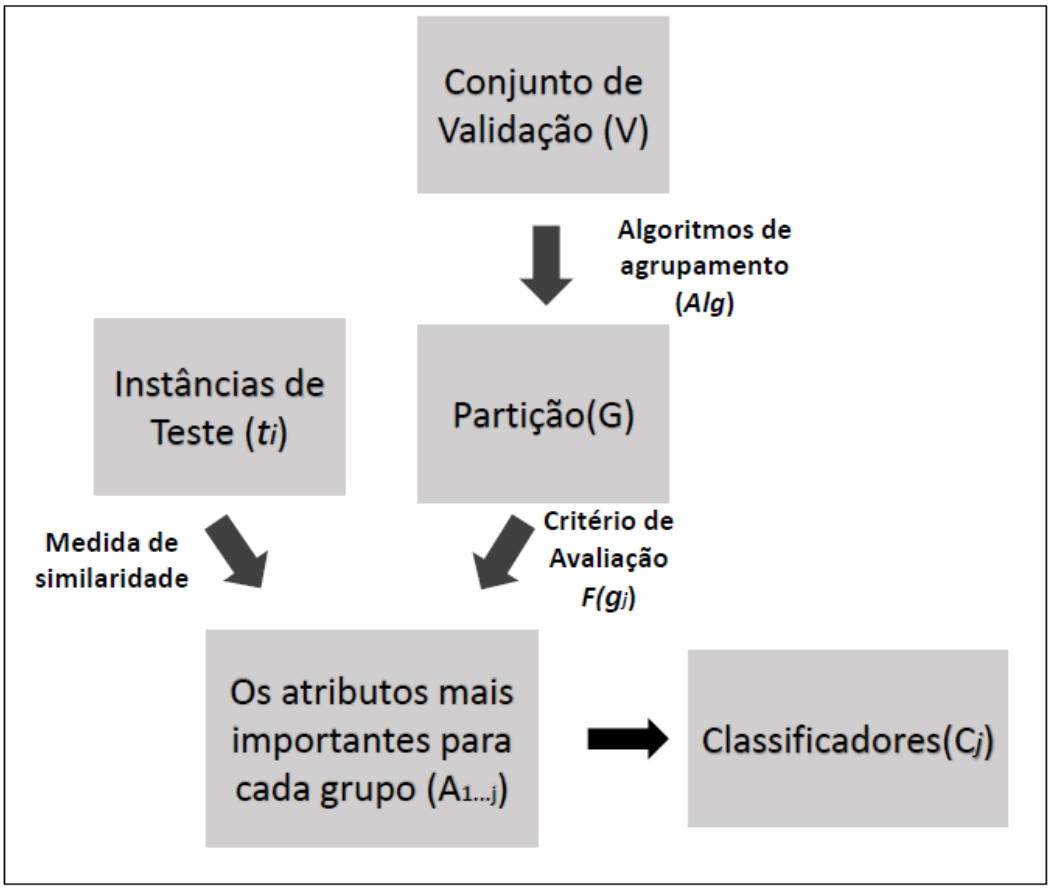
No próximo passo, para cada padrão de teste t_i, é definido a qual grupo g_j esse padrão pertence, e em seguida seus atributos mais relevantes são selecionados (Linhas 16-17). O classificador correspondente é aplicado para classificar o padrão de teste. A saída do algoritmo é a acurácia dos métodos de classificação (Linha 18).

4.2 Arquitetura Geral da Versão Dinâmica

Uma nova versão da abordagem de Seleção de Atributos proposta, chamada de Versão Dinâmica, foi desenvolvida com o objetivo de comparar com a versão anterior. Essa versão consiste em uma adaptação da versão anterior, para dar uma maior dinamicidade no processo de seleção de atributos. Nesta seção será apresentada a ideia central desta nova versão e o seu funcionamento. A Figura 3 apresenta a estrutura geral da versão Dinâmica

do método proposto.

Figura 3: Ilustração da estrutura geral da Versão Dinâmica do método proposto.



Seu funcionamento inicial é o mesmo da versão anterior, onde os atributos mais importantes para cada grupo são definidos através de um critério de avaliação aplicado a cada grupo.

Sua principal diferença está na maneira com que os atributos mais importantes para cada instância de teste são definidos. Na versão anterior, a instância de teste era comparada com todos os grupos, visando definir a qual grupo a mesma pertence. Assim, a instância teria os mesmos atributos do grupo mais próximo.

A ideia central da versão Dinâmica é que a instância de teste possua atributos de todos os grupos, mas de maneira proporcional a sua distância com o ponto central de cada grupo, definindo a probabilidade de a mesma pertencer a cada um dos grupos. Ou seja, quanto mais próxima a instância de teste estiver de um grupo, mais atributos desse grupo ela terá.

Quando testamos uma nova instância t_i, esta é comparada com a partição G fornecida pelo algoritmo de agrupamento Alg, através de uma distância d_j, através da Equação 4.7:

$$d_j = dist_{y=1}^{j}(t_i, p_y). \tag{4.7}$$

Uma função P é aplicada à d_j, para definir a probabilidade $Prob_j$ de t_i pertencer a g_j, através da Equação 4.8:

$$Prob_j = P(t_i, g_j). \tag{4.8}$$

Assim, os novos atributos selecionados NV_i para a instância de teste t_i, são obtidos através dos atributos mais importantes A_j para cada grupo g_j mantendo a proporção estabelecida em $Prob_j$, através da Equação 4.9:

$$NV_i = select(A_j, Prob_j). \tag{4.9}$$

Portanto, para a classificação de t_i será utilizado NV_i como o seu novo conjunto de atributos.

4.2.1 Funcionamento

O Algoritmo 2 apresenta o processo de seleção dos atributos mais importantes para cada instância de teste t_i da Versão Dinâmica.

Algoritmo 2 Versão Dinâmica

```
1:  para cada padrão de teste t_i em T faça
2:      para cada grupo g_j faça
3:          Calcular a distância de T_i para p_j através de d_j = dist(t_i, p_j)
4:      fim para
5:      para cada grupo g_j faça
6:          Prob_j = P(t_i, d_j)
7:      fim para
8:      para cada atributo a de A_j faça
9:          NV_i = select(A_j, Prob_j)
10:     fim para
11: fim para
```

A Versão Dinâmica possui o mesmo funcionamento da versão anterior linhas 1 à 12 do Algoritmo 1. A principal diferença está na forma em que os atributos mais importantes para uma dada instância de teste são selecionados.

Para que uma instância de teste t_i seja classificada, é necessário definir os atributos

mais importantes. Para isso, sua distância para os grupos g_j é calculada através de d_j (Linhas 2 - 4).

Após o cálculo da distância da instância de teste t_i para ponto central p_j, a probabilidade $Prob_j$ da instância t_i pertencer a g_j é calculada (Linhas 5 - 7). Esse valor de probabilidade servirá para estabelecer quantos atributos de cada grupo serão selecionados para a instância t_i, assim, quanto mais próximo t_i estiver de um determinado grupo, mais atributos daquele grupo t_i possuirá.

Uma vez que todas as probabilidades forem calculadas, um novo conjunto de atributos NV_i, para a instância t_i será selecionado (Linhas 8 - 10). Esse novo conjunto consiste nos atributos mais importantes, que foram obtidos através da seleção de uma proporção dos atributos de cada grupo, estabelecida através de $Prob_j$.

4.3 Considerações Finais do Capítulo

Esse Capítulo apresentou a ideia central da abordagem proposta através da apresentação de duas versões desenvolvida, uma Semidinâmica e uma Dinâmica. Foi apresentado a estrutura geral e o funcionamento de cada versão, bem como o principais passos em forma de pseudo-código.

Como apresentado, a principal diferença entre as duas versões é a maneira com que os atributos mais importantes são selecionados. Enquanto a versão Semidinâmica seleciona os melhores atributos apenas do grupo mais próximo, a versão Dinâmica seleciona atributos de todos os grupos, obedecendo uma proporção estabelecida pela distância entre eles e a instância de teste.

5 Análise Experimental

Na análise experimental serão apresentado os parâmetros mais importantes, bem como a adaptação do método proposto para a realização dos experimentos demonstrados no Capítulo 6.

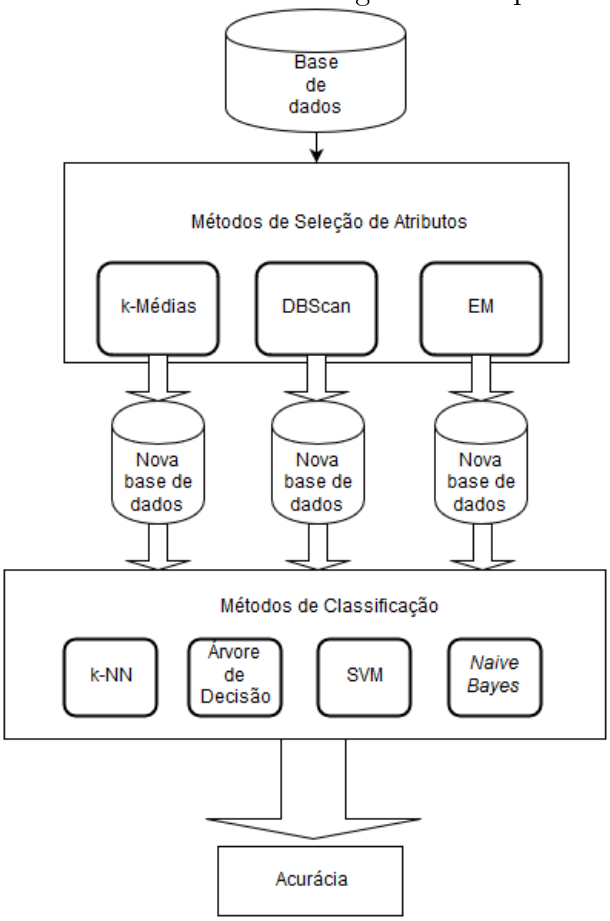
Visando analisar a viabilidade da Versão Semidinâmica, uma análise experimental foi necessária. A Figura 4 ilustra a estrutura metodológica utilizada na análise empírica realizada. Como pode ser observado, a fim de analisar o desempenho do método proposto com diferentes critérios de agrupamento, usamos três algoritmos de agrupamento, que são k-Médias (KM), DBSCAN (DB) e *Expectation Maximization*. Esses algoritmos usam diferentes técnicas de agrupamento permitindo uma análise do desempenho do método proposto com diferentes tipos de algoritmos, baseados na distância, na densidade e na probabilidade, respectivamente.

Um parâmetro importante para o método proposto é o tamanho dos subconjuntos de atributos a serem escolhidos, também referida como taxa de redução. Esses tamanhos podem ser variáveis (uma quantidade diferente para cada grupo) ou fixo, com o mesmo tamanho. Neste trabalho, iremos analisar com a taxa de redução fixa. Para uma análise inicial, utilizamos três taxas de redução, 25%, 50% e 75%, dos atributos para todos os grupos.

Um outro parâmetro importante no método proposto é o critério de avaliação dos atributos (função F). Esse critério é essencial para o funcionamento deste método. Neste trabalho foi utilizado como critério de avaliação o coeficiente de correlação de *Pearson* (FIGUEIREDO et al., 2014), com o intuito de analisar quais são os atributos mais correlacionados com o atributo classe. Uma vez que a partição G é criada, todos os casos pertencentes a um determinado grupo são analisados pelo o cálculo do coeficiente de correlação de *Pearson*.

Os valores do coeficiente de correlação são armazenados em um vetor. Através desse vetor, os atributos são ranqueados e os N melhores atributos são selecionados, permitindo

Figura 4: Estrutura Metodológica dos Experimentos.



que cada instância seja representada por um subconjunto de atributos, apenas os mais correlacionados com o atributo classe.

Cada subconjunto do grupo do atributo selecionado foi armazenado em uma matriz, em que as linhas representam os grupos e as colunas representando os atributos mais importantes para cada grupo.

Depois que a partição é definida, com os subconjuntos de atributos de cada grupo podemos criar uma nova base de dados. Esta nova base de dados consiste nas mesmas instâncias porém contendo os atributos mais relevantes para cada grupo de instâncias, podendo assim aplicá-las a problemas de classificação. A fim de realizar isso, foram utilizados os métodos de classificação mais presentes na literatura, que são: k-NN, árvore de decisão, SVM e *Naive Bayes*.

Após a descoberta do melhor algoritmo de agrupamento como seletor de atributos para a Versão Semidinâmica do método proposto, a mesma foi comparada com a sua Versão Dinâmica, desenvolvida apenas com o melhor algoritmo de agrupamento encontrado, mantendo as mesmas taxas de redução de atributos. Para a análise empírica da Versão

Dinâmica foram utilizados os mesmos parâmetros da primeira versão do método proposto.

5.1 Bases de Dados Utilizadas

Os experimentos deste trabalho foram realizados com 15 bases de dados para serem utilizadas no método proposto. Um pré processamento foi aplicado a todas as bases, visando obter apenas valores reais normalizados entre [0,1]. Todas as bases de dados foram obtidas no repositório UCI, com diferentes características e contendo diferentes números de atributos e instâncias. As próximas subseções apresentam uma breve descrição sobre cada base. Tabela 1 apresenta as características das bases de dados descritas a seguir, são elas: número de atributos, instâncias e atributos classes.

Tabela 1: Bases de Dados Usadas nos Experimentos.

ID	**Base de dados**	**Qtd. de Instâncias**	**Qtd. de Atributos**	**Qtd. de Classes**
b1	Ionosphere	351	34	2
b2	Gaussian	60	600	3
b3	Lung Cancer	32	56	3
b4	Breast Cancer(Prognostic)	198	34	2
b5	Spam	4601	58	2
b6	Arrhythmia	452	279	13
b7	Parkinsons	195	23	2
b8	Jude	248	985	6
b9	Libras Movement	360	91	15
b10	Simulated	60	600	6
b11	Micromass	931	1301	2
b12	ADS	3279	1559	2
b13	Semeion Handwritten Digit	1593	256	2
b14	Protein	583	121	5
b15	Hill-Valley	606	101	2

- ADS

  Este conjunto de dados representa um conjunto de possíveis anúncios em páginas da Internet. A tarefa consiste em prever se uma imagem é ou não de um anúncio publicitário. A ADS é composta por 1559 atributos, sendo dois os atributos classes, e 3279 instâncias.

- Arrhythmia

  Esta base de dados contém 279 atributos, dos quais 206 são valores numéricos e o restante são valores categóricos, e 13 classes. O objetivo é distinguir entre a presença

e a ausência de arritmia cardíaca e classificá-la em um dos 16 grupos. A classe 01 refere-se a eletrocardiogramas normais, da classe 02 à 15 refere-se a diferentes tipos de arritmia e a classe 16 ao restante que não foi classificado.

- Breast Cancer Wisconsin (Prognostic)

  A base de dados Breast Cancer é representado por 34 atributos, sendo 2 atributos classes, e 198 instâncias. Cada registro representa o acompanhamento de dados para um caso de câncer de mama, incluído apenas os casos que apresentam câncer de mama invasivo e sem evidência de metástases.

- Gaussian

  A base de dados Gaussiana possui 600 atributos, 60 instâncias (genes) e 3 classes (0,1,2), sendo todos os atributos numéricos. É importante enfatizar que a quantidade de instâncias é dividido igualmente para cada classe, ou seja, existem 20 padrões para cada classe.

- Hill-Valley

  A base de dados Hill-Valley é composta por 101 atributos, 2 atributos classe e 606 instâncias. Cada registro representa 100 pontos em um gráfico bidimensional, que quando plotado, em ordem (de 1 a 100). Os pontos irão criar tanto um Hill (colina) ou um Valley (vale).

- Ionosphere

  Esta base de dados contêm 351 instâncias, dividas em duas classes: *bad* (fraco) e *good* (bom). Os padrões representam medições recebidas por 16 antenas de alta frequência de um sistema de radar. Cada padrão é formado por 17 pulsos complexos, e existem duas características para cada pulso (coeficientes do número complexo), totalizando 34 atributos. Quando o sinal recebido é fraco, significa que o sinal passa direto pela ionosfera, quando o sinal retornado é forte, significa que o sinal é refletido de volta por um objeto qualquer.

- Jude

  A base Jude foi gerada a partir de dados de células de leucemia de expressão genética. Existem 248 amostras de células de leucemia (instâncias), 985 genes (atributos), que descrevem o nível de expressão de cada gene para cada amostra de célula, e 6 atributos classe.

- Libras Movement

  A base Libras possui 360 instâncias, divididas igualmente em 15 classes, que representam os movimentos da mão na língua brasileira de sinais - LIBRAS. Cada classe representa um tipo de movimento de mão específico. Os atributos usados nas classes são todos numéricos e descrevem as curvas dos movimentos de mão que são mapeadas pelo algoritmo.

- Lung Cancer

  Lung Cancer foi gerada a partir de imagens de diagnósticos de câncer de pulmão. Os dados descreveram 3 tipos de câncer pulmonar patológico. Ela é composta por 56 atributos, um deles representa o identificador do paciente e outro o atributo classe, e 32 instâncias.

- Micromass

  A Micromass é uma base de dados para explorar abordagens de aprendizagem de máquina para a identificação de microorganismos a partir de dados de espectrometria de massa. Sua composição é de 1300 atributos numéricos, sendo dois valores de atributos classes, e 931 instâncias.

- Parkinsons

  A base de dados Parkinsons é formada por uma variedade de medições biomédicas de vozes. Composta por 23 atributos, onde cada um representa uma voz um indivíduo e 195 instâncias que corresponde as gravações de vozes.

- Protein

  A base de dados Protein é formada por 121 atributos e 583 instâncias. Essa base tem como objetivo prever as estruturas secundárias de proteínas globulares. Essas estruturas são previstas através na análise de sequências de aminoácidos.

- Spam

  A Spam é um base de dados composto por uma coleção de e-mails spam, que foram obtidas pelos e-mails dos próprios autores da base, e vinheram da sua caixa de e-mails pessoais. Esta base possui 58 atributos e 4601 instâncias.

- Simulated

  A Simulated é um banco de dados sintético que simula dados de microarranjos, criada para testar os algoritmos de aprendizado de máquina na análise da expressão

do gene. A base de dados Simulated, ao contrário da base gaussian, é desbalanceada, possuindo 8,12,10,15,5,10 padrões em cada uma de suas 6 classes, tendo um total de 60 instâncias e 600 atributos.

- Semeion Handwritten Digit

  A base Semeion Handwritten Digit possui 1593 dígitos escritos à mão que foram digitalizados de cerca de 80 pessoas. Estes dígitos foram estendidos em uma caixa retangular com medidas 16×16 em escala cinza de 256 valores. Esta base é formada com 1593 instâncias e 256 atributos.

5.2 Métodos e Materiais

Todos os algoritmos desta análise, foram desenvolvidos usando linguagem MATLAB. Para a execução desta análise empírica, foi necessário definir os principais parâmetros usados nos algoritmos de agrupamento.

Para o algoritmo k-Médias, a definição do valor de k (número de grupos desejados) foi escolhido considerando o número de classes apresentado no conjunto de dados. Por exemplo, o conjunto de dados *Micromass* tem apenas dois valores para o rótulo de classe, 0 ou 1. Portanto, o valor escolhido para k é 2. A mesma metodologia foi aplicada ao algoritmo *Expectation Maximization*. Para a implementação do *DBSCAN*, é necessário definir o valor de *Eps* (Raio de vizinhança de um ponto) e *MinPts* (número mínimo de pontos dentro do raio). Os valores destes parâmetros são definidos empiricamente através da utilização de uma análise inicial. Várias análises foram conduzidas e os parâmetros escolhidos variou de acordo com o conjunto de dados utilizado. Os parâmetros foram escolhidos com o objetivo de gerar os melhores grupos. Para avaliar a qualidade dos agrupamentos, foi aplicado nos resultados o índice *Davies-Bouldin*(DB). Para os algoritmos de classificação, foram executados com os parâmetros *default* do MATLAB.

Para os experimentos, tivemos 12 configurações para cada conjunto de dados, uma vez que foram aplicados três algoritmos de agrupamento para fornecer os conjuntos de dados reduzidos e quatro algoritmos de classificação para avaliar o desempenho dos métodos de redução de atributos. Além disso, com o intuito de obter uma melhor estimativa nas taxas de precisão, um método de validação cruzada de 10 vezes é aplicado para todos os algoritmos de classificação.

Assim, todos os resultados de precisão apresentados se referem à média das ordena-

ções de 10 conjuntos de testes diferentes. Para os métodos de seleção de atributos foram realizadas 10 execuções, tendo um total de 100 valores de precisão. Quando se deseja realizar a comparação entre diferentes métodos de seleção, cada um dos seus 100 resultados são comparados, ordenando-os entre si, de maneira que o maior valor receba 1 e o menor valor da comparação entre N métodos receba o valor N. Em seguida é calculada a média desses valores, e são esses os resultados que estão presentes neste trabalho. Ou seja, quanto menor a média da ordenação, melhor será o método que está sendo comparado.

Para estabelecer a ordenação é necessário analisar comparativamente os valores de precisão de cada método que está sendo comparado, esses valores estão presentes no Apêndice A.

Após obter os resultados dos métodos de classificação, será realizado uma análise comparativa da abordagem proposta com os outros métodos de redução de dimensionalidade, como PCA, LDA e seleção randômica.

A fim de validar o desempenho do método proposto de uma forma mais significativa, será aplicado o teste de *Friedman* e o seu respectivo *post-hoc*. Uma vez que ambos consistem em um teste estatístico não-paramétrico que é adequado para comparar o desempenho de diferentes algoritmos de aprendizagem quando aplicados para várias bases de dados. O teste de *Friedman* será usado para identificar se os resultados obtidos nas diferentes análises possuem diferença estatística. Caso isso aconteça, foi aplicado um teste *post-hoc* de *Friedman* para realizar uma comparação par-a-par entre todos os métodos, e poder assim, escolher a melhor configuração para a abordagem proposta. Nestes testes, serão considerados todas as 10 execuções, com todos os 10 diferentes *folds*, totalizando 100 observações de acurácias.

6 Resultados

Este capítulo apresenta e discute os resultados dos experimentos obtidos pelo método proposto. Os resultados serão apresentados através de quatro análises, divididas nas seções seguintes.

6.1 Considerações Iniciais

A primeira análise foi feita com diferentes proporções de seleção de atributos, 25%, 50% e 75%, para os três algoritmos de agrupamento utilizados. O objetivo desta análise é descobrir qual a melhor proporção de atributos para obter melhor acurácia para cada método de agrupamento.

A segunda análise visa comparar os diferentes algoritmos de agrupamento utilizados na Seleção Semidinâmica de Atributos e determinar qual deles melhor se adéqua ao método proposto. Esta análise foi feita utilizando a melhor proporção de seleção de atributos descoberta na análise anterior.

Após descobrir o melhor algoritmo de agrupamento para a versão Semidinâmica do método proposto, foi desenvolvida uma nova versão utilizando o mesmo, chamada versão Dinâmica. A terceira análise consiste em comparar a versão anterior com a versão Dinâmica, com as proporções utilizadas anteriormente, visando obter a melhor entre elas para comparar posteriormente com outros métodos de redução de dimensionalidade.

Na quarta análise foi realizada uma comparação entre os conhecidos métodos de redução de dimensionalidade com a versão que obteve um melhor desempenho no momento anterior. Para realizar essa comparação, foram escolhidos métodos de seleção e extração de atributos, são eles: Seleção randômica, PCA e LDA. O objetivo desta análise é detectar quais técnicas obtiveram melhor acurácia em sistemas de classificação.

Nas próximas seções serão apresentados os resultados das análises de desempenho e seus respectivos testes estatísticos. Para obter os resultados estatísticos foi aplicado o teste

de *Friedman* e seu respectivo *post-hoc*, a fim de validar as análises citadas anteriormente.

6.2 Análise dos Percentuais de Seleção de Atributos

Essa seção tem por objetivo descobrir qual a melhor proporção de seleção de atributos para a abordagem proposta em cada um dos algoritmos de agrupamento utilizados. Nas tabelas a seguir, cada coluna representa o resultado para cada proporção, onde o melhor valor da ordenação (menor valor), para cada base de dados está destacado nas células cinzas. A última linha representa a quantidade de bases que tiveram melhores resultados, para cada proporção.

6.2.1 Análise de Proporção do k-Médias

A Tabela 2 apresenta a média das ordenações e o desvio padrão obtidos através da aplicação do k-Médias como seletor de atributos da Seleção Semidinâmica, com diferentes taxas de seleção, para cada conjunto de dados. O valor apresentado em cada célula consiste na média de todos os resultados obtidos, pelos quatro classificadores, para a respectiva base de dados.

Baseado na Tabela 2, analisando comparativamente, podemos observar que o melhor resultado para o k-Médias foi quando se utilizou 50% como taxa de seleção de atributos, visto que, esta proporção alcançou melhores resultados em 9 das 15 bases de dados utilizadas. Já a proporção de 25% obteve apenas 2 melhores resultados e a proporção de 75% obteve 4 melhores resultados.

A Tabela 3 apresenta os resultados da aplicação deste teste de *Friedman* para os experimentos da Tabela 2. Os valores que estão destacados nas células cinza são as performances estatisticamente significativas. Quando o p-valor é menor que *0,05*, significa que houve diferença estatística e foi necessário consultar as médias, que estão dispostas no Apêndice A, para determinar o vencedor da comparação, ou seja, aquele que tiver a maior acurácia média. Além disso, a última célula da tabela apresenta o número de resultados estatisticamente significativos. As demais células da última linha apresentam o número de vitórias/derrotas para a comparação da respectiva coluna.

Tabela 2: Média das ordenações e desvio padrão para diferentes proporções utilizando o k-Médias como seletor de atributos.

	k-Médias		
ID	**25%**	**50%**	**75%**
b1	2,04±0,758	**1,95±0,780**	2,00±0,735
b2	1,92±0,487	**1,89±0,482**	2,17±0,524
b3	**1,98±0,465**	2,01±0,732	2,00±0,445
b4	1,99±0,787	2,08±0,751	**1,92±0,758**
b5	2,14±0,774	**1,81±0,903**	2,04±0,710
b6	2,19±0,760	**1,72±0,770**	2,07±0,755
b7	2,03±0,713	**1,90±0,754**	2,06±0,768
b8	1,99±0,210	2,02±0,238	**1,98±0,200**
b9	2,67±0,524	**1,51±0,644**	1,81±0,687
b10	**1,96±0,432**	2,00±0,431	2,03±0,423
b11	1,92±0,719	**1,62±0,710**	2,44±0,722
b12	1,91±0,746	**1,79±0,758**	2,29±0,799
b13	2,52±0,675	1,86±0,754	**1,61±0,694**
b14	2,29±0,766	1,87±0,742	**1,82±0,786**
b15	2,02±0,379	**1,89±0,337**	2,08±0,396
Quant.	2	9	4

Tabela 3: Resultados do Teste de *Friedman* (p-valor) para diferentes proporções utilizando o k-Médias como seletor de atributos.

	Friedman	**25% - 50%**	**25% - 75%**	**50% - 75%**
b1	0,213	-	-	-
b2	**0,001**	0,711	**0,001**	**0,001**
b3	0,761	-	-	-
b4	**0,046**	0,298	0,591	**0,036**
b5	**0,003**	**0,001**	0,992	**0,001**
b6	**0,001**	**0,001**	0,196	**0,001**
b7	**0,001**	**0,040**	0,907	**0,001**
b8	0,090	-	-	-
b9	**0,001**	**0,001**	**0,001**	**0,001**
b10	0,160	-	-	-
b11	**0,001**	**0,001**	**0,001**	**0,001**
b12	**0,001**	0,215	**0,001**	**0,001**
b13	**0,001**	**0,001**	0,766	**0,001**
b14	**0,001**	**0,001**	0,091	**0,001**
b15	**0,001**	**0,001**	**0,001**	**0,001**
Placar	11	1/7	4/1	9/2

Podemos observar que a diferença de desempenho foi estatisticamente significante para a maioria das bases de dados (11 das 15 bases de dados). Além disso, quando comparamos as proporções entre si, pode-se observar que 50% obteve melhores resultados do que 25% e 75% em 7 e 9 bases respectivamente, perdendo em apenas 1 base para 25% e 2 bases para 75%. Podemos observar também, que 25% é estatisticamente melhor que 75% em 4 das 5 bases em que houve diferença estatística na comparação entre elas. Assim, pode-se concluir que para o k-Médias, a melhor proporção foi a de 50% dos atributos selecionados.

6.2.2 Análise de proporção do DBSCAN

A mesma análise das proporções foi realizada para o DBSCAN. A Tabela 4 apresenta os resultados obtidos para cada proporção quando o DBSCAN foi utilizado na seleção de atributos. É possível observar pela média das ordenações, assim como para o k-Médias, a proporção que obteve melhores resultados foi a de 50%, obtendo um menor valor em 8 das 15 base de dados.

Tabela 4: Média das ordenações e desvio padrão para diferentes proporções utilizando o DBSCAN como seletor de atributos

	DBSCAN		
ID	**25%**	**50%**	**75%**
b1	**1,92±0,779**	2,02±0,761	2,05±0,763
b2	1,85±0,623	**1,82±0,578**	2,32±0,617
b3	**1,90±0,669**	2,06±0,682	2,02±0,652
b4	1,98±0,781	**1,97±0,798**	2,04±0,761
b5	2,39±0,812	**1,65±0,704**	1,95±0,735
b6	2,24±0,760	**1,66±0,776**	2,08±0,729
b7	2,03±0,757	2,03±0,703	**1,89±0,800**
b8	1,96±0,211	2,08±0,323	**1,95±0,195**
b9	2,42±0,596	**1,22±0,511**	2,34±0,619
b10	**1,94±0,452**	2,09±0,493	1,96±0,447
b11	2,25±0,747	2,24±0,743	**1,49±0,635**
b12	1,76±0,747	**1,67±0,656**	2,56±0,687
b13	2,34±0,578	**1,21±0,609**	2,44±0,593
b14	2,46±0,644	1,75±0,721	**1,68±0,699**
b15	2,02±0,335	**1,95±0,291**	2,02±0,329
Quant.	3	8	4

A Tabela 5 apresenta os resultados do teste estatístico aplicado às comparações da Tabela 4. Podemos observar que a diferença de desempenho foi estatisticamente significante para todas as bases de dados. Analisando cada comparação, é possível perceber que a proporção de 50% obteve os melhores resultados, pois a mesma ganhou de 25% e 75%

em 6 e 7 bases respectivamente, perdendo apenas em 5 para 25% e 4 para 75%.

Tabela 5: Resultados do Teste de *Friedman* (p-valor) para diferentes proporções utilizando o DBSCAN como seletor de atributos.

	Friedman	**25% - 50%**	**25% - 75%**	**50% - 75%**
b1	**0,002**	**0,001**	0,106	0,300
b2	**0,001**	**0,001**	**0,001**	**0,001**
b3	**0,026**	**0,024**	0,133	0,769
b4	**0,036**	0,211	0,671	**0,030**
b5	**0,001**	**0,001**	**0,001**	0,303
b6	**0,001**	**0,001**	0,061	**0,001**
b7	**0,003**	0,987	**0,013**	**0,008**
b8	**0,001**	**0,001**	0,866	**0,001**
b9	**0,001**	**0,001**	0,559	**0,001**
b10	**0,001**	**0,001**	0,970	**0,001**
b11	**0,001**	0,994	**0,001**	**0,001**
b12	**0,001**	0,546	**0,001**	**0,001**
b13	**0,001**	**0,001**	0,945	**0,001**
b14	**0,001**	**0,001**	**0,001**	0,649
b15	**0,001**	**0,001**	0,433	**0,001**
Placar	15	5/6	2/4	7/4

6.2.3 Análise de proporção do EM

A mesma análise das proporções realizada nas subseções anteriores, também foi executada para o EM. A Tabela 6 apresenta os resultados obtidos pela análise das proporções de seleção de atributos utilizando o EM durante a seleção de atributos. Assim como para os demais algoritmos de agrupamento, a melhor proporção foi a de 50%, obtendo melhor resultado em 8 das 15 bases de dados utilizadas.

Tabela 6: Média das ordenações e desvio padrão para diferentes proporções utilizando o EM como seletor de atributos.

	EM		
ID	**25%**	**50%**	**75%**
b1	1,95±0,760	**1,91±0,781**	2,13±0,744
b2	1,96±0,557	**1,83±0,486**	2,19±0,717
b3	2,11±0,652	1,98±0,674	**1,90±0,647**
b4	**1,88±0,763**	2,10±0,776	2,00±0,762
b5	**1,89±0,942**	1,94±0,616	2,16±0,548
b6	2,24±0,751	1,91±0,805	**1,83±0,752**
b7	2,06±0,712	1,99±0,797	**1,94±0,765**
b8	**1,93±0,195**	2,00±0,252	2,06±0,302
b9	2,26±0,675	**1,59±0,839**	2,14±0,697
b10	**1,99±0,485**	2,00±0,566	2,00±0,539
b11	2,38±0,677	1,93±0,884	**1,68±0,628**
b12	2,24±0,747	**1,85±0,825**	1,89±0,776
b13	2,63±0,698	**1,41±0,527**	1,95±0,677
b14	1,94±0,852	**1,73±0,688**	2,13±0,734
b15	2,02±0,373	**1,96±0,379**	2,01±0,352
Quant.	3	8	4

A Tabela 7 apresenta os resultados do teste estatístico aplicado às comparações da Tabela 6. Podemos observar que a diferença de desempenho foi estatisticamente significante para a maioria das bases de dados (12 das 15). Analisando cada comparação, é possível perceber que, assim como para os outros algoritmos de agrupamento, a proporção de 50% obteve os melhores resultados, pois a mesma ganhou de 25% e 75% em 7 e 6 bases respectivamente, perdendo apenas em 2 para 25% e 2 para 75%.

6.2.4 Conclusão da Primeira Análise

Nas subseções anteriores, foi realizada análises com diferentes algoritmos de agrupamento em diferentes proporções de seleção de atributos. Após analisar o resultado das análises prévias, fica evidente que a proporção de seleção de atributos que foi mais adequada para a Seleção Semidinâmica foi de 50%, independente do algoritmo de agrupamento utilizado.

Tabela 7: Resultados do Teste de *Friedman* (p-valor) para diferentes proporções utilizando o EM como seletor de atributos.

ID	Friedman	25% - 50%	25% - 75%	50% - 75%
b1	**0,001**	0,965	**0,004**	**0,001**
b2	**0,002**	0,151	**0,001**	**0,015**
b3	**0,001**	**0,049**	**0,001**	**0,042**
b4	**0,016**	**0,014**	0,138	0,643
b5	**0,001**	0,591	**0,001**	**0,001**
b6	**0,001**	**0,001**	**0,001**	0,585
b7	0,152	-	-	-
b8	**0,001**	**0,004**	**0,001**	**0,012**
b9	**0,001**	**0,001**	0,220	**0,001**
b10	0,759	-	-	-
b11	**0,001**	**0,001**	**0,001**	**0,001**
b12	**0,001**	**0,001**	**0,001**	0,843
b13	**0,001**	**0,001**	0,983	**0,001**
b14	**0,001**	**0,001**	**0,001**	0,648
b15	0,081	-	-	-
Placar	12	2/7	5/4	6/2

Acreditamos que esse resultado ocorreu devido às outras proporções analisadas, 25% e 75%, serem taxas extremas de seleção, onde na primeira houve uma grande redução do número de atributos, perdendo informações importantes para a classificação, e na segunda uma pequena redução, permitindo a presença de atributos poucos relevantes para serem usados pelos métodos de classificação.

6.3 Análise entre os Diferentes Algoritmos de Agrupamento

Esta seção tem por objetivo descobrir qual dos algoritmos de agrupamento melhor se adéqua a seleção Semidinâmica de atributos. Como na análise anterior a proporção de seleção que obteve melhores resultados foi a de 50%, as comparações entre os algoritmos de agrupamento realizadas nessa seção foram executadas apenas com essa mesma proporção.

Essa análise foi realizada em duas etapas. A primeira etapa consistiu na comparação entre os algoritmos de agrupamento sem a separação entre os classificadores utilizados, separando os resultados por base de dados. Enquanto a segunda etapa visa comparar os algoritmos de agrupamento separando-os por classificador, visando descobrir para qual algoritmo de agrupamento os classificadores obtiveram melhores resultados individualmente, independente da base de dados.

A Tabela 8 apresenta a média das ordenações e o desvio padrão obtidos quando comparados os três algoritmos de agrupamento utilizados na Seleção Semidinâmica, utilizando a melhor proporção encontrada, 50%.

Tabela 8: Média das ordenações e desvio padrão para todos algoritmos de agrupamento, separado por base.

ID	k-Médias	DBSCAN	EM
b1	**1,89 ± 0,732**	2,08± 0,780	2,03± 0,757
b2	1,95± 0,493	2,14± 0,492	**1,91± 0,509**
b3	**1,88± 0,652**	2,12± 0,654	1,99± 0,657
b4	2,03± 0,766	**1,96± 0,781**	2,00± 0,762
b5	**1,89± 0,794**	2,06± 0,790	2,03± 0,825
b6	2,05± 0,772	**1,91± 0,816**	2,02± 0,764
b7	**1,88± 0,751**	2,08± 0,741	2,03± 0,729
b8	1,97± 0,257	2,05± 0,330	**1,97± 0,255**
b9	**1,95± 0,762**	2,03 ± 0,805	2,00± 0,779
b10	**1,96± 0,762**	2,03± 0,805	2,00± 0,779
b11	2,07± 0,765	2,20± 0,740	**1,72 ± 0,812**
b12	**1,94± 0,775**	2,08± 0,804	1,97± 0,796
b13	**1,94± 0,798**	1,98± 0,817	2,07± 0,779
b14	**1,92± 0,770**	2,13± 0,814	1,94± 0,776
b15	1,98± 0,269	**1,97 ± 0,253**	2,04± 0,304
Quant.	9	3	3

Cada coluna representa o resultado de um algoritmo de agrupamento, onde o melhor resultado (menor valor médio de ordenação) para cada base de dados está destacado nas células cinzas. A última linha representa a quantidade de bases que tiveram melhores resultados, para cada algoritmo de agrupamento.

Baseado na Tabela 8, podemos observar que o melhor desempenho é obtido quando usamos o k-Médias como método de seleção de atributos, visto que o mesmo obteve melhores resultados em 9 das 15 bases de dados utilizadas. Já o algoritmo EM obteve 3 melhores resultados e o DBSCAN obteve 3 melhores resultados também.

A Tabela 9 apresenta os resultados da aplicação do teste de *Friedman* para os experimentos da Tabela 8.

Podemos observar que a diferença de desempenho foi estatisticamente significante para a maioria das bases de dados(11 das 15 bases de dados). Além disso, o método proposto utilizando o k-Médias foi estatisticamente melhor que a do DB em 8 base de dados e melhor que o EM em 3 base de dados, perdendo apenas em uma. Portanto, podemos concluir estatisticamente que o melhor algoritmo de agrupamento para a Seleção Dinâmica foi o k-Médias.

Tabela 9: Resultados do Teste de *Friedman* (p-valor) para todos algoritmos de agrupamento, separado por base.

ID	Friedman	KM - DB	KM - EM	DB - EM
b1	**0,008**	**0,007**	0,100	0,596
b2	**0,002**	**0,011**	0,988	**0,023**
b3	**0,004**	**0,030**	0,129	**0,043**
b4	**0,036**	**0,769**	0,167	**0,043**
b5	0,150	-	-	-
b6	0,073	-	-	-
b7	**0,002**	**0,002**	**0,007**	0,634
b8	**0,002**	**0,009**	0,977	**0,004**
b9	0,549	-	-	-
b10	**0,008**	**0,005**	0,790	**0,001**
b11	**0,002**	0,293	**0,002**	**0,001**
b12	0,128	-	-	-
b13	**0,004**	0,300	**0,002**	0,155
b14	**0,002**	**0,007**	1	**0,007**
b15	**0,001**	**0,001**	**0,001**	0,573
Placar	11	8/1	3/1	6/1

Após as análises anteriores, sem a separação entre os algoritmos de classificação, foi realizado uma análise dos diferentes métodos de classificação utilizados após a seleção de atributos. O objetivo é verificar o desempenho para cada classificador individualmente sem levar em consideração a separação dos dados entre as bases de dados.

A Tabela 10 apresenta a média das ordenações dos quatro métodos de classificação, obtido por todas as bases de dados. As colunas representam os resultados de quando cada algoritmo de agrupamento foi utilizado como método de seleção de atributos. O melhor resultado (menor valor médio das ordenações) está destacado na célula cinza.

Tabela 10: Média das ordenações e desvio padrão dos algoritmos de agrupamento para todos os métodos de classificação.

Classificador	k-Médias	Dbscan	EM
K-NN	**1,93±0,641**	2,13±0,675	1,93±0,668
AD	**1,97±0,724**	2,02±0,717	2,00±0,72
SVM	1,98±0,657	2,05±0,687	**1,95±0,660**
NB	**1,93±0,694**	2,03±0,721	2,03±0,716
Quant.	3	0	1

Analisando a referida tabela, pode-se observar que o melhor resultado ocorreu quando

o algoritmo k-Médias foi usado como método de seleção, tendo em vista que ele obteve melhores resultados em 3 dos 4 classificadores analisados, perdendo apenas no SVM, que obteve o melhor resultado quando se utilizou o EM.

A Tabela 11 apresenta os resultados da comparação estatística da Tabela 10. Pode-se observar, que 3 dos 4 classificadores tiveram resultados com significância estatística. Os resultados enfatizam que o k-Médias foi o melhor método de agrupamento para a Seleção Dinâmica, pois o mesmo ganhou do DBSCAN e do EM em 3 e 2 classificadores respectivamente, não perdendo em nenhum.

Tabela 11: Resultados do teste de *Friedman* (p-valor) para todos os algoritmos de agrupamento, separado por métodos de classificação

Classificador	**Friedman**	**KM - DB**	**KM - EM**	**DB - EM**
k-NN	**0,001**	**0,001**	**0,006**	**0,001**
AD	0,262	-	-	-
SVM	**0,001**	**0,031**	0,632	**0,001**
NB	**0,003**	**0,010**	**0,009**	0,999
Placar	3	3/0	2/0	0/2

6.3.1 Conclusão da Segunda Análise

Sendo assim, pelas análises realizadas entre os algoritmos de agrupamento utilizados, conclui-se que através dos resultados produzidos, o melhor algoritmo de agrupamento a ser utilizado como seletor de atributos da abordagem proposta foi o k-Médias, por ter obtido um melhor desempenho em relação aos demais nas análises realizadas.

Nós acreditamos que isso ocorreu devido a similaridade de processo existente entre a abordagem de seleção de atributos apresentada neste trabalho e o algoritmo do k-Médias. Tanto o algoritmo proposto como o k-Médias utilizam um processamento baseado em centro, em que as instâncias são alocadas ao grupo de centro mais próximo a elas.

Esse fato deve estar favorecendo o k-Médias, que fornece uma partição com grupos baseados em centros. Para o DBScan e o EM, onde os grupos são formados baseado em densidade e probabilidade, respectivamente, a escolha do grupo ao qual o padrão de teste pertence, por exemplo, não favorece esse tipo de partição. Uma possível solução seria uma mudança na Equação 4.5 do Capítulo 4, definindo uma nova medida de similaridade que favoreça os outros algoritmos.

6.4 Comparação entre a Versão Semidinâmica e Dinâmica

Após a descoberta de que o melhor algoritmo de agrupamento para a abordagem proposta foi o k-Médias, como dito anteriormente, uma nova versão foi produzida utilizando apenas o k-Médias durante a seleção de atributos, versão chamada de Dinâmica. Durante a próxima análise, a versão Semidinâmica será chamada de V1, enquanto a nova versão, Dinâmica, será chamada de V2.

A Tabela 12, apresenta os resultados da comparação entre as duas versões para diferentes proporções de seleção de atributos.

Tabela 12: Média das ordenações e desvio padrão da comparação das duas versões.

	25%		**50%**		**75%**	
ID	**V1**	**V2**	**V1**	**V2**	**V1**	**V2**
b1	**1,37±0,427**	1,62±0,427	**1,36±0,432**	1,64±0,432	**1,34±0,435**	1,65±0,435
b2	1,51±0,270	**1,48±0,270**	**1,49±0,295**	1,51±0,295	1,54±0,297	**1,46±0,297**
b3	1,50±0,379	**1,49±0,379**	1,52±0,365	**1,48±0,365**	**1,49±0,384**	1,50±0,384
b4	**1,40±0,457**	1,59±0,457	**1,34±0,440**	1,65±0,440	**1,34±0,434**	1,65±0,434
b5	**1,31±0,459**	1,68±0,459	1,64±0,473	**1,36±0,473**	1,61±0,483	**1,38±0,483**
b6	**1,46±0,478**	1,53±0,478	**1,32±0,445**	1,68±0,445	1,50±0,476	**1,49±0,476**
b7	**1,34±0,428**	1,65±0,428	**1,36±0,431**	1,63±0,431	**1,43±0,453**	1,56±0,453
b8	**1,49±0,117**	1,50±0,117	**1,48±0,175**	1,51±0,175	**1,43±0,199**	1,56±0,199
b9	**1,46±0,473**	1,53±0,473	**1,28±0,428**	1,71±0,428	**1,47±0,478**	1,52±0,478
b10	**1,45±0,289**	1,54±0,289	**1,48±0,267**	1,51±0,267	**1,49±0,256**	1,50±0,256
b11	**1,33±0,449**	1,66±0,449	**1,46±0,477**	1,53±0,477	1,74±0,419	**1,25±0,419**
b12	**1,41±0,480**	1,58±0,480	**1,37±0,471**	1,62±0,471	1,52±0,490	**1,47±0,490**
b13	1,93±0,245	**1,06±0,245**	1,74±0,430	**1,25±0,430**	1,79±0,400	**1,20±0,400**
b14	**1,49±0,482**	1,51±0,482	1,95±0,201	**1,04±0,201**	**1,35±0,464**	1,64±0,464
b15	**1,48±0,245**	1,51±0,245	**1,46±0,188**	1,53±0,188	1,51±0,226	**1,48±0,226**
Nº	12	3	11	4	8	7

Através dos resultados da Tabela 12, pode-se observar que a versão V1, obteve melhores resultados em todas as proporções analisadas. A proporção em que as duas versões obtiveram resultados mais semelhantes foi a de 75%, acreditamos que isso ocorreu devido a pequena redução do número de atributos, o que levava as duas versões a ter um comportamento mais semelhante do que quanto se reduzia um maior número de atributos.

Visando analisar esses resultados de maneira estatística, o teste de *Friedman*, foi aplicado a cada comparação realizada. Os resultados dessa análise estatística está presente na Tabela 13.

Tabela 13: Resultados do teste de *Friedman* (p-valor) para comparação entre as duas versões do método proposto.

Friedman	**V1-V2**		
ID	**25%**	**50%**	**75%**
b1	**0,001**	**0,001**	**0,032**
b2	0,169	0,310	**0,007**
b3	0,895	0,274	0,649
b4	**0,001**	**0,001**	**0,001**
b5	**0,001**	**0,001**	**0,002**
b6	0,117	**0,001**	0,833
b7	**0,001**	**0,001**	**0,003**
b8	0,669	**0,008**	**0,001**
b9	0,126	**0,001**	0,324
b10	**0,001**	0,131	0,922
b11	**0,001**	0,215	**0,001**
b12	**0,001**	**0,001**	0,308
b13	**0,001**	**0,001**	**0,001**
b14	0,678	**0,001**	**0,001**
b15	0,662	**0,001**	0,099
Placar	7/1	8/3	6/3

Assim, a versão Semidinâmica (V1) obteve melhores resultados que a versão Dinâmica (V2) em todas as proporções testadas. Tendo em vista que para 25%, a mesma ganhou em 7, perdendo apenas em 1 base de dados. Bem como ara 50%, onde a mesma ganhou em 8, perdendo apenas em 3 bases de dados, e para 75%, onde a mesma ganhou em 6, perdendo apenas em 3 bases de dados.

6.4.1 Conclusão da Terceira Análise

Podemos observar que a versão V1, foi a melhor versão para o método proposto. Em análises anteriores ficou evidente que o método proposto obteve melhores resultados quando se utilizou a versão Semidinâmica com o k-Médias como algoritmo de agrupamento e uma taxa de seleção de atributos de 50%.

Acreditamos que a versão Semidinâmica obteve melhores resultados do que a versão

Dinâmica por duas razões, o tamanho do conjunto de validação e o critério de avaliação utilizado. Para selecionar os melhores atributos, um algoritmo de agrupamento é aplicado a um conjunto de validação, que consiste em 10% da base de dados. A aplicação do algoritmo de agrupamento nesse pequeno conjunto, pode gerar grupos com poucas instâncias.

Como alguns grupos podem ficar com poucas instâncias, os atributos desses grupos terão um baixo valor de correlação. Como a versão Dinâmica utiliza atributos de todos os grupos, esse atributos podem estar prejudicando o desempenho da mesma. Assim, uma possível solução seria aumentar o conjunto de validação ou alterar o critério de avaliação utilizado.

Assim, agora por diante, a melhor versão do método proposto (V1) , obtida com a utilização do k-Médias, será abreviada com SDFS .

6.5 Comparação da Abordagem Proposta com Outros Métodos de Redução de Dimensionalidade

Após a realização das análises anteriores e sua conclusão, foi realizada uma nova análise, comparando o desempenho do método proposto com com outros métodos de redução de dimensionalidade.

Como a versão que obteve o melhor resultado foi a Semidinâmica, a mesma será utilizada para realizar a comparação com os demais métodos. A análise consiste em utilizar o método proposto na versão Semidinâmica com o k-Médias como método de seleção (SDFS) para comparar com os seguintes métodos de redução de características: seleção randômica de atributos (RD), PCA e LDA, respectivamente. Além desses métodos, também foi comparado sem seleção de atributos (No-FS).

O objetivo desta análise é identificar o método de redução de dimensionalidade que obteve melhor desempenho e verificar se o método proposto obteve bons resultados comparados com métodos já existentes e consagrados na literatura.

A Tabela 14 apresenta a acurácia média de todos os métodos de classificação, para cada base de dados.

Podemos observar através das médias que o SDFS obteve melhores resultados quando comparado com todos os demais métodos, pois obteve melhores resultados em 7 de 15 base de dados. Em seguida, temos os resultados do LDA com 3 melhores resultados, depois o

Tabela 14: Acurácia média e desvio padrão da comparação da melhor versão da Seleção Semidinâmica com os métodos existentes de redução de dimensionalidade

ID	SDFS	No-FS	RD	PCA	LDA
b1	2,07±1,030	2,84±1,124	3,15±1,181	**2,04±1,282**	3,72±1,364
b2	**1,66±0,948**	1,89±1,193	1,97±1,165	2,08±1,542	3,94±1,308
b3	**1,68±0,994**	2,14±1,144	2,21±1,188	2,32±1,346	2,29±1,408
b4	**2,56±1,153**	2,71±1,281	3,07±1,300	2,60±1,433	2,74±1,474
b5	2,75±1,014	**2,02±1,259**	3,45±1,027	3,00±1,674	3,38±1,456
b6	2,53±1,032	2,90±0,998	3,03±1,090	4,35±1,215	**1,79±1,265**
b7	**2,04±1,140**	2,72±1,241	2,95±1,166	2,40±1,376	3,17±1,529
b8	**1,08±0,312**	1,31±0,813	1,25±0,724	3,15±1,428	4,04±1,15
b9	2,92±1,111	2,83±1,126	3,49±1,254	3,17±1,490	**1,85±1,269**
b10	**1,35±0,818**	1,43±0,912	1,66±1,080	2,27±1,619	3,86±1,325
b11	2,74±1,231	**2,64±1,240**	3,32±1,243	3,07±1,567	2,71±1,521
b12	2,56±1,128	**2,42±1,042**	3,13±1,154	3,29±1,459	2,65±1,636
b13	3,45±0,627	3,02±0,694	4,90±0,379	**1,46±0,644**	1,76±0,776
b14	2,71±1,007	3,01±0,999	3,56±0,989	4,22±1,088	**1,21±0,613**
b15	**1,11±0,533**	1,15±0,639	1,34±1,036	2,78±1,980	1,16±0,684
Quant.	7	3	0	2	3

No-FS com 3 e finalmente o PCA com apenas 2 entre os melhores resultados. A seleção randômica (RD) não obteve o melhor resultado em nenhuma base de dados.

Na Tabela 15 estão presentes os resultados do teste de *Friedman* para a Tabela 14. Quando observamos a Tabela 15, podemos ver que o método proposto (SDFS) obteve melhores resultados quando comparado com todos os outros métodos. Em todas as 15 bases utilizadas, o desempenho dos diferentes métodos de seleção de características provou ser estatisticamente significativo. Quando comparamos o método proposto, em cada uma das bases com os demais métodos, o mesmo supera o método randômico em 13 bases, superando também o No-FS, PCA e LDA em 6, 9 e 7 bases, respectivamente.

Em suma, os resultados apresentados neste trabalho são muito promissores, uma vez que o método proposto alcançou um melhor desempenho que outros métodos de redução de características bem estabelecidas.

Além disso, o método proposto conseguiu ser melhor do que utilizando os conjuntos de dados originais, mostrando que a redução de atributos ruidosos e/ou redundantes pode ter um efeito positivo no desempenho de uma tarefa de classificação.

Tabela 15: Resultados do teste de *Friedman* (p-valor) da comparação da melhor versão da Seleção Semidinâmica com os métodos existentes de redução de dimensionalidade

ID	**Friedman**	**SDFS-NoFS**	**SDFS-RD**	**SDFS-PCA**	**SDFS-LDA**
b1	**0,002**	**0,001**	**0,001**	0,998	**0,001**
b2	**0,001**	**0,001**	**0,002**	0,304	**0,001**
b3	**0,001**	**0,001**	**0,001**	**0,001**	**0,001**
b4	**0,001**	0,660	**0,001**	0,998	0,561
b5	**0,007**	**0,001**	**0,002**	0,270	**0,001**
b6	**0,001**	**0,005**	**0,001**	**0,001**	**0,001**
b7	**0,001**	**0,001**	**0,001**	**0,021**	**0,001**
b8	**0,001**	0,079	0,207	**0,001**	**0,001**
b9	**0,001**	0,930	**0,001**	0,511	**0,001**
b10	**0,001**	0,817	**0,021**	**0,001**	**0,001**
b11	**0,001**	0,809	**0,001**	**0,018**	0,992
b12	**0,001**	0,682	**0,001**	**0,001**	0,906
b13	**0,001**	**0,012**	**0,001**	**0,001**	**0,001**
b14	**0,001**	**0,003**	**0,001**	**0,001**	**0,001**
b15	**0,001**	0,995	0,183	**0,001**	0,997
Numb	15	6/2	13/0	9/1	7/4

7 Conclusão

Este capítulo apresenta as considerações finais, apresentando as conclusões obtidas com os experimentos realizados, como também as principais contribuições e o que ainda pode ser explorado futuramente deste trabalho.

7.1 Considerações Finais

Este trabalho apresentou um novo método de seleção de atributos não supervisionada para tarefas de classificação. O principal objetivo do método proposto foi selecionar os melhores atributos para um grupo de instâncias, em vez de todo o conjunto de dados, utilizando uma maneira dinâmica de selecionar atributos. Para realizar a seleção dos atributos mais relevantes, como base do método proposto foram utilizados algoritmos de agrupamento.

A afim de avaliar a viabilidade do método proposto, uma análise empírica foi realizada. Para isto, foram aplicados três diferentes algoritmos de agrupamento com diferentes proporções de seleção de atributos. Além disso, também foi desenvolvida duas versões para o método proposto, uma Semidinâmica e outra Dinâmica. Assim, o subconjunto de características obtido foi aplicado à quatro métodos de classificação diferentes, em 15 conjuntos de dados (obtidos a partir do repositório UCI).

Com a realização da análise empírica, pode-se verificar que dos algoritmos de agrupamento utilizados, o que melhor se adaptou a essa nova técnica de seleção de atributos foi o k-Médias, que obteve um melhor desempenho que os demais algoritmos de agrupamento nas diversas análises realizadas.

Para fins de comparação, o desempenho da melhor versão do método proposto foi comparado com três métodos existentes de redução de dimensionalidade, a redução de atributos aleatória, e os métodos de extração de características PCA e LDA, assim como a utilização do conjunto de dados originais (sem redução de características). Para realizar

essa comparação, o método proposto utilizou o k-Médias em sua versão Semidinâmica de seleção de atributos, tendo em vista que esse algoritmo de agrupamento obteve os melhores resultados em análises anteriores. Através desta análise empírica, pudemos observar que o método proposto superou estatisticamente o método randômico em 13 bases de dados, bem como o No-FS, PCA e LDA em 6, 9 e 7 conjuntos de dados, respectivamente.

Os resultados apresentados neste trabalho são bastantes promissores, um vez que o método proposto conseguir ter um melhor desempenho do que métodos de redução de dimensionalidade bem estabelecidos, sejam eles de redução ou extração de características.

Além disso, a utilização do método proposto conseguiu melhor desempenho do que utilizando os conjuntos de dados originais, mostrando que a redução pode ter um efeito positivo no desempenho de uma tarefa de classificação, reduzindo significativamente o tamanho do conjunto de dados, sem afetar o desempenho, obtendo na maioria dos casos uma melhoria na sua acurácia de classificação.

7.1.1 Principais Contribuições

Este trabalho traz uma contribuição às pesquisas sobre seleção de atributos, apresentando uma nova abordagem com a utilização de algoritmos de agrupamento na seleção do melhor subconjunto de atributos de uma base de dados. Como principais contribuições deste trabalho, podemos citar:

- Criação de uma nova abordagem para seleção de atributos, que visa escolher os melhores atributos para uma instância, ou grupo de instância. Diferentemente dos métodos tradicionais, quer selecionam os melhores atributos para uma base de dados de maneira geral, sem divisão por instâncias.
- Criação de duas versões da abordagem proposta. A primeira versão, Semidinâmica, seleciona os melhores atributos de acordo com o grupo a qual a instância mais se assemelha, enquanto a segunda versão, Dinâmica, seleciona de maneira probabilística de acordo com a distância para todos os grupos, selecionando uma maior quantidade de atributos dos grupos mais próximos.
- Análise dos diferentes algoritmos de agrupamento, visando determinar qual mais se adequa ao método proposto.
- Análise de diferentes proporções de seleção de atributos, buscando encontrar a taxa de seleção que melhor se adapta a abordagem proposta.

- Comparação nova abordagens com outras existentes, mostrado que a mesma conseguiu reduzir a dimensionalidade dos dados e obter um aumento na acurácia de classificação, mostrando que uma seleção adequada dos atributos mais relevantes de uma base de dados, eleva a precisão de um classificador.

7.1.2 Trabalhos Futuros

A seguir são apresentados alguns pontos que podem ser desenvolvidos com o método de seleção de atributos apresentado neste trabalho:

- Comparar a abordagem proposta com outros métodos de seleção de atributos.
- Investigar o uso de novas medidas de similaridades.
- Utilizar outros critérios de avaliação para definir os atributos mais relevantes.
- Utilizar outros algoritmos de agrupamento.
- Investigar o baixo desempenho do DBSCAN quando utilizado na abordagem proposta.
- Utilizar novas bases de dados, para analisar o comportamento do método proposto com outros conjuntos de dados.

Referências

BELLMAN, R. Dynamic programming and lagrange multipliers. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 42, n. 10, p. 767–769, 1956.

BERRY, M. J. A.; LINOFF, S. L. *Mastering Data Mining.* USA: Wiley and Sons, 2000.

BHONDAVE, R. et al. Improvement of expectation maximization clustering using select attribute. *International Journal of Computer Science and Mobile Computing*, Issue, v. 3, n. 4, p. 503–508, 2014.

BOUTSIDIS, C.; MAGDON-ISMAIL, M. Deterministic feature selection for k-means clustering. *IEEE Transactions on Information Theory*, IEEE, v. 59, n. 9, p. 6099–6110, 2013.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Electrical and Microelectronic Engineering,*, v. 40, n. 1, p. 16–28, 2014.

CHU, C. et al. Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, Elsevier, v. 60, n. 1, p. 59–70, 2012.

CLIFFORD, H.; STEPHENSON, W. *An introduction to numerical taxonomy.* London: Academic Press, 1975.

CONOVER, W. J. *Practical Nonparametric Statistics*. [S.l.]: Wiley New York, 1980.

COVõES, T. F. Dissertação de Mestrado, *Seleção de atributos via agrupamento*. São Paulo, SP, Brasil: [s.n.], Janeiro 2010.

COVõES, T. F.; HRUSCHKA, E. R. An experimental study on unsupervised clustering-based feature selection methods. In: IEEE. *2009 Ninth International Conference on Intelligent Systems Design and Applications*. [S.l.], 2009. p. 993–1000.

DAS, S.; PANIGRAHI, B. K. Multi-objective evolutionary algorithms. *Encyclopedia of artificial intelligence*, v. 3, p. 1145–1151, 2009.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *KDD*. [S.l.]: AAAI Press, 1996. p. 226–231.

FIGUEIREDO, F. D. B. et al. Desvendando os mistérios do coeficiente de correlação de pearson: o retorno. *Leviathan: Cadernos de Pesquisa Política*, n. 8, p. 66–95, 2014.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.

FORMAN, G. An extensive empirical study of feature selection metrics for text classication. *The Journal of machine learning research*, n. 3, p. 1289–1305, 2003.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, v. 32, n. 200, p. 675–701, 1937.

GAN, G.; MA, C.; WU, J. *Data clustering: theory, algorithms*. Philadelphia, Pennsylvania: ASA-SIAM Series on Statistics and Applied Probability, 2007.

GUNN, S. et al. Support vector machines for classification and regression. *ISIS technical report*, v. 14, 1998.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, n. Mar, p. 1157–1182, 2003.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. 3. rd. USA: Elsevier, 2011.

HANSEN, M.; DUBAYAH, R.; DEFRIES, R. Classification trees: An alternative to traditional land cover classifiers. *International Journal of Remote Sensing*, v. 17, n. 5, p. 1075–1081, 1996.

HARRISON, T. *Intranet data warehouse*. Brasil: Editora Berkeley, 1998.

LAW, M. H. C.; FIGUEIREDO, M. A. T.; JAIN, A. K. Simultaneous feature selection and clustering using mixture models. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 26, n. 9, p. 1154–1166, 2004.

LIU, H.; MOTODA, H. *Feature extraction, construction and selection: A data mining perspective*. [S.l.]: Springer Science & Business Media, 1998.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.]: Berkeley, CA: University of California Press, 1997. p. 281–297.

MARDIA, K. V.; KENT, J. T. B. J. M. *Multivariate Analysis*. 2th. ed. USA: Academic Press, 1979.

MITCHELL, T. M. Machine learning. *McGraw-Hill Science*, 1997.

MLADENOVIC, V. M. et al. The direct execution of expecttion-maximization algorithm using symbolic processing. In: *Telecommunication in Modern Satellite Cable and Broadcasting Services (TELSIKS)*. [S.l.]: IEEE, 2011. p. 265–268.

OLIVEIRA, J. A.; DUTRA, L. V.; RENNó, C. D. Aplicação de métodos de extração e seleção de atributos para classificação de regiões. *XII Simpósio Brasileiro de Sensoriamento Remoto*, INPE, n. 1, p. 4201–4208, 2005.

PAPPA, G. L. *Seleção de atributos utilizando Algoritmos Genéticos multiobjetivos*. Tese (Doutorado) — Pontifícia Universidade Católica do Paraná, 2002.

PARK, C. H.; KIM, S. B. Sequential random k-nearest neighbor feature selection for high-dimensional data. *Expert Systems with Applications*, Elsevier, v. 42, n. 5, p. 2336–2342, 2015.

PEARSON, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, v. 2, n. 6, p. 559–572, 1901.

PEREIRA, R. B. *Seleção lazy de atributos para a tarefa de classificação.* Tese (Doutorado) — UFF-Universidade Federal Fluminense, Brazil, 2009.

QUINLAN, J. R. Induction of decision trees. In: *Machine Learning 1.* [S.l.]: Kluwer Academic, Boston, 1986. p. 81–106.

SANTHANAM, T.; PADMAVATHI, M. Application of k-means and genetic algorithms for dimension reduction by integrating svm for diabetes diagnosis. *Procedia Computer Science*, Elsevier, v. 47, p. 76–83, 2015.

THOM, A.; KRAMER, O. Acceleration of DBSCAN-based clustering with reduced neighborhood evaluations. *KI 2010: Advances in Artificial Intelligence*, v. 6359, p. 195–202, 2010.

WITTEN, I.; FRANK, E. *Data Mining: pratical machine learning toolsand techniques.* 2th. ed. USA: Elsevier, 2005.

YU, L.; LIU, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: *ICML.* [S.l.]: Proceedings of the Twentieth International Conference on Machine Learning, 2003. v. 3, p. 856–863.

ZHOU, A. *et al.* Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation*, Elsevier, v. 1, n. 1, p. 32–49, 2011.

ZITZLER, E.; THIELE, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE transactions on Evolutionary Computation*, IEEE, v. 3, n. 4, p. 257–271, 1999.

APÊNDICE A – Resultados das Médias das Acurácias

As tabelas a seguir apresentam a acurácia média e desvio padrão das análises realizadas no Capítulo 6.

Tabela 16: Acurácia média das proporções do k-Médias

	k-Médias		
ID	**25%**	**50%**	**75%**
b1	90,2	90,6	90,5
b2	87,5	87,8	87,1
b3	69,0	68,3	67,7
b4	70,4	69,2	71,3
b5	73,3	73,2	73,1
b6	52,3	56,7	53,2
b7	86,0	87,1	85,7
b8	99,7	99,5	99,8
b9	52,2	60,2	57,9
b10	92,2	91,1	91,0
b11	86,6	87,9	84,2
b12	89,5	90,5	83,9
b13	75,4	77,5	75,1
b14	68,0	73,2	74,2
b15	99,7	99,9	99,6

Tabela 19: Acurácia média dos algoritmos de agrupamento para todos os métodos de classificação

Classificador	***k*-Médias**	**Dbscan**	**EM**
k-NN	88,0	85,6	87,8
AD	76,6	75,5	75,5
SVM	79,6	78,8	87,3
NB	75,8	73,8	74,9

Tabela 17: Acurácia média das proporções do DBSCAN

	DBSCAN		
ID	**25%**	**50%**	**75%**
b1	90,1	89,6	89,4
b2	84,9	83,3	80,6
b3	65,0	57,8	60,5
b4	69,4	69,9	69,2
b5	69,7	72,3	71,6
b6	52,4	58,3	53,5
b7	85,1	85,1	86
b8	99,8	99,0	99,8
b9	55,5	59,4	56,2
b10	91,4	89,0	90,8
b11	85,3	87,5	90,2
b12	91,1	88,9	85,3
b13	81,9	85,5	79,9
b14	68,3	71,1	72,5
b15	99,8	99,9	99,8

Tabela 18: Acurácia média das proporções do EM

	EM		
ID	**25%**	**50%**	**75%**
b1	89,9	90,0	88,7
b2	84,5	88,3	80,8
b3	57,0	63,5	67,1
b4	72,0	70,1	70,9
b5	73,0	72,3	72,4
b6	52,8	56,9	57,3
b7	84,7	85,8	85,8
b8	99,7	99,5	99,2
b9	56,7	58,9	57,8
b10	92,9	91,6	91,2
b11	84,3	88,3	88,4
b12	86,8	91,0	91,5
b13	83,5	85,3	85,1
b14	70,2	73,2	66,5
b15	99,7	99,8	99,7

Tabela 20: Acurácia média das proporções da Versão Semidinâmica (V1)

	V1		
ID	**25%**	**50%**	**75%**
b1	90,1	90,7	90,5
b2	87,5	87,8	87,2
b3	69,0	68,3	69,1
b4	70,4	70,2	71,3
b5	73,3	73,2	73,1
b6	52,6	56,7	53,2
b7	86,0	87,1	85,7
b8	99,7	99,5	98,8
b9	56,6	60,2	58,9
b10	92,2	91,1	91,0
b11	86,6	88,0	84,3
b12	89,5	90,6	88,7
b13	82,6	85,7	83,3
b14	72,6	73,2	74,2
b15	99,7	99,8	99,6

Tabela 21: Acurácia média das proporções da Versão Dinâmica (V2)

	V2		
ID	**25%**	**50%**	**75%**
b1	87,7	85,5	87,4
b2	89,1	86,1	88,5
b3	69,0	70,2	67,0
b4	66,2	64,0	65,1
b5	70,2	74,5	72,9
b6	51,3	51,4	53,4
b7	79,9	82,7	83,3
b8	99,7	99,0	98,5
b9	50,6	52,8	55,8
b10	89,2	89,7	89,4
b11	82,0	86,3	88,3
b12	83,7	84,1	83,7
b13	85,6	88,0	90,1
b14	68,6	70,1	70,0
b15	99,7	99,7	99,7

Tabela 22: Acurácia média da comparação da melhor versão com alguns métodos existentes de redução de dimensionalidade

ID	**SDFS**	**NoFS**	**RD**	**PCA**	**LDA**
b1	90,6	87,9	86,3	91,3	80,7
b2	74,9	68,6	69,0	73,2	39,5
b3	68,3	53,0	51,7	50,3	51,8
b4	69,2	67,9	65,1	68,8	68,1
b5	73,2	75,5	71,1	59,7	70,8
b6	56,7	54,9	54,3	53,7	68,5
b7	87,1	82,9	81,3	86,4	79,5
b8	99,5	99,0	98,7	70,0	65,6
b9	60,2	60,61	56,7	61,2	74,7
b10	91,1	89,6	88,1	71,3	55,7
b11	87,9	88,1	86,3	86,0	86,9
b12	90,5	91,2	87,4	84,8	90,3
b13	77,4	78,8	69,1	94,5	93,5
b14	73,2	71,3	68,2	51,7	85,7
b15	99,9	99,8	52,8	97,2	99,8