UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DEPARTAMENTO ACADÊMICO DE INFORMÁTICA

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JOÃO PEDRO EVARISTO DE JULIO

REDUÇÃO DE DIMENSIONALIDADE: APLICAÇÃO DE ALGORITMOS DE SELEÇÃO E EXTRAÇÃO DE ATRIBUTOS

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA

2019

JOÃO PEDRO EVARISTO DE JULIO

REDUÇÃO DE DIMENSIONALIDADE: APLICAÇÃO DE ALGORITMOS DE SELEÇÃO E EXTRAÇÃO DE ATRIBUTOS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação, do Departamento Acadêmico de Informática, da Universidade Tecnológica Federal do Paraná.

Orientador: Profª. Dra Helyane Bronoski Borges

PONTA GROSSA

2019

TERMO DE APROVAÇÃO

REDUÇÃO DE DIMENSIONALIDADE: APLICAÇÃO DE ALGORITMOS DE SELEÇÃO E EXTRAÇÃO DE ATRIBUTOS

por

JOÃO PEDRO EVARISTO DE JULIO

Este Trabalho de Conclusão de Curso (TCC) foi apresentado em 19 de Novembro de 2019 como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Prof.ª Dra Helyane Bronoski Borges
Orientadora

_____
Prof.ª Dra Simone Nasser Matos
Membro titular

_____
Prof. MSc Geraldo Ranthum
Membro titular

_____
Prof. MSc Geraldo Ranthum
Responsável pelo Trabalho de Conclusão de Curso

_____
Profª. Dra Mauren Louise Sguario
Coordenador do curso

AGRADECIMENTOS

Primeiramente aos meus pais por me fornecerem todo apoio e motivação além de todos os sacrifícios que fizeram em prol dos meus objetivos.

Á minha namorada Jéssica que esteve ao meu lado durante todos estes anos me motivando.

A minha orientadora Prof.ª Dr.ª Helyane Bronoski Borges, pela atenção, persuasão e ensinamentos.

Aos meus companheiros da Casa Verde e ao meu amigo Renan pela descontração, auxílio e pela palavra amiga.

Obrigado.

RESUMO

DE JULIO, João Pedro E. **Redução de Dimensionalidade:** Aplicação de algoritmos de seleção e extração de atributos. 2019. 77 f. Trabalho de Conclusão de Curso. Bacharelado em Ciência da Computação - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2019.

O diagnóstico de doenças genéticas como câncer tem avançado com a evolução de técnicas de obtenção de dados genéticos, e a quantidade de genes mapeados tem aumentado significativamente e consequentemente a complexidade na análise destes dados devido ao pouco número de amostras. Por meio de técnicas como a Seleção (com as abordagens Filtro, *Wrapper* e *Embedded*), e a Extração de atributos é possível realizar a redução da dimensionalidade, que além de remover atributos irrelevantes ou redundantes, torna mais fácil a compreensão dos resultados. A Seleção de atributos tem como objetivo encontrar atributos relevantes para aumentar a capacidade preditiva dos classificadores enquanto a Extração de atributos realiza operações de transformação sem a perda das características dos dados. Este trabalho apresenta uma aplicação de técnicas de Extração de atributos sobre subconjuntos selecionados por meio da Seleção de atributos, realizando assim uma combinação das técnicas. A combinação proposta utiliza a busca sequencial para selecionar os atributos com dois algoritmos da abordagem Filtro e sete formas de redução da abordagem *Wrapper*. Em cada subconjunto foi-se aplicado a Análise de Componentes Principais (PCA) com os 90, 95 e 99% dos atributos. Para os experimentos, foram utilizadas 5 bases de dados genéticas com milhares de atributos por amostra. Ao realizar a análise da taxa de classificação com sete diferentes classificadores, pode-se notar um aumento significativo na taxa de classificação dos dados após a aplicação da combinação de técnicas, obtendo-se um aumento de até 12% no pior caso.

Palavras-chave: Seleção de Atributos. Extração de Atributos. Análise de Componentes Principais. Redução de Dimensionalidade.

ABSTRACT

DE JULIO, João Pedro E. **Dimensionality reduction:** Application of attribute selection and attribute extraction algorithms. 2019. 77 p. Work of Conclusion Course. Graduation in Computer Science - Federal Technology University - Paraná. Ponta Grossa, 2019.

The diagnosis of genetic diseases such as cancer has advanced with the evolution of techniques for obtaining genetic data, and the number of mapped genes has increased significantly and consequently the complexity in the analysis of these data due to the small number of samples. Techniques such as Selection (with the Filter, Wrapper, and Embedded approaches) and Attribute Extraction make it possible to reduce dimensionality, which in addition to removing irrelevant or redundant attributes, makes it easier to understand the results. Attribute Selection aims to find relevant attributes to increase the predictive capacity of classifiers while Attribute Extraction performs transformation operations without losing data's properties. Thus, this paper presents an application of Attribute Extraction techniques on selected subsets through Attribute Selection. The proposed combination uses sequential search to select attributes with two algorithms of the Filter approach and seven ways to reduce the Wrapper approach. In each subset, PCA was applied with 90, 95 and 99% of the attributes. For the experiments, five genetic databases with thousands of attributes per sample were used. When analyzing the classification rate with seven different classifiers, can be noted a significant increase in the data classification rate after applying the combination of techniques, resulting in an increase of up to 12% in the worst case.

Keywords: Attribute Selection. Attribute Extraction. Principal Component Analysis. Dimensionality Reduction.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

ALDE	*Angle Linear Discriminant Embendding*
ALL	Leucemia Linfoblática Aguda
AML	Leucemia Mielóide Aguda
BLOGREG	*Bayesian Logistic Regression*
CFG	*Compount feature generation*
CFS	*Correlation Feature Selection*
dICA	*Discriminant Independent Component Analysis*
DLBCL	*Diffuse Large B-Cell Lymphoma*
DNA	Ácido Desoxirribonucleico
ELM	*Extreme Learning Machine*
Fi	*Fisher Score*
FL	Linfoma Folicular
ICA	Análise de Componentes Independentes
KDD	*Knowledge-Discovery in Databases*
LDA	*Linear Discriminant Analysis*
LLE	*Locally Linear Embendding*
MIM	*Mutual Information Maximization*
MIMAGA	*Mutual Information Maximization with Adaptive Genetic Algorithm*
MMC	*Maximum Margin Criterion*
NPE	*Neighborhood Preserving Embendding*
OLDA	*Orthogonal LDA*
ONPPs	*Orthogonal Neighborhood Preserving Projections*
PCA	Análise de Componentes Principais
PLS	*Partial Least Square*
PLSRFE	*PLS-based feature selection*
RELM	*Regularized Extreme Learning Machine*
SFS	*Sequential Forward Selection*
SLDA	*Sparse Linear Discriminant Analysis*
TR	*Trace Ratio*
Weka	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Microarranjo de Ácido Desoxirribonucleico (DNA) é uma forma de se obter dados de expressões gênicas que vem sendo popularizada devido a possibilidade de analisar grandes quantidades de dados em tempo viável. Esta análise, realizada através do processo de Extração de Conhecimento de Bases de Dados (KDD - *Knowledge-Discovery in Databases*), pode auxiliar na identificação de padrões referentes a doenças apesar de apresentar certos desafios para a mineração de dados em tarefas como agrupamentos de atributos e amostras, classificação e regressão (JIANG; XU, 2015).

Um dos desafios enfrentados na tarefa de classificação dos dados de microarranjos de DNA é o tamanho da base de dados, pois apresenta um número elevado de genes (atributos), da ordem de milhares, e relativamente poucas amostras. Devido a estas características, torna-se problemática a aplicação de classificadores tradicionais devido ao alto custo computacional e o comprometimento da precisão do classificador(BORGES; NIEVOLA, 2012).

Para lidar com este problema é utilizada a redução de dimensionalidade que tem como objetivo reduzir o volume de dados - em relação ao número de atributos, a serem tratados. Eliminando dados irrelevantes ou redundantes e melhorando a capacidade de generalização de métodos de aprendizagem (BORGES, 2006). Segundo Kastrin e Peterlin (2010) existem duas técnicas para redução do número de atributos: a extração de atributos e a seleção de atributos.

A extração de atributos realiza uma transformação ou combinações nos atributos originais, construindo novas representações para estes dados sem perder as características originais dos mesmos. Esta transformação pode ser realizada com métodos lineares como: Análise de Componentes Principais (PCA), análise de discriminantes lineares (LDA) e Análise de Componentes Independentes (ICA). E não lineares como: Isomap, Análise de Componentes Curvilíneos e Incorporação localmente linear (LLE) (GUYON et al, 2006).

O método PCA pode realizar a redução do número de atributos mantendo o máximo da variação entre os dados, transformando uma grande quantidade de atributos em um pequeno subconjunto a partir da perspectiva da significância destes atributos. Esta redução visa facilitar a tarefa de classificação, visto que, os dados resultantes possuem alta correlação com a categoria da amostra (DING et al, 2012).

A seleção de atributos é uma técnica que seleciona o melhor subconjunto do grupo original de dados, conforme a estratégia de busca e a medida de avaliação utilizada. É um processo presente em grande parte dos algoritmos de aprendizagem de máquina utilizados na mineração de dados pois determinam os atributos que mais impactam na tomada de decisão (LLERENA, 2013). Para a seleção de atributos é possível realizar o teste dos subconjuntos de formas distintas como por exemplo por meio das abordagens Filtro e *Wrapper*.

Diversos estudos são encontrados a respeito da aplicação de métodos de seleção e de extração de atributos em microarranjos de DNA (ALMEIDA, 2018), porém pouco se encontra a respeito da combinação das abordagens de seleção de atributos com a técnica de análise de componentes principais.

Pode-se observar em Borges (2006) a grande quantidade de subconjuntos gerados à partir da utilização de múltiplas abordagens de seleção de atributos, com duas diferentes formas de busca (sequencial e aleatória), combinado com a extração de atributos. Com isso, este trabalho apresenta uma análise da aplicação de combinação de técnicas de seleção de atributos com busca sequencial, como apresentado em Borges (2006), juntamente com a PCA. Esta combinação busca alcançar uma melhor taxa de classificação em bases de dados ligadas a doenças possíveis de serem identificadas a partir de microarranjos de DNA.

1.1 DESCRIÇÃO DO PROBLEMA

A análise de bases de dados de microarranjos vem se mostrando importante na identificação de diversos tipos de câncer devido a capacidade de representar expressões gênicas completas (PEREZ-DIEZ, MORGUN, SHULZHENKO, 2007). Porém, estas bases possuem uma limitação devido à alta quantidade de atributos presentes pois, além de tornar a análise lenta e computacionalmente custosa, possuem um grande número de atributos redundantes que podem interferir no processo de classificação desses dados. Os dados redundantes acabam prejudicando a tarefa de mineração, pois confundem os algoritmos com dados repetitivos (BÓLON-CANEDO, 2017).

Para a redução da dimensionalidade pode-se destacar duas técnicas: a seleção e a extração de atributos. Segundo Borges e Nievola (2012), a seleção de atributos é uma técnica que visa encontrar o menor subconjunto de atributos que melhor representem as características da base de dados original. A extração de atributos visa transformar um conjunto de atributos em um conjunto com um número menor destes, porém sem perder as principais características dos dados (HAND, MANNILA e SMYTH, 2001).

Com o objetivo de classificar as amostras, a seleção e a extração de atributos buscam maximizar a taxa de acerto do classificador e minimizar o número de atributos utilizados na classificação (AHUJA, 2017).

Algoritmos de seleção e extração de atributos são utilizados, tanto individualmente quanto combinados, em dados de microarranjos de DNA como alguns trabalhos publicados em (DASH, 2017; ARUNKUMAR, KAMAKRISHNAN, 2018; KUMAR, RATH, RATH, 2016; GUO et al, 2017; LATKOWSKI, OSOWSKI, 2017; HE et al, 2015; EBRAHIMPOUR et al, 2017; AZIZ, VERNA, SRIVASTAVA, 2016; LU et al, 2017; MOLLAEE, MOATTAR, 2016; YOU et al, 2014; NANNI, BRAHNAM, LUMINI, 2012). Com estes trabalhos, é possível observar que técnicas de seleção, extração e ambas combinadas podem trazer melhores resultados em relação à métricas da tarefa de classificação quando comparados com a base original de dados.

1.2 OBJETIVOS

Esta seção apresenta o objetivo geral e os objetivos específicos deste trabalho.

1.2.1 Objetivo Geral

Analisar os resultados da aplicação de métodos de seleção de atributos juntamente com extração de atributos por meio da Análise de Componentes Principais em dados de microarranjos de DNA.

1.2.2 Objetivos específicos

- Realizar o mapeamento sistemático da literatura sobre técnicas de redução de dimensionalidade em bases de dados de expressão gênica;
- Aplicar os algoritmos de seleção de atributos e Análise de Componentes Principais;
- Realizar experimentos em bases de expressão gênica;
- Comparar os resultados estatisticamente.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está estruturado em seis capítulos. O Capítulo 2 apresenta os conceitos sobre bases de dados, atributos, redução de dimensionalidade e microarranjos de DNA.

O Capítulo 3 apresenta a revisão sistemática executada na literatura, possibilitando a identificação de trabalhos relacionados que são referência para este trabalho.

O Capítulo 4 apresenta a descrição das bases de dados utilizadas, a metodologia aplicada para os experimentos e para a avaliação dos resultados.

O Capítulo 5 apresenta a realização dos experimentos utilizando a metodologia definida e a análise dos resultados obtidos.

O Capítulo 6 apresenta as conclusões e os trabalhos futuros sugeridos.

2 REDUÇÃO DE DIMENSIONALIDADE E CLASSIFICAÇÃO EM BASES DE DADOS

Este Capítulo apresenta os conceitos fundamentais da redução de dimensionalidade e sobre a técnica de microarranjos de DNA. Na Seção 2.1 são descritos os conceitos introdutórios sobre bases de dados, atributos e microarranjos. Na Seção 2.2 são apresentadas as técnicas e abordagens de seleção de atributos e na Seção 2.3 as estratégias de busca que podem ser utilizadas na seleção. Na Seção 2.4 é dada uma visão geral a respeito da extração de atributos. Por fim, a Seção 2.5 apresenta as considerações finais do capítulo.

2.1 CONCEITOS FUNDAMENTAIS

Uma base de dados pode ser definida como uma coleção de objetos que podem ser eventos, observações ou registros. Estes objetos de dados são caracterizados por valores em um conjunto de características pré-determinadas chamadas de atributos (WITTEN; FRANK; HALL, 2011).

As bases de dados possuem três características gerais: Dispersão, Resolução e Dimensão. A dispersão ocorre quando há muitos atributos com valor zero, que são geralmente ligadas a dados assimétricos. A resolução é referente ao padrão dos dados e a dimensão está relacionada a quantidade de atributos presentes nos objetos. A alta dimensão pode ocasionar a “Maldição da dimensionalidade” (AHUJA, 2017).

Os valores dos atributos em um determinado registro são uma representação simbólica ou numérica das características reais do objeto. Diferentes tipos de atributos são utilizados para verificar a consistência entre as propriedades dos valores medidos e as propriedades do atributo (TAN; STEINBACH; KUMAR, 2006). Segundo Han, Kamber e Pei (2012) este tipo é determinado pelo conjunto de valores possíveis para aquele atributo, sendo que os tipos mais comuns de atributos são quantitativos e qualitativos.

Os atributos quantitativos podem ser representados por valores reais ou valores inteiros possuindo propriedades de ordem e distância entre os valores. Por sua vez, os atributos qualitativos possuem como valores símbolos ou nomes (rótulos) que mesmo sendo representados com números, não possuem a maioria das

propriedades dos atributos numéricos, considerando que estes valores podem ser iguais ou diferentes entre si (KANTARDZIC, 2011).

Os tipos de atributos podem ainda serem divididos em atributos nominais e ordinais (no caso dos atributos qualitativos) e atributos intervalares e proporcionais (no caso dos atributos quantitativos) (TAN; STEINBACH; KUMAR, 2006).

Atributos nominais nos fornecem apenas informações para distinção de um objeto a outro não possuindo uma ordem significativa. Cada valor representa uma categoria, código ou estado sendo possível apenas realizar operações de igualdade e desigualdade (KANTARDZIC, 2011).

Atributos Ordinais possuem valores que fornecem informações suficientes para ordenar os objetos, sem existir relação de distância entre os valores, não sendo possível dizer quão diferente um valor é de outro (WITTEN; FRANK; HALL, 2011).

Os atributos intervalares são representados em unidades fixas e de igual tamanho permitindo valores positivos e negativos. As diferenças entre os valores são significativas, sendo possível a realização de operação de soma e subtração. Além disso, o valor zero não é considerado um “zero-real”, já que o zero não é definido como ausência de característica (HAN; KAMBER; PEI, 2012).

Já os atributos proporcionais, como o nome indica, se “preocupam” tanto com a proporção quanto a diferença entre os valores, possuindo a representação do “zero-real”. É possível realizar a ordenação dos valores assim como o cálculo da diferença entre eles (LIU, MOTODA, 1998).

Segundo Kantardzic (2011), outra maneira de diferenciar os atributos é pela quantidade de valores possíveis, em discretos e contínuos. Os atributos discretos possuem um conjunto finito de valores possíveis e são muitas vezes representados por números inteiros. Os atributos discretos são especificados pelos atributos nominais e ordinais. Um caso especial dos atributos discretos são os atributos binários que podem assumir apenas dois valores (geralmente representados por Zero e Um) (KANTARDZIC, 2011). Atributos contínuos podem conter valores do universo dos reais, implicando em infinitas possibilidades de valores. Muitas vezes são representados com números em ponto flutuante, possuindo precisão limitada. São especificados pelos atributos intervalares e proporcionais (TAN; STEINBACH; KUMAR, 2006).

Ainda há a definição de vetor de atributos, que se refere a um conjunto de atributos como os microarranjos de DNA, que são vetores bidimensionais, onde seus

atributos representam genes, os elementos da matriz representam o nível da expressão em um gene particular e sua classe representa a classificação deste gene (Tabela 1) (AROWOLO et al, 2017).

Tabela 1 - Exemplo de base de dados de Microarranjos de DNA

Instância	Atributo1	Atributo2	Atributo3	....	Atributo m-1	Atributo m	Classe
1	$X_{1\,1}$	$X_{1\,2}$	$X_{1\,3}$	⋯	$X_{1\,m-1}$	$X_{1\,m}$	A
2	$X_{2\,1}$	$X_{2\,2}$	$X_{2\,3}$	⋯	$X_{2\,m-1}$	$X_{2\,m}$	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	$X_{n\,1}$	$X_{n\,2}$	X_{n3}		$X_{n\,m-1}$	$X_{n\,m}$	A

Fonte: Autoria própria

Estes microarranjos possuem centenas ou milhares de atributos e apenas algumas dezenas de amostras, gerando um alto custo computacional e diminuindo o desempenho de generalização de classificadores, tornando mais difícil a análise destes dados devido a “maldição da dimensionalidade” (BÓLON-CANEDO, 2017). Uma das técnicas utilizadas para preparação dos dados de microarranjos para evitar esta maldição é a redução de dimensionalidade (AHUJA, 2017).

A redução de dimensionalidade, também conhecida como redução de dados vertical, é uma tarefa do pré-processamento que vem ganhando importância por reduzir a quantidade de variáveis ou atributos aleatórios de acordo com o objetivo da tarefa (FAYYAD,1996). Esta redução se faz importante em casos onde as bases de dados possuem o número de atributos (características) muito maior do que a quantidade de instâncias (amostras) (HASTIE *et al*, 2001).

Nestas bases os atributos irrelevantes, que não possuem forte relação para a definição da classe (CHUANG *et al*, 2011), acabam ocupando uma grande quantidade de memória, diminuindo a performance dos algoritmos de aprendizagem como árvores de decisão e regras, regressão linear, aprendizado baseado em instâncias e clusterização (GOLDSCHIMIDT; PASSOS, 2005).

Além disso, umas das preocupações é a alta variância e o sobre ajuste, que ocorre quando o modelo se ajusta muito bem ao conjunto de treinamento, porém, se mostra incapaz de classificar novas instâncias (HASTIE *et al*, 2001). Outros problemas podem ser considerados como o alto nível de ruídos na forma de dados redundantes (AHUJA, 2017).

Segundo Goldschimidt e Passos (2005) a redução de dimensionalidade é implementada pela eliminação ou substituição dos atributos, com o objetivo de encontrar um conjunto mínimo de atributos para que a informação original seja mantida. Outros autores já separam a redução em abordagens que transformam ou projetam os dados originais em um espaço de menor dimensionalidade e abordagens que detectam e removem atributos irrelevantes, fracamente relevantes ou redundantes (HAN; KAMBER; PEI, 2012).

O termo redução de dimensionalidade pode ser também relacionado apenas a técnica de transformação dos atributos (extração de atributos), enquanto a obtenção de um subconjunto dos dados originais é conhecida como seleção de subconjunto ou apenas seleção de atributos (TAN; STEINBACH; KUMAR, 2006). Assim, a redução de dimensionalidade traz algumas vantagens para dados de microarranjos, como por exemplo, a diminuição do custo computacional, redução de ruído para melhorar a acurácia do classificador (MAULIK, 2014).

Para realizar a redução destes atributos é possível aplicar duas das principais abordagens comumente utilizadas, a seleção de atributos e a extração de atributos.

2.2 SELEÇÃO DE ATRIBUTOS

Seleção de atributos é o processo de selecionar um subconjunto com os atributos mais relevantes para construção de modelo ou para interpretação do resultado. Este processo já faz parte da maioria dos algoritmos de aprendizagem de máquina para determinar quais são os atributos mais indicados a fazerem parte de suas decisões, porém, pode haver alguns efeitos negativos na seleção destes algoritmos, como a escolha de algum atributo menos relevante que leva a uma confusão no aprendizado (LIANG et al, 2018).

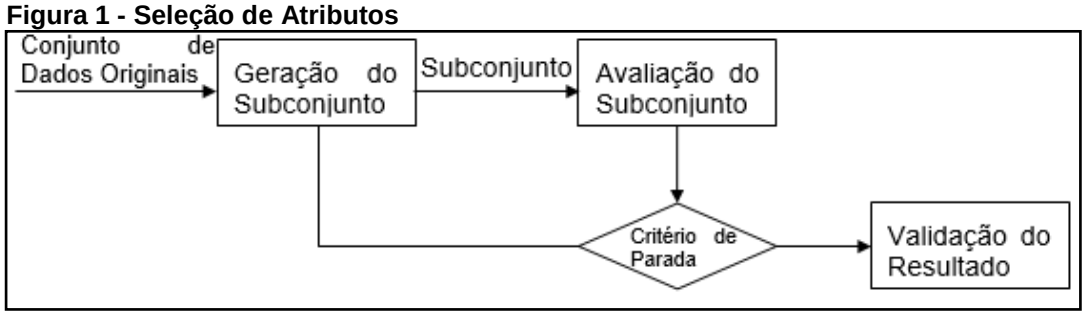
Este processo é importante no pré-processamento, principalmente no reconhecimento de padrões e na descoberta e predição de regras (MAULIK, 2014). Pode trazer resultados imediatos como o aumento da velocidade de algoritmos de mineração de dados, aumento da precisão e a facilidade de compreensão dos resultados (BORGES; NIEVOLA, 2012)

Alguns autores defendem ainda que a seleção ideal é experimentar todos os subconjuntos possíveis como entrada no algoritmo de mineração de dados e

posteriormente escolher o subconjunto com melhor resultado, porém como o número de subconjuntos com n atributos é de 2^n se torna inviável, na maioria das vezes, a prática desta técnica (FREITAS, 2002). Outra alternativa seria realizar a seleção manualmente baseando-se em um conhecimento prévio do problema de aprendizado e o que cada atributo realmente significa, porém, a dificuldade aumenta especialmente se o comportamento dos dados é desconhecido (WITTEN; FRANK; HALL, 2011).

O processo de seleção de atributos é normalmente realizado em quatro passos – geração de um subconjunto, avaliação, análise do critério de parada e validação dos resultados (Figura 1), começando com a estratégia de pesquisa para controlar a geração de um novo subconjunto e depois realizar uma avaliação deste subconjunto. Com isso é analisado um critério de parada e, caso tal critério seja satisfeito, é passado para a última fase com a validação do subconjunto.

Figura 1 - Seleção de Atributos



Fonte: Adaptado de Liu e Motoda (1998)

Há diversas estratégias que podem ser utilizadas para realizar uma busca de subconjunto, devendo-se evitar buscas computacionalmente custosas, e tem como objetivo encontrar um conjunto de atributos ótimos (com maior chance de prever a classe) ou próximos disso (CHUANG et al, 2011). Porém, como é difícil se obter resultados ótimos com baixo custo computacional, é necessário encontrar um balanço entre eles.

Entre as estratégias é possível destacar a busca exaustiva, a seleção para frente (*forward*) e a eliminação para trás (*backward*). É possível encontrar algumas estratégias mais sofisticadas como a combinação da *forward+backward* em uma busca bidirecional, buscas *BestFirst, busca Beam* e buscas baseadas em algoritmos genéticos que serão vistas na seção 2.3 (TAN; STEINBACH; KUMAR, 2006).

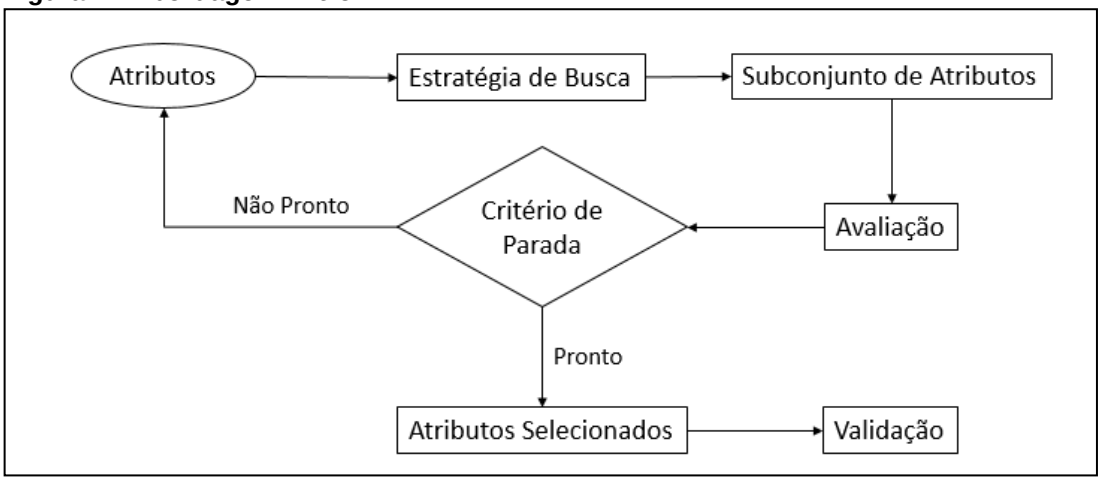
A seleção de atributos pode ser dividida em três abordagens de validação: Filtro, *Wrapper* e Embutido, considerando que alguns autores categorizam a

combinação das duas primeiras. Filtro e *Wrapper* se diferem basicamente na forma de avaliação dos subconjuntos (LIU; YU, 2005).

2.2.1 Filtro

Na abordagem filtro a seleção dos atributos é realizada sem considerar o algoritmo de mineração de dados, avaliando cada atributo (ou um subconjunto de atributos) determinando o nível de correlação entre os atributos e as classes ou avaliando os subconjuntos buscando de forma heurística o melhor, confiando em características gerais dos dados para validar e selecionar subconjuntos (ALMEIDA, 2018; LIU; YU, 2005). Em outras palavras é possível dizer que um método para validar o quão bom é um subconjunto pode ser feito observando apenas características intrínsecas dos dados (BÓLON-CANEDO *et al*, 2014). Na Figura 2 é mostrado um esquema da abordagem filtro.

Figura 2 - Abordagem Filtro



Fonte: Tan, Steinbach, Kumar (2009)

Alguns filtros clássicos são *Correlation Feature Selection (CFS), ReliefF* e Ganho de informação (BÓLON-CANEDO et al, 2014) onde:

- *Correlation Feature Selection (CFS):* Partindo do pressuposto de que um bom subconjunto possui uma alta relação com a classe, mas com baixa correlação entre si, o CFS avalia um subconjunto de atributos baseado na habilidade de predição de cada atributo junto com o grau de redundância entre eles. Ordenando os subconjuntos de acordo com a correlação baseada em uma função de validação

heurística, utilizando a equação de mérito para realizar esta análise (1) (ALMEIDA, 2018; BÓLON-CANEDO *et al*, 2014).

$$Merito_s = \frac{k\bar{r}_{cf}}{\sqrt{k+k(k+1)\bar{r}_{ff}}} \quad (1)$$

O mérito é a heurística do subconjunto S com k atributos tendo $\bar{r}_{cf}$ como media entre a relação atributo-classe e $\bar{r}_{ff}$ como a média da relação atributo-atributo (BORGES; NIEVOLA, 2012).

- *Consistency Subset Evaluation (CSE):* Tem como objetivo encontrar o menor subconjunto de atributos mantendo a consistência original do subconjunto. Um subconjunto é considerado consistente quando, para todas as instâncias, não é encontrado dois ou mais exemplos iguais com classes diferentes (DASH; LIU, 2003).
- *ReliefF:* É, segundo Bólon-Canedo (2014), uma extensão do *relief,* onde a ideia principal é avaliar o quão bem os atributos distinguem entre instâncias de classes diferentes e a qualidade em que eles agrupam instâncias da mesma classe. Utiliza amostras aleatórias de instâncias e localiza seus vizinhos mais próximos, sendo eles de classes iguais ou diferentes, e comparam-se os valores entre eles para então atualizar a pontuação de relevância do atributo (WANG et al, 2005).
- Ganho de informação: Um dos métodos mais comuns de validação dos atributos, que gera uma classificação ordenada de todos os atributos e define uma medida para esses atributos serem escolhidos, como por exemplo, o ganho de informação ser positivo ou ter o maior ganho de informação. Pode ser definido pela diferença entre a incerteza anterior e posterior (BORGES, 2006; BÓLON-CANEDO *et al*, 2014). É baseado no conceito de entropia, em que é feito o cálculo do atributo por (2):

$$H(A) = \sum_{i=1}^{k}[p_i * \log_2 p_i] \quad (2)$$

Em que pi é a probabilidade de $1 \leq i \leq k$ de cada valor que o atributo assume. A entropia da classe (H (C)) pode ser calculada da mesma forma. Já a entropia da classe dado um atributo pode ser calculada com a equação a seguir (ALMEIDA, 2018) (3):

$$H(C|A) = -\sum_{i=1}^{k}\sum_{j=1}^{m}[p_{j|i} * \log(\frac{p_{j|i}}{p_i})] \quad (3)$$

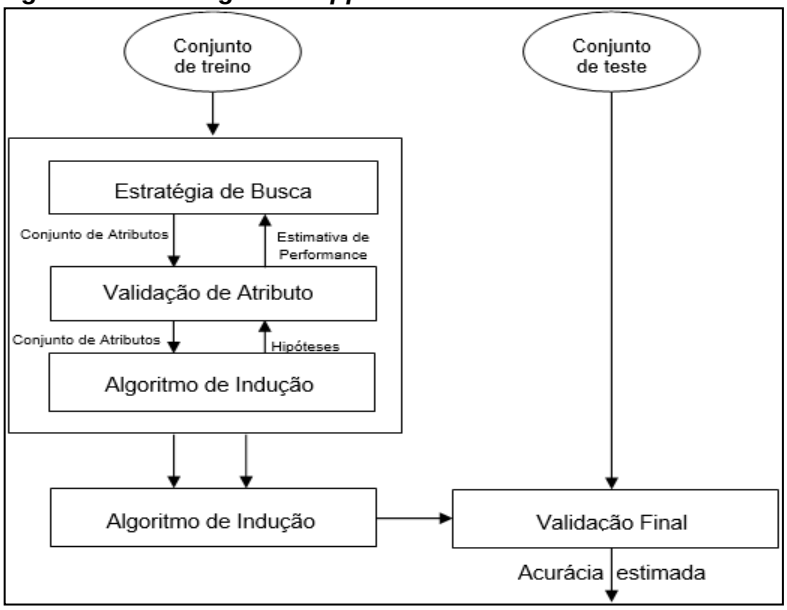
O ganho é definido como a diferença entre a entropia da classe e a entropia condicional da classe (ALMEIDA, 2018) (4):

$$GANHO(C|A) = H(C) - H(C|A) \quad (4)$$

2.2.2 Wrapper

O método *Wrapper* (‘embrulho’) é conhecido assim, pois o algoritmo de indução está envolto com a seleção dos atributos (Figura 3) (WITTEN; FRANK; HALL, 2011) utilizando estes algoritmos, que devem ser o mesmo do alvo, como uma caixa-preta no processo (TAN; STEINBACH; KUMAR, 2006).

Figura 3 - Abordagem *Wrapper*



Fonte: Kohavi; John (1997)

O algoritmo de indução, geralmente os de classificação, é realizado diversas vezes cada vez com um subconjunto de dados diferentes (FREITAS, 2002) com o objetivo de minimizar a taxa de erro do classificador (LIU; MOTODA, 1998). Porém, devido a sua execução repetida, possui um alto custo computacional (ALMEIDA, 2018).

O subconjunto com melhor avaliação é escolhido como o subconjunto final que é utilizado no algoritmo de indução fora do “pacote” (KOHAVI; JOHN, 1997) sendo a maioria das avaliações realizadas de acordo com o nível de precisão de predição feita pelo algoritmo (ALMEIDA, 2018).

O método de indução utilizado deve fornecer uma estimativa melhor do que uma medida independente com uma diferente linha indutiva para que seja vantajoso o uso do método (CHIZI; MAIMON, 2010)

Para a realização da busca pelo melhor subconjunto é necessário determinar os atributos possíveis, um atributo inicial, a condição de parada e qual são os mecanismos de busca (KOHAVI; JOHN, 1997). Sendo utilizadas heurísticas e meta heurísticas como mecanismo de busca (ALMEIDA, 2018).

2.2.3 Embedded

Os métodos embutidos possuem como parte da construção do modelo de classificação, a seleção do subconjunto de atributos mais relevantes (ALMEIDA, 2018). Buscando reduzir o tempo levado para reclassificar os diferentes subconjuntos como no método *Wrapper*, os métodos embutidos determinam se um atributo é relevante durante a execução do algoritmo de classificação (CHANDRASHEKAR, 2014).

Este método pode ser visto em algoritmos de indução de conjunções lógicas, redes neurais, árvores de decisão onde particularmente nas árvores a seleção é realizada no momento de criação do nó (MOLINA; BELANCHE; NEBOT, 2002).

2.3 ESTRATÉGIAS DE BUSCA

Há diversas estratégias que são utilizadas para a realização da busca para tentar encontrar um conjunto de atributos mais próximos dos ótimos possíveis. É encontrar na literatura algumas estratégias como a busca exaustiva, a seleção para frente (*forward*), a eliminação para trás (*backward*), a busca bidirecional, buscas *BestFirst, busca Beam* e buscas baseadas em algoritmos genéticos (TAN; STEINBACH; KUMAR, 2006).

- Busca Exaustiva

Esta busca percorre o conjunto inteiro de atributos, este fato combinado com o crescimento exponencial da dimensionalidade das bases de dados, torna esta busca impraticável (BORGES; NIEVOLA, 2012; WITTEN; FRANK; HALL, 2011);

- Seleção para frente (*Forward*)

Um tipo de busca gulosa que encontra sempre o atributo que mais contribui para a melhora da performance do subconjunto, se inicia com um subconjunto de atributos vazio e vai sendo adicionado um a um validando este atributo. Caso seja

escolhido, é incorporado ao subconjunto e passa para o próximo atributo que ainda não foi escolhido. Este processo é repetido até que não se possa melhorar o subconjunto (BORGES; NIEVOLA, 2012; HAN; KAMBER; PEI, 2012; ALMEIDA, 2018);

- Eliminação para trás (*Backward*)

É a técnica contrária à *forward,* começando com um subconjunto com todos os atributos que são removidos um por vez até que o subconjunto não melhore (BORGES; NIEVOLA, 2012);

- Busca bidirecional

Realiza uma combinação das duas técnicas anteriores, podendo ser iniciado com um conjunto vazio ou não e em cada interação o algoritmo seleciona o melhor atributo, removendo o pior dentre os atributos restantes (HAN; KAMBER; PEI, 2012);

- *Best First*

É semelhante ao *forward*, porém, além de acabar quando a performance começa a diminuir, armazena uma lista de todos os subconjuntos já validados e ordenados por ordem de medida de performance (WITTEN; FRANK; HALL, 2011);

- Busca *Beam*

A busca *Beam* (‘feixe’) funciona de forma similar ao *Best First* mas realiza truncamentos das listas de subconjuntos de atributos onde em cada passo, apenas um número fixo de atributos mais promissores é mantido (WITTEN; FRANK; HALL, 2011);

- Algoritmos Genéticos

Esta técnica é baseada na seleção natural tendo a evolução baseada em aleatoriedade de um bom subconjunto, além da combinação dos subconjuntos baseada no desempenho de cada um (WITTEN; FRANK; HALL, 2011).

2.4 EXTRAÇÃO DE ATRIBUTOS

O processo de extração de atributos tem como objetivo de alterar a forma de representar os dados, criando novas variáveis a partir de combinações de outras pertencentes ao conjunto original, para que o número de atributos seja menor, porém, sem perder as características dos dados (CLARKE *et al*, 2009).

Matematicamente pode-se dizer que realiza a transformação de um vetor N-dimensional $X = [x_1, x_2, x_3, ..., x_n]^T$ por um mapeamento f e mapeá-lo em um vetor m-dimensional $Y= [y_1, y_2, y_3, ..., y_m]^T$ com n>m (DING et al, 2011). Segundo Kohavi e John (1998), a redução é chamada de linear caso $f(x) = Ax$, para toda matriz $A_{m;n}$ e não linear caso contrário.

2.4.1 Não linear

Na redução não linear, o mapeamento dos dados tem com o objetivo de manter na nova representação algum padrão de distância ou topologia dos dados sem se limitar à linearidade da transformação (LEE; VERLEYSEN, 2007).

Tem como vantagem o alto poder de discriminação dos atributos extraídos, o que gera uma boa taxa de classificação, e o controle de '*overfitting'* que diminui as chances de um modelo se adequar muito bem a uma base de treinamento enquanto tem baixa taxa de acerto na base de testes.

Porém, uma desvantagem é o fato de perder a interpretabilidade dos dados, além de que esta transformação pode ser custosa do ponto de vista computacional (AZIZ; VERMA; SRIVASTAVA, 2016).

2.4.2 Linear

A extração linear assume que os dados estão em um subespaço linear de dimensão d, projetando-os neste subespaço utilizando uma fatoração matricial, em que com um conjunto de dados X_{mxn} e uma dimensão escolhida r<d para produzir uma transformação linear P_{rxd}, obtendo $Y_{rxn} = PX$ como os dados projetados em uma dimensão menor (CUNNINGHAM; GHAHRAMANI, 2015).

Estes métodos assumem que os dados estão linearmente separados no subespaço em que foi projetado. Uma forma interessante de se realizar esta projeção é analisar os principais componentes que represente uma proporção da variância dos dados (CLARKE et al, 2009).

2.4.2.1 Análise de Componentes Principais (PCA)

Esta técnica busca reduzir os dados em que as variáveis têm um alto grau de dependência entre si, mantendo o máximo da variação. Para isso realiza uma transformação dos atributos em componentes principais que não possuem relação entre si e são ordenados em relação à variação presente nos dados originais. Então os *n* componentes principais contém a maior parte da informação nos *m* dados originais (*n*>*m*).

A transformação é realizada obtendo *m* vetores ortonormais da base de dados normalizada, em que estes vetores (os componentes principais) são ordenados em ordem decrescente de variância para então, o conjunto de dados possa ser reduzido ao realizar a remoção de componentes com variação mais baixa (HAN; KAMBER; PEI, 2012).

Estes componentes são extraídos por meio de autovalores e auto vetores que podem ser obtidos segundo Nurfalah, Adiwijaya e Ardiyanti (2016), com o cálculo da média $\bar{X}$ do conjunto de entrada (5):

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \tag{5}$$

Sendo X_i os atributos de entrada e *n* o tamanho da entrada. Após encontrada a média é possível calcular a matriz de covariância C_x (6):

$$C_x = \sum_{i=1}^{n} \frac{(X_i - \bar{X}) - (X_i - \bar{X})^T}{n-1} \tag{6}$$

A partir da matriz de covariância é possível calcular os autovalores e auto vetores com (7):

$$C U_n = \lambda_n U_n \tag{7}$$

Onde U é o auto vetor e λ o auto valor.

Com isso é possível realizar a transformação dos dados utilizando a fórmula (8):

$$Y = U^T * (X - \bar{\bar{X}}) \tag{8}$$

Umas das vantagens do PCA é a extração sequencial da maior parte da variação dos dados, assim espera-se que apenas os últimos (menos que os *m* atributos originais) contenham a maior parte das informações. Além disso, se torna de mais fácil interpretação devido à natureza ortogonal dos dados extraídos (HAND; MANNILA; SMYTH, 2001).

2.5 CONSIDERAÇÕES DO CAPÍTULO

Este Capítulo apresentou uma visão geral sobre o que são atributos, bases de dados e microarranjos. Além disso, foram introduzidos os conceitos de redução de dimensionalidade, passando por suas principais abordagens (Seleção e Extração de Atributos) e suas técnicas (Filtro, *Wrapper* e *Embedded* para Seleção e Linear e não Linear para Extração).

3 MAPEAMENTO SISTEMÁTICO

Este Capítulo traz um levantamento bibliográfico com o objetivo de estabelecer o atual estado da arte a respeito da redução de dimensionalidade em microarranjos de DNA. A Seção 3.1 traz uma breve introdução sobre o método de revisão da literatura, enquanto a Seção 3.2 traz a aplicação do método. Já a Seção 3.3 apresenta um as considerações a respeito do capítulo.

3.1 DESCRIÇÃO DO MÉTODO DE MAPEAMENTO SISTEMÁTICO

Foi realizado um método de revisão sistemática, apresentando etapas de classificação das atividades que tornam a busca mais focada e organizada, facilitando o encontro de resultados mais efetivos. Não há uma ordem exata dos passos a serem seguidos, podendo ser adaptado por cada pesquisador. As etapas desse processo são (WAZLAWICK, 2017):

- Estabelecimento da Intenção de Pesquisa: delimitação do tema e elaboração das questões de pesquisa a serem trabalhadas pela revisão sistemática. São definidas as características da pesquisa para classificá-la como mapeamento sistemático.
- Definição das Palavras-chave e Bases de Dados: palavras utilizadas para a realização de uma busca preliminar exploratória com o objetivo de validar os resultados obtidos pelas mesmas, avaliando a necessidade de troca ou combinações de palavras-chave.
- Pesquisa na Base de Dados: utilizando as palavras-chave definidas anteriormente, são realizadas as buscas efetivas nas bases de dados para a definição de um gerenciador de bibliografia.
- Procedimento de Filtragem: primeiras filtragens de todos os resultados obtidos nas bases de dados, descartando resultados duplicados, trabalhos que não possuem nenhuma relação com o tema definido na etapa 1, trabalhos apresentados em conferência e capítulos de livros, além de outros filtros foram definidos (ALMEIDA, 2018).
- Leitura dos Resumos: em função dessa leitura, classificar os artigos de acordo com a relevância com o tema, sendo elas “alta”, “médica” e “baixa”.

- Leitura de Alta Relevância: leitura dos artigos classificados com “alta” relevância, fazendo anotações sobre os principais conceitos e aquilo que pareça relevante para o tema. Separá-los na lista de artigos que devem ser lidos por completo. Se for necessário, os resumos dos artigos de “média” e “baixa” podem ser lidos, começando sempre pela que possui maior relevância.

Caso o pesquisador ache necessário, é possível realizar uma análise das referências bibliográficas citadas nos trabalhos considerados de “alta” relevância, aumentando assim a base de dados. Após a elaboração de todos esses passos, o pesquisador decide se o material adquirido é o suficiente para dar continuidade à pesquisa, devendo essa decisão ser tomada em conjunto com um orientador (WAZLAWICK, 2017).

3.2 APLICAÇÃO DO MÉTODO

Esta Seção apresenta os passos realizados para a aplicação do método de mapeamento sistemático. Inicialmente com as questões de pesquisa, seguida das bases de pesquisas utilizadas. Os termos de pesquisa são detalhados e logo depois é possível encontrar a realização das buscas e os procedimentos de filtragem do material encontrado. Por último são mostrados os resultados do mapeamento, respondendo assim as questões definidas.

3.2.1 Questão de pesquisa

Foi realizado um mapeamento sistemático focado em identificar as características dos métodos buscando por novos métodos de seleção ou possíveis melhorias em métodos já disponíveis, sendo necessário a elaboração de perguntas que devem ser respondidas acerca de cada trabalho encontrado, apresentado no Quadro 1.

As questões elaboradas tem como objetivo levantar as principais características dos trabalhos encontrados por meio das buscas. Estes questionamentos são importantes para a análise de abordagens já exploradas anteriormente.

Quadro 1 - Definição das perguntas do mapeamento sistemático

ID	Pergunta
Q1	Quais as abordagens de redução de dimensionalidade utilizadas? (Seleção/Extração de atributos)
Q2	Quais as bases de dados de microarranjos utilizadas?
Q3	Qual foi a contribuição científica dos autores em cada trabalho escrito por eles? (Criação de um novo método ou aprimoramento de um método existente)
Q4	Quais os classificadores e as medidas de avaliação utilizados?
Q5	O resultado obtido pelo conjunto de dados reduzidos foi superior ao conjunto de dados formados por todos os atributos?
Q6	Qual a quantidade de trabalhos referentes ao tema da pesquisa foram publicados por ano?

Fonte: Autoria própria

3.2.2 Seleção das bases de pesquisa

Para iniciar as buscas foram definidas as bases de dados utilizadas no decorrer do processo. Foram escolhidas bases que atendessem os seguintes requisitos de filtragem: seleção de publicações na área de Ciência da Computação, artigos na língua inglesa e realizados por período de publicação.

Para a definição da base de pesquisa para esse processo foi realizado buscas no sistema CAPES, podendo ser analisado no Quadro 2 a relação da base de pesquisa, seus respectivos endereços e filtros utilizados.

Quadro 2 - Definição das bases de pesquisa

Base de Pesquisa	Sites	Filtros utilizados
Science Direct	<https://www.sciencedirect.com>	*"Title, abstract or keywords"*
Web of Science	<https://www.webofknowledge.com>	*"Topic"*
Springer	<https://www.springer.com>	*"Title"*
InderScience	<https://www.inderscienceonline.com>	*"Anywhere"*
IEEE	<https://ieeexplore.ieee.org/Xplore/home.jsp>	*"Abstract"*
Scopus	<https://www.scopus.com>	*"Title, abstract or keywords"*

Fonte: Autoria própria

3.2.3 Definição dos termos de busca

Etapa de definição das palavras-chave que compõem as questões anteriormente definidas. Essas palavras são utilizadas na execução das pesquisas

buscando resultados mais precisos dentro da área estudada. As palavras definidas nesse estudo estão apresentadas no Quadro 3.

Quadro 3 - Definição das palavras-chave

ID	Língua Portuguesa	Língua Inglesa
1	Microarranjo	*"Microarray"*
2	Seleção de Atributos	*"Attribute selection"*
3	Extração de Atributos	*"Feature extraction"*
4	Redução de Dimensionalidade em bases de dados	*"Dimensionality reduction in database"*
5	Seleção de Atributos	*"Feature selection"*

Fonte: Autoria própria

Com as palavras-chave definidas, é possível analisar a combinação das mesmas, utilizando um operador lógico AND. Através dessa combinação de termos, obteve-se *strings* de busca, que são apresentadas no Quadro 4.

Quadro 4 - Definição das strings de busca

ID	String de Busca
S1	*"Microarray" AND("Attribute selection" OR "Feature selection")*
S2	*"Microarray" AND "("Feature extraction" OR "Attribute extraction")"*
S3	*"Microarray" AND "Dimensionality reduction in database"*
S4	*"Microarray" AND ("Feature selection" OR "Attribute selection") AND ("Feature extraction" OR "Attribute extraction")*

Fonte: Autoria própria

Além do método de pesquisa já apresentado, foi realizado buscas na ferramenta *Google Scholar,* utilizando as *strings* apresentadas no Quadro 5, com o objetivo de filtrar artigos publicados em conferências e/ou seminários.

Quadro 5 - Definição das strings de buscas da pesquisa complementar

ID	String de Busca
S5	*"Attribute selection in microarray"*
S6	*"Feature extraction in microarray"*
S7	*"Dimensionality reduction in microarray"*

Fonte: Autoria própria

3.2.4 Realização das buscas

Buscas realizadas no período de 2010 a 2018 apenas com os termos em inglês. Na Tabela 2 é possível analisar que a *string* S1 é a que retorna a maior quantidade de trabalhos, por este motivo, foi realizada a busca complementar no *Google Scholar* como pode ser visto na Tabela 3.

Tabela 2 - Total de resultados por string

Base de pesquisa	S1	S2	S3	S4	Busca primária
Science Direct	114	17	3	6	140
Web of Science	391	55	3	22	471
SpringerLink	18	3	0	2	23
InderScience	126	53	0	38	217
IEEE	43	1	0	7	51
Scopus	337	211	0	178	726
Total	1029	340	6	253	1628

Fonte: Autoria própria

A tabela 3 traz os resultados obtidos na busca complementar utilizando as *strings* 5, 6 e 7.

Tabela 3 - Total de resultados pela busca complementar

Base de pesquisa	S5	S6	S7	Total de trabalhos
Google Scholar	0	0	2	2

Fonte: Autoria própria

3.2.5 Procedimento de filtragem

Para a etapa de filtragem, foram definidos os seguintes critérios.

1. Exclusão de artigos duplicados pela junção dos resultados obtidos nas bases de dados.
2. Eliminação de todas as publicações de livros, capítulos de livros e conferências.
3. Eliminação de artigos com índice *Qualis* menor do que B2
4. Exclusão de artigos que não possuem relação com o tema deste trabalho, levando em consideração para isso os seus títulos, resumos e palavras-chave.
5. Exclusão de artigos que não possuem avaliação quantitativa dos resultados.

Para a realização das etapas de filtragem foi criado uma biblioteca na ferramenta *Zotero* e feita a junção dos dados obtidos pelas buscas principais, os resultados das etapas 1 e 2 da filtragem pode ser observada na Tabela 4.

Tabela 4 - Aplicação das etapas 1 e 2 do processo de filtragem

String	Busca Primária	Filtragem 1 e 2
S1	1029	459
S2	340	65
S3	6	4
S4	253	76
Total	1628	604

Fonte: Autoria própria

Após a realização da filtragem 1 e 2, os resultados são classificados de acordo com seu qualis Capes (de A1 à C) e são removidos trabalhos publicados em publicações com avaliação menor que o qualis B2. Os resultados são apresentados na Tabela 5.

Tabela 5 - Aplicação da etapa 3 do processo de filtragem

String	A1	A2	B1	B2	Total
S1	108	71	39	4	222
S2	11	14	7	3	35
S3	1	0	0	0	1
S4	20	14	4	2	40
S5	0	0	0	0	0
S6	0	0	0	0	0
S7	0	0	0	0	0
Total	140	99	50	9	298

Fonte: Autoria própria

Após este procedimento, é realizada a filtragem 4 em que foram analisados os títulos, resumos e palavras-chave, sendo retirados resultados que não possuem relação com o tema pesquisado e a filtragem 5 que remove artigos que não possuem uma avaliação quantitativa dos resultados. Como poder ser visto na Tabela 6, após os procedimentos de filtragem 4 e 5 foram removidos 290 trabalhos, mantendo apenas os trabalhos a serem lidos integralmente.

Tabela 6 - Aplicação da etapa 4 e 5 do processo de filtragem

Qualis	Filtragem 4 e 5
A1	4
A2	2
B1	2
B2	0
Total	8

Fonte: Autoria própria

Os artigos de periódicos encontrados são listados no Quadro 6.

Quadro 6 - Artigos de Periódicos

ID Publicação	Autor	Ano	Título
1	Borges, H. B., Nievola, J. C	2012	*Comparing dimensionality reduction methods in gene expression databases*
2	Nanni, L., Brahnam, S., Lumini, A.	2012	*Combining multiple approaches for gene microarray classification*
3	You, W., Yang, Z., Yuan, M., Ji, G.	2014	*TotalPLS: Local Dimension Reduction for Multicategory Microarray Data*
4	Mollaee, M., Moattar, M. H.	2016	*A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification*
5	Guo, S., Guo, D., Chen, L., Jiang, Q.	2017	*A L1-regularized feature selection method for local dimension reduction on microarray*
6	Latkowski, T., Osowski, S.	2017	*Gene selection in autism – comparative study*
7	Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y.	2017	*A Hybrid Feature Selection Algorithm for Gene Expression Data*
8	Sreevani, Murthy, C.A., Chanda, B.	2018	*Generation of compound features based on feature interaction for classification*

Fonte: Autoria própria

3.2.6 Resultados

Nesta etapa da revisão sistemática são apresentadas as respostas obtidas para as questões definidas pela seção 3.2.1. Para isso as questões são identificadas pelos identificadores Q1 a Q6.

Q1 - Quais as abordagens de redução de dimensionalidade utilizadas? (Redução ou extração de atributos)

Em Borges e Nievola (2012) é realizada uma comparação em relação à acurácia de classificação a partir de bases de dados de microarranjo de DNA que tiveram a quantidade de atributos reduzidos com a seleção de atributos e com o método de projeção aleatória. Para a seleção de atributos, usando a abordagem filtro, foram utilizadas as buscas sequencial e aleatória juntamente com as medidas de dependência e consistência. Além disso, aplicou a abordagem *wrapper* utilizando os classificadores com diferentes metodologias como árvores de decisão, classificadores probabilísticos e não probabilísticos.

Em Nanni, Brahnam e Lumini (2012) primeiramente são comparados alguns algoritmos de seleção de atributos como *Fisher Score*, *Gini Index*, mRMR, *T-test* e SB. Na segunda etapa o algoritmo com melhor resultado, no caso Fisher Score, na fase anterior é escolhido para compor os testes com a aplicação da extração de atributos nos dados já selecionados. Para comparação, são aplicados outros algoritmos para transformar os dados como *Locally Lunear Embendding* (*LLE*), *Orthogonal LDA* (*OLDA*), *Orthogonal Neighborhood Preserving Projections* (*ONPP*s) e Neighborhood Preserving Embendding. Obtiveram bons resultados em comparação a outros algoritmos já consolidados.

Em You et al (2014) é proposto um *framework* onde primeiro é realizada uma seleção de atributos na base de alta dimensionalidade para se obter um subconjunto de informações importantes. Depois é aplicada a extração neste subconjunto encontrado na primeira fase. Na fase da seleção é utilizado o algoritmo PLSRFE (*PLS-based feature selection*) e depois o subconjunto gerado é transformado por meio da *PLS-based Feature Extraction* e este *framework* foi chamado de *TotalPLS*.

Em Mollaee e Moattar (2016) foi proposto um *framework* (*PSO-dICA*) composto por uma primeira etapa que realiza uma seleção hibrida de atributos com alguns métodos *ranking* como *Bayesian Logistic Regression* (*BLOGREG*), *T-test* e *Fisher-Test* e selecionando os atributos com a maior soma dos resultados dos 3 algoritmos aplicados. Na segunda etapa é aplicado o método *Discriminant Independent Component Analysis* (dICA) e para contornar a limitação de métodos baseados em gradiente de custo como o *dICA*, utiliza o algoritmo de otimização

Particle Swarm Optimization como função de ajuste para aumentar a sintropia e a taxa de *Fisher* ao mesmo tempo.

Em Guo et al (2017) propõe um método de redução de dimensionalidade que une a seleção de atributos *embedded* chamada Regularização L1, que utiliza regressão logística para tentar encontrar um subconjunto ótimo a partir da separabilidade das classes, com o método de extração de atributos chamada *Partial Least Square* (PLS) que busca encontrar uma decomposição linear para os atributos. É um método não dependente do classificador, por isso foram utilizados um teste estatístico de significância e teste-t para comparação com métodos da literatura.

Em Latkowski e Osowski (2017) foram realizados testes em base de dados da área de bioinformática aplicando separadamente diversos algoritmos de seleção de atributos como *Fisher Discriminant Analysis*, *ReliefF*, *Two Sample two-test, Kolmogorov-Smirnov test, Kruskal-Wallis test, Stepwise Regression method, Feature Correlation with Class e Support Vector Machine.* Com cada subconjunto gerado foi realizada a extração dos atributos com os algorítmos *K-means, Random Forest e* Algoritmo Genético*.*

Em Lu et al (2017) é proposta uma abordagem de seleção de atributos híbrida. Nesta abordagem inicialmente é realizada a seleção dos genes com maior correlação por meio do algoritmo *Mutual Information Maximization*. Depois, é realizada a extração de atributos com os subconjuntos gerados com a seleção utilizando o algoritmo genético adaptativo. Os experimentos foram realizados com bases de dados de diversos tipos de câncer e encontraram uma ótima compatibilidade do método proposto com o classificador *Extreme Learning Machine*, obtendo os melhores resultados quando comparados a outras abordagens de redução de dimensionalidade e outros classificadores.

Sreevani, Murthy e Chanda (2018) propõe um algoritmo de redução de dimensionalidade chamado *Information Maximization and Redundancy Maximization through Feature Interaction for CFG* (*Compount Feature Generation*), onde consegue diminuir a dimensionalidade a partir da produção do máximo de semi-atributos informativos, da remoção de atributos irrelevantes e depois considerando os atributos redundantes, selecionar os mais significativos. O algoritmo gera os chamados atributos compostos, uma combinação de atributos transformados e selecionados utilizados posteriormente para classificação das instâncias.

Q2 - Quais as bases de dados de microarranjos utilizadas?

Foram identificadas nos estudos 26 bases de dados utilizadas para realização de testes. Na tabela 8 são apresentadas as bases em ordem alfabética.

Tabela 7 - Bases de dados x quantidade de trabalhos

Base de dados	Total
ALML	1
ALL-AML	1
Brain	1
Breast	2
CLL-SUB-111	2
Colon	3
DLBCL	4
DLBC-*Tumor*	1
DLBC-*Outcome*	1
DLBCL-NIH	1
Duke	1
GCM	2
GLA-BAR-180	1
Glioma	1
Leukemia	1
Leukemia-ALL	1
Lung Cancer	6
Medullobastoma	1
MLL	2
NCBI	1
NCI60	2
Ovarian	1
Prostate	3
SBRCT	3
Stjude	1
TOX-171	1

Fonte: Autoria própria

É possível observar uma grande variedade nas bases de dados utilizadas, porém nota-se que a maioria das bases que são utilizadas em mais de um trabalho são bases ligadas a algum tipo de câncer.

Q3 - Qual foi a contribuição científica dos autores em cada trabalho escrito por eles?

Os artigos encontrados podem ser divididos de acordo com sua contribuição científica. No primeiro grupo estão os trabalhos que desenvolvem novas técnicas de redução de dimensionalidade. No segundo os trabalhos que realizam adaptações em técnicas de redução já disponíveis na literatura. Esta divisão pode ser observada no Quadro 7.

Os trabalhos são identificados pelo ID especificados na tabela 6.

Quadro 7 - Tipos de contribuição científica x Artigos selecionados

Tipo de contribuição	ID Publicação
Criação de um novo método	6
	5
	2
	4
	7
Aprimoramento de um método existente	1
	8
	3

Fonte: Autoria própria

Q4 - Quais os classificadores e as medidas de avaliação utilizadas?

As medidas de avaliação utilizadas são pouco variadas, onde Borges e Nievola (2012) utiliza a validação cruzada estratificada 10-vezes. Latkowski, Osowski (2017) e Nanni, Brahnam, Lumini (2012), utilizam a validação cruzada 10-vezes e Guo et al (2017) com apenas 5-vezes. Para fortalecer os resultados, You et al utiliza a validação cruzada *holdout* e a validação cruzada k-vezes.

A leitura e análise dos estudos identificou a utilização de 10 classificadores diferentes nos 8 trabalhos selecionados. Na Tabela 8 são apresentados os classificadores, as medidas de avaliação e a quantidade de trabalhos por técnica encontrada.

Tabela 8 - Técnicas de classificação x Número de trabalhos

Técnica	Total
C5	1
Extreme Learning Machines	1
KNN	2
LDA	2
Multiclass Support Vector Machine	1
Naive Bayes	1
Random Forest	1
Redes Neurais	1
Regularized Extreme Learning Machines	1
Support Vector Machines	5

Fonte: Autoria própria

Q5 - O resultado obtido pelo conjunto de dados reduzidos foi superior ao conjunto de dados formados por todos os atributos?

Em Borges e Nievola (2012), com a comparação dos métodos de redução de dimensionalidade, foi observado que as taxas de acerto dos algoritmos de classificação aumentam significativamente quando aplicados em atributos já selecionados. Tendo encontrado como melhor combinação a que foi construída com a abordagem *Wrapper* aplicado juntamente com a busca sequencial. Foi observado que o algoritmo de Projeção Aleatória fica como uma alternativa, tendo visto que o método de seleção de atributos se mostra mais efetivo.

Em Nanni, Brahnam e Lumini (2012) não foi realizada a classificação com todos os atributos, os resultados foram apenas comparados com resultados de outros estudos já publicados. Encontrou bons resultados quando avaliados acurácia e AUC tanto da seleção com *Fisher Score* (Fi) quanto com a combinação do Fi com *Neighborhood Preserving Embendding* (NPE) que obteve os melhores resultados durante os testes. Além disso, validou a ideia de combinar diferentes abordagens por meio da estatística-Q.

You et al (2014) obteve melhores resultados do que a combinação de outros métodos de redução (*F-test, Relief, MSVMRFE*) com a extração. O ponto principal da abordagem é a redução considerável do número de atributos redundantes, aprimorando a semelhança entre amostras dentro de uma mesma categoria.

Em Mollaee e Moattar (2016) foi comparado alguns métodos conhecidos na literatura com foco no algoritmo *dICA*, para observar o comportamento da aplicação do método PSO. Foi observado que o *framework* proposto (Seleção de atributos + *PSO-dICA+SVM*) obtém melhores resultados do que a combinação da seleção de atributos com outros métodos de extração principalmente quando comparados aos resultados da aplicação apenas do *dICA + SVM*.

Em Guo et al (2017) não foi realizada a classificação utilizando todos os atributos. Porém, realizou a classificação de instâncias que tiveram a dimensionalidade reduzida por meio de diferentes abordagens além da abordagem proposta. Os autores verificaram que o algoritmo proposto obteve resultados estatisticamente semelhantes ou superiores do que abordagens encontradas na literatura. Além disso, a abordagem proposta se mostra bastante eficiente quando utilizada em dados com muitos ruídos e *outliers.*

Latkowski e Osowski (2017) analisou, a partir de comparação entre diversas combinações, o desempenho da fusão dos resultados obtidos pela seleção de atributos utilizando três algoritmos de extração. Na primeira etapa encontrou um melhor resultado em acurácia e especificidade na fusão dos seis melhores métodos utilizando *Random Forest*. Utilizou referências na literatura para determinar que o método proposto possui uma acurácia média superior às outras formas de redução de dimensionalidade semelhantes.

Lu et al (2017) compara o método proposto (MIMAGA – *Mutual Information Maximization with Adaptive Genetic Algorithm*) com outros algoritmos conhecidos de redução de dimensionalidade (*Relief*, SFS (*Sequencial Forward Selection*) e *Mutual Information Maximization* (*MIM*)), obtendo maior acurácia que os algoritmos comparados. Além disso utilizou os dados reduzidos pelo método proposto em 4 classificadores diferentes: *Extreme Learning Machine* (*ELM*), SVM, *Regularized Extreme Learning Machine* (RELM) e rede neural com *back propagation*. Chega à conclusão que o método proposto funciona melhor quando utilizado em conjunto com o classificador *RELM*.

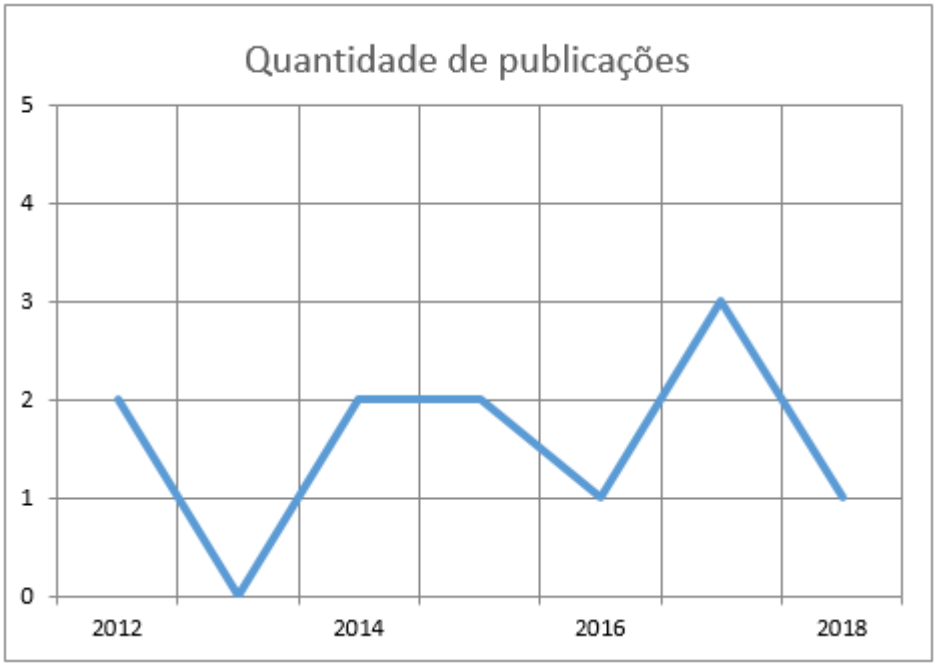
Apesar de não realizar a comparação com a base completa, Sreevani, Murthy e Chanda (2018) realizaram por meio de comparação com outros métodos de seleção e extração de atributos como *Relief*, *MRSF* (*Minimum Redundancy Spectral Feature selection*), entre outros. Utilizou os classificadores 1-NN e C4.5 para obter resultados que mostram que o método proposto possui superioridade sobre os métodos de

seleção *ReliefF*, *Trace Ratio* (TR) e *MRSF* além dos métodos de extração *MFA*, *ALDE* (*Angle Linear Discriminant Embendding*) e *SLDA* (*Sparse Linear Discriminant Analysis*). Além disso possui uma performance equiparável com o método *MMC* (*Maximum Margin Criterion*).

Q6 - Qual a quantidade de trabalhos referentes ao tema da pesquisa foram publicados por ano?

A Figura 4 apresenta a distribuição de trabalhos encontrados por ano de publicação.

Figura 4 - Quantidades de publicações por ano



Fonte: Autoria Própria

3.3 CONSIDERAÇÕES DO CAPÍTULO

Neste Capítulo foram apresentados os trabalhos relacionados a partir da revisão sistemática e buscas complementares na literatura por artigos publicados em periódicos.

Com a revisão sistemática foi possível identificar, analisar e interpretar todos os artigos de periódicos relevantes para o problema definido. Com a busca

complementar foi possível verificar e acrescentar no levantamento bibliográfico outros trabalhos relacionados a esta pesquisa.

Após o levantamento, foi verificado a ausência de trabalhos em que são aplicadas as técnicas de seleção de atributos em conjunto com a extração de atributos utilizando PCA.

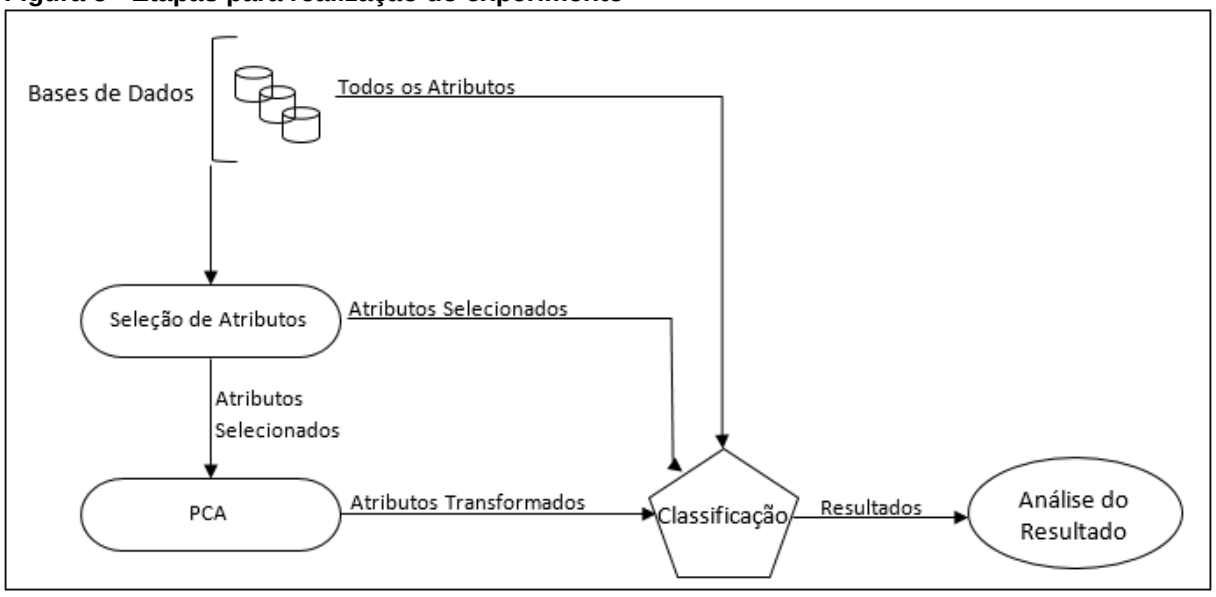
4 REDUÇÃO DE DIMENSIONALIDADE EM BASES DE DADOS DE MICROARRANJOS

Este Capítulo apresenta os procedimentos aplicados para a realização dos experimentos de redução de dimensionalidade em bases de dados de microarranjo. Na Seção 4.1 é apresentada a metodologia empregada para a realização dos experimentos. A Seção 4.2 traz uma visão geral sobre as bases de dados utilizadas. As Seções 4.3 mostra a aplicação da seleção. A Seção 4.4 demonstrada como foram aplicadas as abordagens de seleção e extração em conjunto. A Seção 4.5 traz os algoritmos de classificação utilizados e, por fim, a Seção 4.6 apresenta as considerações do capítulo.

4.1 METODOLOGIA

Para a realização dos experimentos foram seguidos as seguintes etapas: 1 - Classificação da base original com todos os atributos, 2 - Aplicação dos Métodos de Seleção de Atributos e a classificação dos subconjuntos e 3 - Aplicação em conjunto dos métodos (Seleção de Atributos e PCA) e a classificação dos subconjuntos. Ao fim destes processos é realizada a análise dos resultados, como pode ser visualizado na Figura 5.

Figura 5 - Etapas para realização do experimento



Fonte: Autoria Própria

Na primeira etapa é realizada a classificação utilizando todos os atributos para melhores comparações dos resultados. Na segunda etapa foram aplicados os métodos de seleção de atributos na base original e os subconjuntos gerados foram utilizados para a classificação. Na segunda etapa aplicou-se o PCA nos atributos selecionados na etapa anterior e em seguida os classificou. Por fim os resultados das classificações foram comparados a fim de analisar a taxa de acerto obtida pelos subconjuntos.

Foi utilizada a versão 3.8 da ferramenta *Weka* (*Waikato Environment for Knowledge Analysis*). Este é um *software* de código aberto desenvolvido em linguagem JAVA pela Universidade de Waikato que reúne diversos algoritmos para realização da tarefa de mineração de dados. Os parâmetros dos algoritmos aplicados foram mantidos com valores padrões.

4.2 DESCRIÇÃO DOS CONJUNTOS DE DADOS

Foram selecionadas 5 bases de dados de microarranjo (BORGES, 2006). Os dados são disponibilizados em sua forma bruta e em formato próprio para análise no software *Weka* ("*arff*").

As bases foram analisadas a fim de se verificar a quantidade de atributos, número de amostras e a divisão das amostras de acordo com sua classe como.

Tabela 9 - Conjuntos de dados

Conjuntos de dados	Amostras	Atributos
AML-ALL (GOLUB T. et al, 1999)	72	7129
DLBCL (ALIZADEH et al, 2000)	47	4026
DLBCL-NIH (ROSENWALD et al, 2002)	240	7339
DLBCL-Tumor (SHIPP et al, 2002)	77	7129
DLBCL-Outcome (SHIPP et al, 2002)	58	7129

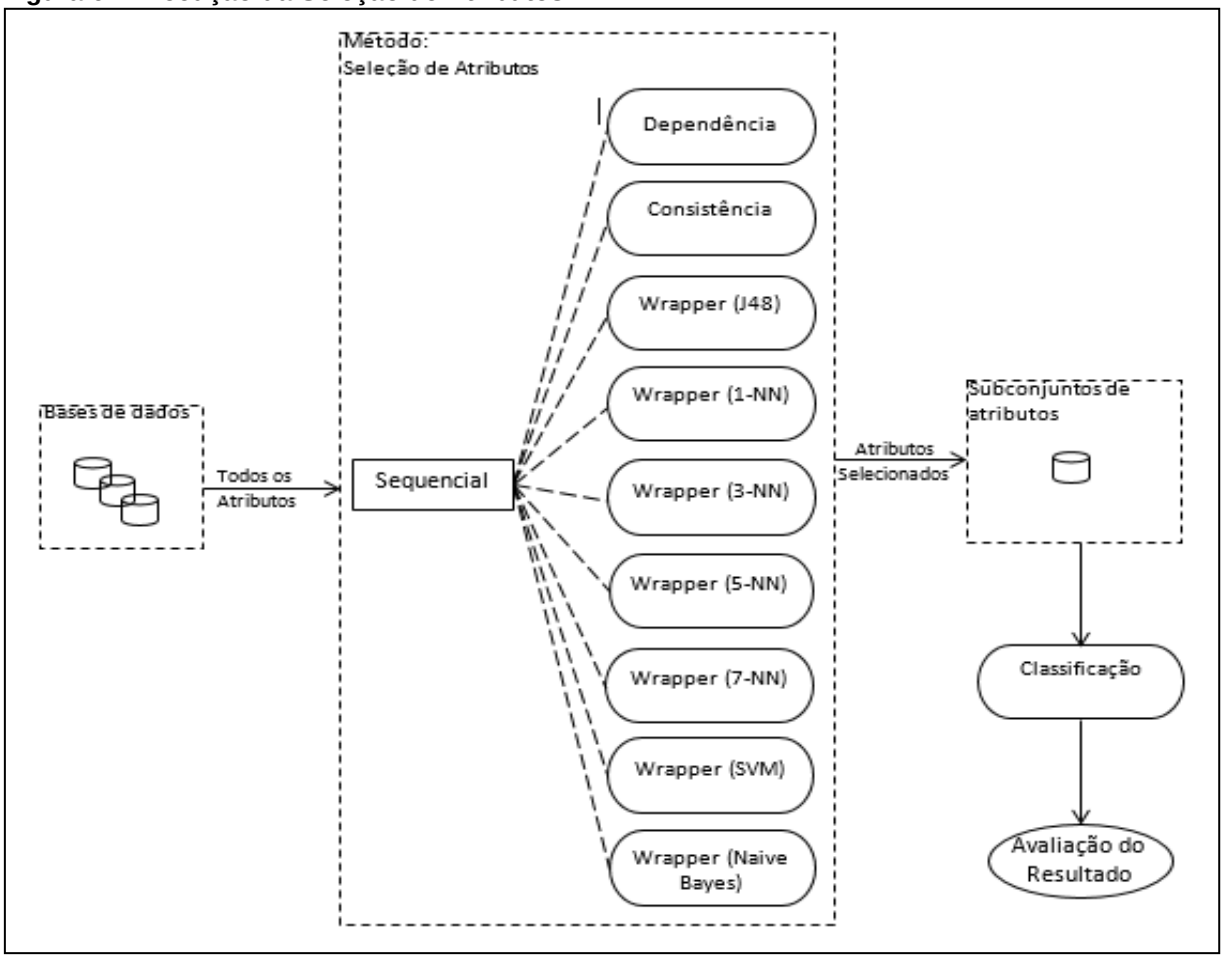
Fonte: Autoria própria

A Tabela 9 apresenta a relação entre quantidade de amostras e atributos de cada conjunto de dados utilizados.

4.3 EXECUÇÃO DA SELEÇÃO DE ATRIBUTOS

Para a execução da seleção de atributos foram aplicadas duas abordagens: Filtro e *Wrapper*. Para a abordagem filtro foram escolhidas as medidas de consistência (CSE) e dependência (CFS). Para a abordagem *wrapper*, que utiliza classificadores para avaliar os subconjuntos, foram escolhidos: J48 (implementação do algoritmo C4.5 no *software Weka*), *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM) e *Naive Bayes*. A metodologia aplicada para esta etapa foi definida por Borges (2006) com a aplicação de duas formas de busca. Neste trabalho, a busca empregada apenas a busca sequencial. Com a aplicação da seleção, cada base de dados resulta em 9 subconjuntos de dados como mostrado na Figura 6.

Figura 6 - Execução da Seleção de Atributos

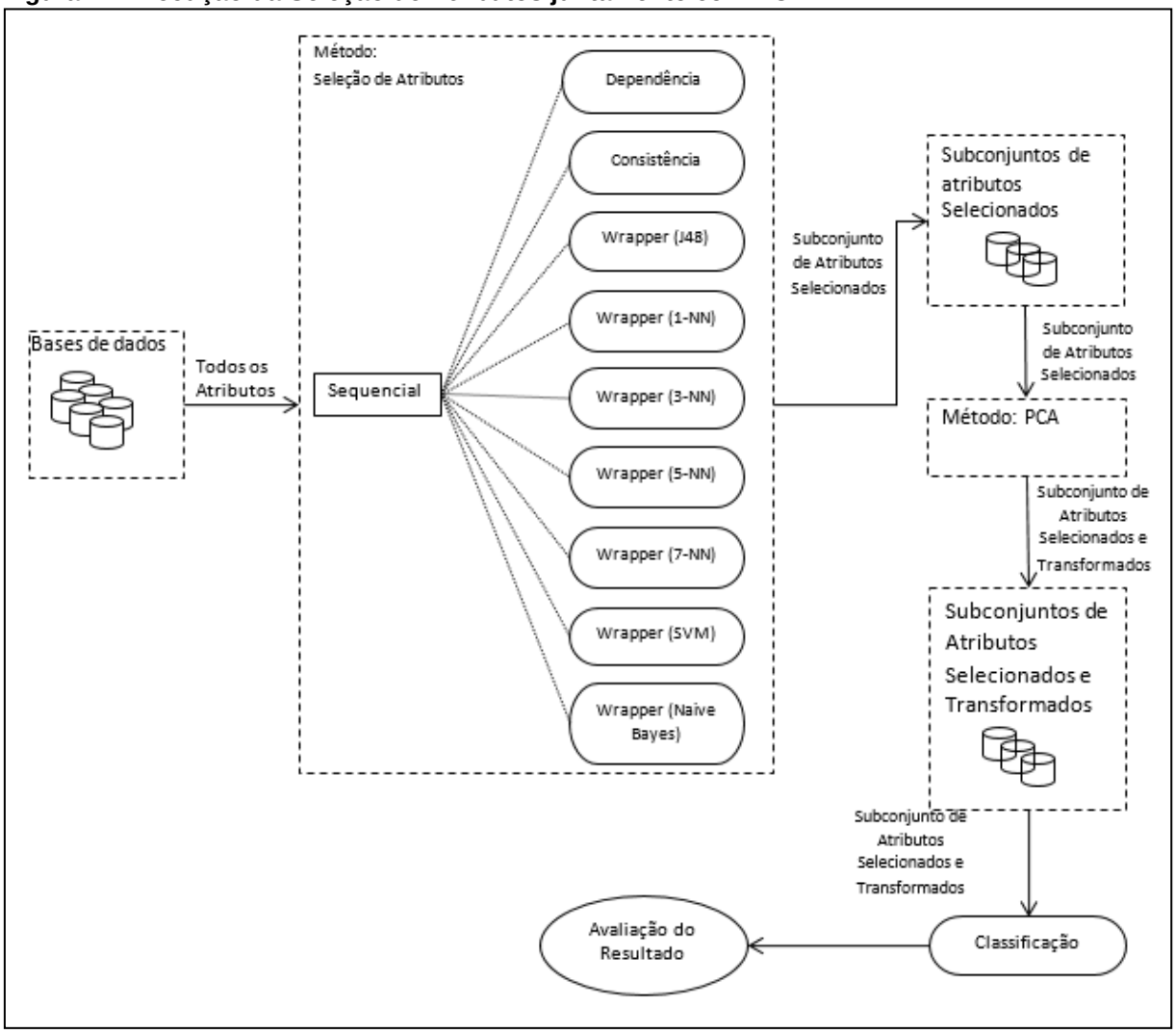


Fonte: Adaptado de Borges (2006)

4.4 EXECUÇÃO CONJUNTA DA SELEÇÃO DE ATRIBUTOS E ANÁLISE DE COMPONENTES PRINCIPAIS

Os subconjuntos gerados pela seleção de atributos foram submetidos ao algoritmo PCA. Dessa forma, cada subconjunto encontrado na seleção de atributos resulta em um novo subconjunto com dados transformados. Para adaptação para esta abordagem, o algoritmo de projeção aleatória encontrado em Borges (2006) foi substituída pela Análise de Componentes Principais. A utilização dos métodos de forma conjunta pode ser observada na Figura 7.

Figura 7 - Execução da Seleção de Atributos juntamente com PCA



Fonte: Adaptado de Borges (2006)

Além disso, o método PCA teve seus critérios de utilização definidos com base na porcentagem de variância dos dados de entrada, sendo 90, 95% e 99% os valores escolhidos.

Com a aplicação das três configurações da Extração de atributos nos subconjuntos resultantes da Seleção de atributos, cada base de dados resulta em 27 subconjuntos de dados.

4.5 CLASSIFICAÇÃO

Foram aplicados algoritmos de classificação nos subconjuntos gerados nas etapas anteriores. Os algoritmos utilizados foram o C4.5, KNN (com K=1, K=3, K=5 e K=7), *Naive Bayes* e SVM. Para uma classificação menos tendenciosa, foi utilizado o método de validação cruzada estratificada com 10 partições (*10 fold cross-validation*) que realiza a partição aleatória dos dados originais em 10 partes iguais.

Como se trata de modelos de predição, é necessário a análise da taxa de acerto do algoritmo. Esta taxa de acerto pode ser obtida a partir do quociente entre o número de instâncias corretamente classificadas e o número total de instancias, onde a taxa de acerto final é a média aritmética das taxas de acerto obtidas para cada partição.

4.6 CONSIDERAÇÕES DO CAPÍTULO

Neste Capítulo foram apresentadas a metodologia geral dos testes aplicados e as bases de dados utilizadas. Além disso foi demonstrado como as abordagens técnicas de Seleção e de Extração de atributos foram aplicadas nas bases de dados escolhidas. Foram apresentados também os classificadores utilizados na realização dos experimentos.

5 EXPERIMENTOS E RESULTADOS

Este Capítulo apresenta os resultados dos experimentos realizados. A Seção 5.1 apresenta os resultados obtidos das bases de dados com todos os atributos. A Seção 5.2 mostra os resultados quando aplicado os algoritmos de seleção de atributos e na seção 5.3 mostra os resultados com a aplicação da extração de atributos por meio do algoritmo de Análise de Componentes Principais. Por fim, a Seção 5.3 aborda as considerações finais do capítulo.

5.1 TODOS OS ATRIBUTOS

Para realizar comparações com os métodos de redução de dimensionalidade, foi realizada a classificação nas bases de dados com todos os atributos. Foram utilizados os classificadores: J48, KNN (com *k*=1, *k*=3, *k*=5 e *k*=7), Naive Bayes e SVM. Os resultados foram avaliados estatisticamente por meio do teste-t emparelhado (MONTGOMERY; RUNGER, 2018), um procedimento estatístico utilizado para determinar se a diferença média entre duas amostras é zero. Neste trabalho o teste foi aplicado utilizando o classificador Naive Bayes como base, onde resultados estatisticamente piores que os do algoritmo base são sinalizados com '*' enquanto os resultados em negrito indicam que são significativamente superior ao algoritmo base.

A Tabela 10 mostra os resultados obtidos a partir da aplicação da classificação em cada base de dados.

Tabela 10 - Classificação com todos os atributos

Bases de dados	Naive Bayes	SVM	1-NN	3-NN	5-NN	7-NN	J48
AML-ALL	99,18±3,27	65,36±6,38*	87,63±12,90*	85,64±11,14*	84,16±11,06*	81,79±11,23*	81,43±10,98*
DLBCL	96,00±9,27	91,25±11,38	75,15±18,75*	75,35±16,44*	72,30±16,72*	75,00±18,03*	77,90±18,99*
DLBCL-NIH	59,00±10,32	63,29±7,48	52,42±10,31	49,54±9,42	51,46±8,97	49,83±9,44*	56,00±10,62
DLBCL-Outcome	41,10±18,37	**55,33±6,90**	49,13±21,77	49,53±20,71	49,67±20,38	55,60±21,77	51,67±17,29
DLBCL-Tumor	80,50±12,37	75,36±3,75	82,41±12,42	88,98±11,27	87,25±12,03	89,30±10,14	80,23±14,28

Fonte: Autoria própria

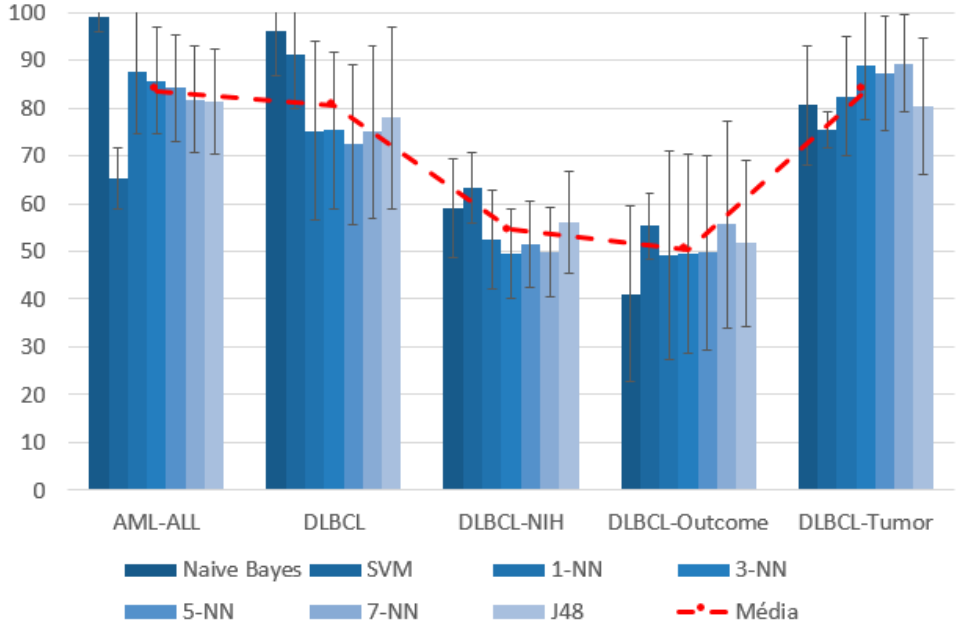
Nota-se que as bases DLBCL-NIH e DLBCL-Outcome obtiveram as menores taxas de classificação, com valores menores que 60%. É possível observar que dentro destas duas bases, a menor classificação foi obtida pelo classificador Naive Bayes na base DLBCL-Outcome, com cerca de 41% de acerto, enquanto a maior foi obtida pelo classificador SVM na base DLBCL-NIH com cerca de 63%.

Na base AML-ALL constata-se que o algoritmo de Naive Bayes obteve resultados estatisticamente superiores aos demais classificadores. Além disso o classificador Naive Bayes aplicado na base AML-ALL foi o classificador com maior taxa de classificação, sendo estatisticamente similar à taxa de classificação obtida com a base DLBCL também com o classificador Naive Bayes além do SVM.

A base DLBCL-Outcome foi a única base que apresentou um classificador estatisticamente superior ao classificador base, com uma taxa de classificação de cerca de 56%.

Na Gráfico 1 é possível observar que a maior média de classificação na base de dados AML-ALL. Tendo uma baixa taxa de classificação média as bases DLBCL-NIH e DLBCL-Outcome.

Gráfico 1 - Média das taxas de acerto todos os atributos, nas bases de dados analisadas



Fonte: Autoria própria

5.2 SELEÇÃO DE ATRIBUTOS

Para a seleção de atributos, como citado anteriormente, foram utilizadas as abordagens Filtro e *Wrapper*. A Tabela 11 mostra a diferença no número de atributos selecionados com cada algoritmo com o método de busca sequencial.

É possível observar que houve uma redução significativa dos atributos por meio da seleção de atributos. Isso se deve em parte pela natureza da busca sequencial, que ao realizar a inserção de um atributo por vez tende a selecionar poucos atributos.

Dentro da abordagem Filtro é possível observar uma significante diferença entre as medidas de avaliação utilizadas. A medida de avaliação CFS foi a que selecionou o maior número de atributos, com uma média de 67 atributos enquanto a medida CSE selecionou em média de 6 atributos.

Na abordagem *Wrapper* nota-se uma boa uniformidade entre as medidas de avaliação, selecionando em média de 4 a 6 atributos, sendo a base DLBCL-NIH a exceção com uma média de 13 atributos escolhidos.

As menores reduções foram realizadas pela abordagem Filtro com medida de avaliação CFS, com redução de 98,8% dos atributos na base AML-ALL; 98,01% na base DLBCL; 99,3% na base DLBCL-NIH; 99,4% na base DLBCL-Outcome e 98,6% para a base DLBCL-Tumor.

Tabela 11 - Quantidade de atributos selecionados por meio da Seleção de Atributos nas bases de dados analisadas

			Bases de dados				
		Medida de Avaliação	**AML-ALL**	**DLBCL**	**DLBCL-NIH**	**DLBCL-Outcome**	**DLBCL-Tumor**
	Total	-	7129	4026	7339	7129	7129
	Filtro	CFS	81	80	45	36	93
		CSE	3	3	14	6	4
Sequencial	**Wrapper**	J48	1	1	13	4	2
		1-NN	4	3	26	7	7
		3-NN	7	5	11	8	4
		5-NN	3	5	2	8	5
		7-NN	3	4	15	7	5
		Naive Bayes	4	4	13	5	5
		SVM	1	3	9	1	1

Fonte: Autoria Própria

Nota-se, também, alguns casos em que apenas um atributo foi selecionado, como observado com a abordagem *Wrapper* com o classificador J48 como medida de avaliação nas bases AML-ALL, DLBCL além do classificador SVM nas bases AML-ALL, DLBCL-Outcome e DLBCL-Tumor.

5.2.1 Abordagem Filtro

Para a aplicação da abordagem Filtro foram utilizadas as medidas de consistência (CSE) e dependência (CFS). A Tabela 12 mostra os resultados da classificação dos dados selecionados utilizando as medidas de CFS e CSE para a base de dados AML-ALL. É possível observar que a medida CFS resultou em taxas de classificação iguais ou superiores à medida CSE em quase todos os classificadores, exceto com o classificador J48.

Observa-se que os menores resultados foram obtidos utilizando o classificador SVM onde ambas as medidas obtiveram uma taxa de acerto abaixo de 70%.

Tabela 12 - Resultado em % dos classificadores – Abordagem Filtro, base AML-ALL

Medida de avaliação	Naive Bayes	SVM	1-NN	3-NN	5-NN	7-NN	J48
CFS	99,87±1,25	65,36±6,38*	98,45±4,89	99,46±3,33	98,36±4,92	97,93±5,36	84,38±10,98*
CSE	94,73±8,90	65,36±6,38*	94,16±8,70	95,04±7,95	95,16±7,93	94,04±8,16	93,04±10,39

Fonte: Autoria Própria

Na Tabela 13 são apresentados os resultados da base DLBCL. Destaca-se a taxa de classificação estatisticamente abaixo obtida para o classificador J48 em relação ao classificador base de Naive Bayes, que é a menor taxa de classificação obtida nos testes aplicados nesta base.

É possível notar o melhor desempenho da medida CFS na maioria dos classificadores, exceto no J48 como comentado anteriormente. Além de uma taxa de 100% de acerto utilizando Naive Bayes e SVM. Apesar disso, os resultados são estatisticamente semelhantes ao algoritmo base e entre as duas diferentes abordagens escolhidas.

Tabela 13 - Resultado em % dos classificadores – Abordagem Filtro, base DLBCL

Medida de avaliação	Naive Bayes	SVM	1-NN	3-NN	5-NN	7-NN	J48
CFS	100,00±0,00	100,00±0,00	97,90±6,36	99,55±3,19	99,00±4,38	99,80±2,00	84,55±17,91*
CSE	90,65±11,63	91,65±12,21	93,65±11,35	95,30±8,96	93,40±11,12	93,20±11,18	87,80±14,24

Fonte: Autoria própria

Na Tabela 14 apresenta-se os resultados da classificação dos dados selecionados utilizando as medidas CFS e CSE para a base de dados DLBCL-NIH. Nota-se uma baixa taxa de classificação em ambas as medidas com os classificadores KNN, em que todos obtiveram resultados estatisticamente inferior ao classificador base. O classificador SVM foi o único que apresentou resultados estatisticamente equivalentes ao classificador base. Além disso, a abordagem CFS apresenta taxas de classificação iguais ou superiores do que a medida CSE para todos os classificadores utilizados.

Tabela 14 - Resultado em % dos classificadores – Abordagem Filtro, base DLBCL-NIH

Medida de avaliação	Naive Bayes	SVM	1-NN	3-NN	5-NN	7-NN	J48
CFS	72,75±8,87	71,63±8,26	57,13±8,25*	56,38±7,82*	58,37±8,36*	59,37±8,24*	66,67±9,79
CSE	66,12±9,30	64,79±9,37	55,17±10,89*	56,04±9,79*	56,75±8,96*	56,75±9,62*	65,04±8,58

Fonte: Autoria própria

Na Tabela 15 verifica-se que, para ambas as medidas de avaliação, o classificador SVM obteve resultados estatisticamente inferiores a seus classificadores base, e os menores resultados obtidos.

É possível notar também que o a maior taxa de classificação da medida CFS foi obtida utilizando o classificador Naive Bayes, com 84,3% de acerto, enquanto para a medida CSE foi obtida pelo classificador 3-NN, com 75.4% de acerto.

Tabela 15 - Resultado em % dos classificadores – Abordagem Filtro, base DLBCL-Outcome

Medida de avaliação	Naive Bayes	SVM	1-NN	3-NN	5-NN	7-NN	J48
CFS	84,33±16,79	55,33±6,90*	72,20±17,36	76,77±15,56	77,27±15,57	76,53±14,94	64,20±17,05*
CSE	72,30±17,39	55,33±6,90*	74,67±16,22	75,40±17,41	67,83±16,86	69,00±15,99	72,23±17,34

Fonte: Autoria própria

Na Tabela 16, utilizando a base DLBCL-Tumor, é possível notar a alta taxa de classificação do classificador 7-NN, apesar de não ser um resultado estatisticamente superior ao classificador base.

Verifica-se que para ambas as medidas, o único classificador estatisticamente inferior ao classificador base é o SVM, com uma taxa inferior a 80%. A medida de classificação CFS obteve resultados estatisticamente iguais ou superiores a medida CSE com todos os classificadores.

Tabela 16 - Resultado em % dos classificadores – Abordagem Filtro, base DLBCL-Tumor

Medida de avaliação	Naive Bayes	SVM	1-NN	3-NN	5-NN	7-NN	J48
CFS	96,16±7,12	75,36±3,75*	96,11±7,41	96,77±6,42	98,46±4,55	100,00±0,00	88,86±11,76
CSE	93,25±7,74	75,36±3,75*	92,80±9,24	92,20±9,49	90,09±9,86	89,88±10,07	90,46±10,45

Fonte: Autoria própria

Identifica-se que, no geral, as maiores taxas de classificação foram obtidas pela medida de CFS. Isto pode ocorrer pela forma que a medida escolhe os atributos, baseados na dependência da classe com o atributo. Com isso, pode-se concluir que a medida de dependência apresentou indicadores de desempenho maiores do que a medida de consistência, porém com diferenças minimamente significativas.

5.2.2 Abordagem Wrapper

Para a abordagem *Wrapper* foram utilizadas como medida de avaliação os algorítmos utilizados na classificação. Para facilitar a discussão, usar-se-á os identificadores das abordagens utilizadas com o prefixo ‘Wp+’ simbolizando a utilização da abordagem *Wrapper*, seguido pelo nome do classificador utilizado como medida de avaliação.

Na Tabela 17 são apresentados os resultados da classificação, utilizando seus respectivos classificadores, dos subconjuntos obtidos com a abordagem *wrapper* a partir das bases de dados originais.

É possível observar que todos os classificadores obtiveram resultados estatisticamente semelhante em cada base aplicada.

Tabela 17 – Resultado em % dos classificadores – Seleção *Wrapper*

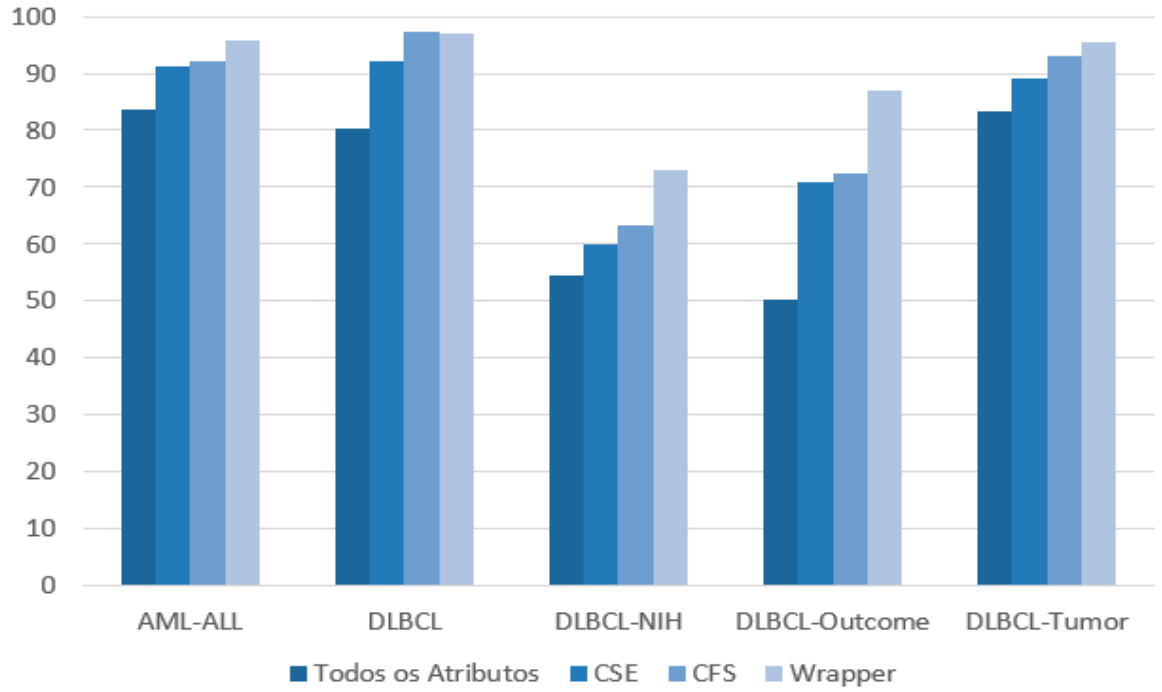
Base de Dados	Wp+*Naive Bayes*	Wp+SVM	Wp+1-NN	Wp+3-NN	Wp+5-NN	Wp+7-NN	Wp+J48
AML-ALL	100,00±0,00	78,93±12,29	99,57±3,18	99,30±4,19	98,95±3,60	100,00±0,00	94,64±9,11
DLBCL	97,40±7,12	97,95±6,20	97,55±7,05	97,50±8,42	97,85±6,52	98,70±5,20	91,20±11,79
DLBCL-NIH	75,00±9,14	74,83±7,65	76,67±8,48	73,33±8,82	60,29±10,99	74,75±7,56	75,50±8,28
DLBCL-Outcome	88,37±13,49	70,60±16,93	87,10±14,42	90,17±12,57	94,73±9,23	87,57±13,67	90,40±12,33
DLBCL-Tumor	99,75±1,76	82,07±8,15	98,73±4,91	97,91±4,82	99,25±2,98	98,46±4,88	92,82±9,17

Fonte: Autoria própria

Destaca-se as abordagens Wp+NaiveBayes e Wp+7-NN que obtiveram em média, entre as bases, as maiores taxas de classificação com valores em torno de 92%. Nas abordagens *Wrapper* foi observado que a aplicação do Wp+SVM obteve uma média entre as bases de 80,7%.

No Gráfico 2 é apresentado, de acordo com os dados descritos em ambas as abordagens, uma comparação entre as abordagens Filtro e *Wrapper* para as bases de dados analisadas.

Gráfico 2 - Média das taxas de acerto para as abordagens Filtro e *Wrapper*, nas bases de dados analisadas



Fonte: Autoria própria

Com a análise dos resultados das duas abordagens, é possível concluir que a abordagem filtro obteve uma média de 81,9% enquanto a abordagem *Wrapper* obteve uma média de 89,6. Portanto, a abordagem *Wrapper* se mostra superior quando comparada a abordagem Filtro para as bases de dados analisadas.

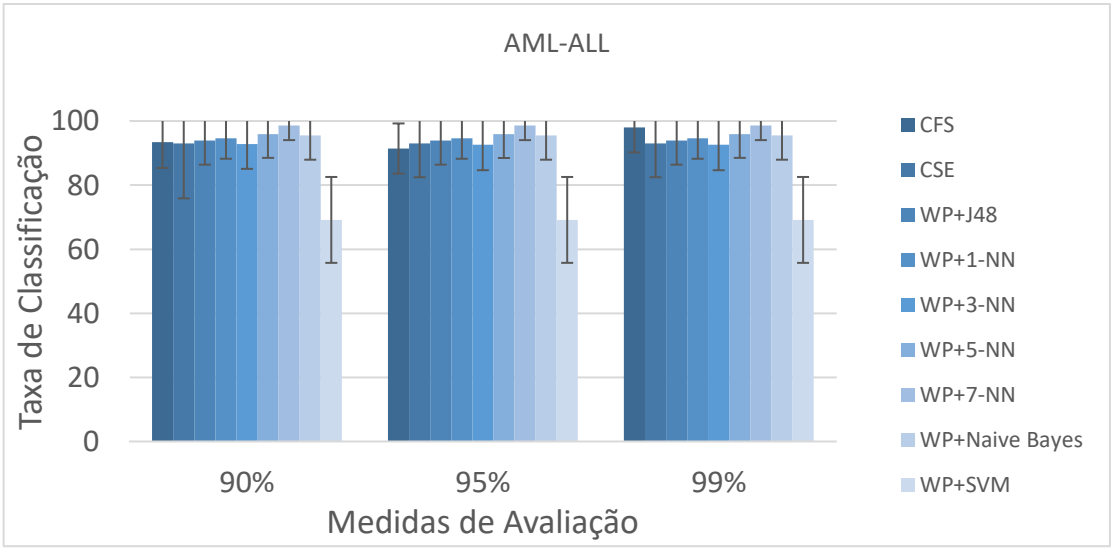
5.3 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

Nesta Seção, são apresentados os resultados referentes a aplicação da Análise de Componentes Principais nos subconjuntos gerados pelas diferentes abordagens da Seleção de atributos. Além de retratar qual combinação entre abordagens de seleção e o método PCA apresentou melhor desempenho preditivo.

O critério de aplicação do PCA foi definido de acordo com o número de atributos finais, sendo foi estabelecido em 90%, 95% e 99%. A fim de facilitar o entendimento e a comparação entre as técnicas de redução de atributos, foi utilizada a taxa média de classificação obtida para cada abordagem.

No Gráfico 3 são apresentadas as taxas médias de classificação com subconjuntos da base de dados AML-ALL.

Gráfico 3 – Média (em %) das taxas de acerto para a Seleção + Extração de atributos, na base AML-ALL



Fonte: Autoria própria

Com os resultados obtidos é possível observar que as melhores taxas médias de classificação com a aplicação da análise de componentes principais juntamente

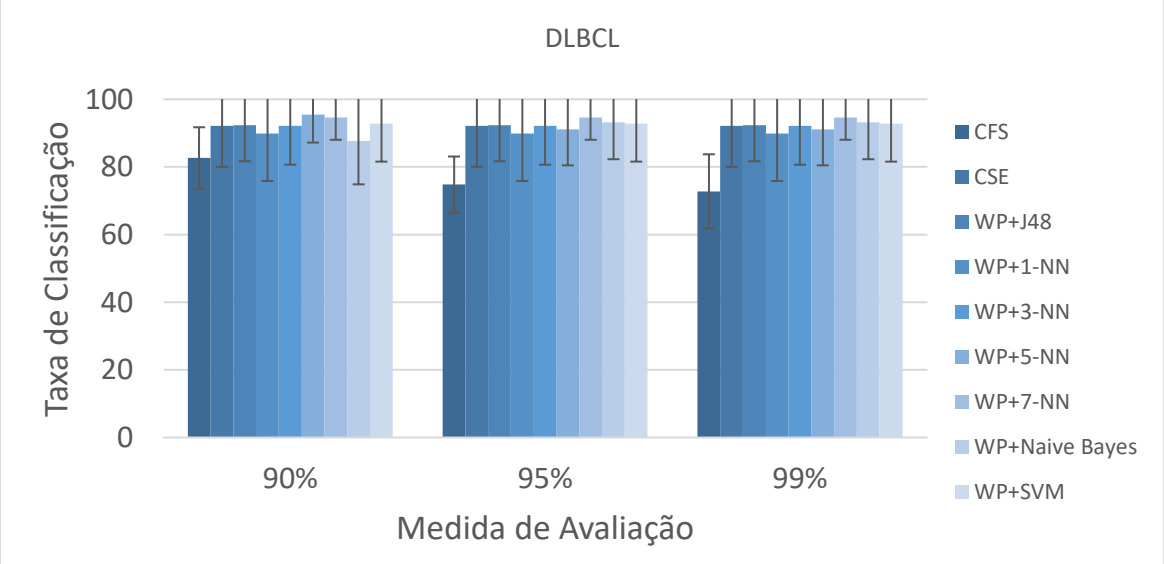
com a seleção de atributos *Wrapper* com o classificador 7-NN como medida de avaliação do subconjunto.

Nota-se que, em média, os resultados obtidos com a utilização de diferentes variâncias no PCA não possuem diferença significativa.

Em comparação com a classificação de todos os atributos, que obteve média de classificação de 81,73%, a média das abordagens de redução de dimensionalidade utilizadas resultaram aumento na taxa de acerto dos classificadores. Apenas o subconjunto obtido com da abordagem Wp+SVM+PCA apresentou uma média inferior.

Nos Gráficos 4, 5, 6 e 7 são apresentadas as taxas de classificação média após a aplicação do PCA.

Gráfico 4 - Média (em %) das taxas de acerto para a Seleção + Extração de atributos, na base DLBCL

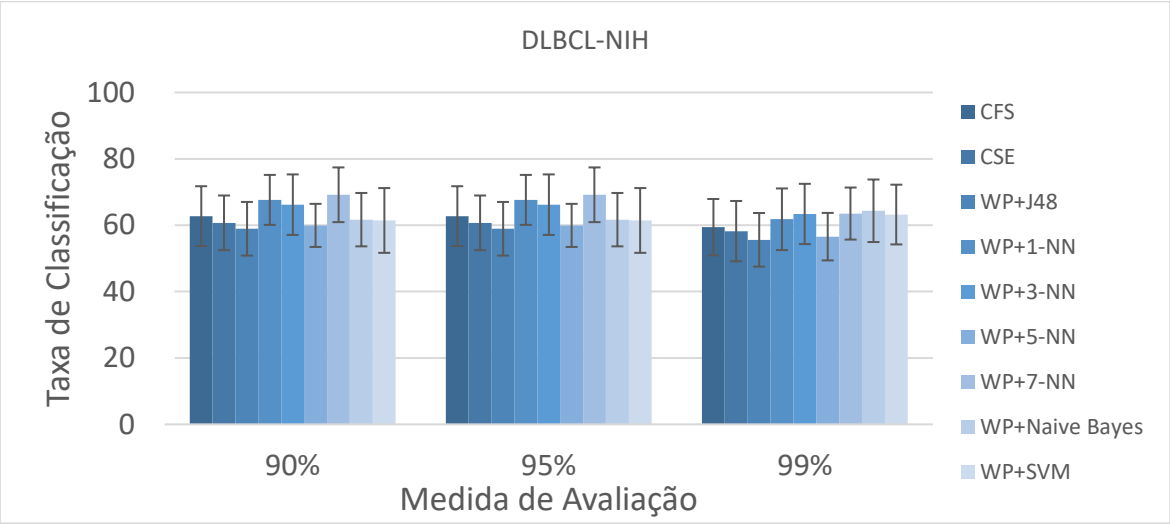


Fonte: Autoria própria

No Gráfico 4 é possível observar que mesmo a abordagem com a média mais baixa de classificação supera a taxa de classificação média de todos os atributos, que é um valor de aproximadamente 81% para a base DLBCL.

A maior média pode ser observada com a aplicação conjunta da seleção *Wrapper* com 5-NN como medida de avaliação e o PCA com 90% dos atributos, obtendo um valor de aproximadamente 95%. Enquanto isso a menor média pode ser observada na combinação da abordagem Filtro utilizando as medidas CFS e PCA utilizando 99% dos atributos.

Gráfico 5 - Média (em %) das taxas de acerto para a Seleção + Extração de atributos, na base DLBCL-NIH

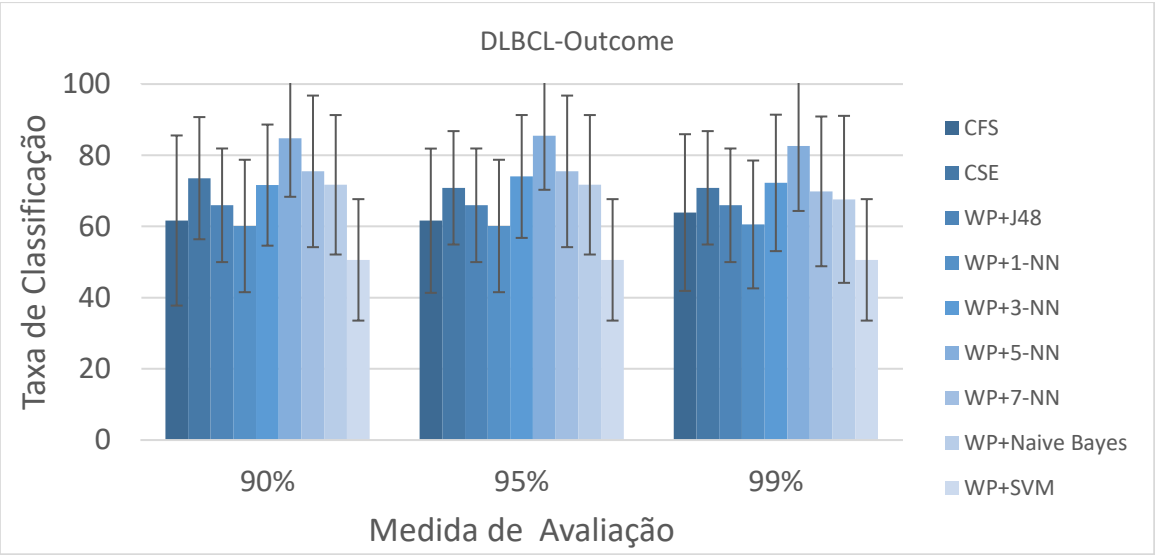


Fonte: Autoria própria

No Gráfico 5, apesar dos resultados relativamente baixos comparados com abordagens anteriores, tem-se melhora significante para com a média de classificação com todos os atributos. A classificação média de 50% sobe, no pior caso, para cerca de 55%.

Tem-se também, no pior caso, uma classificação média de 55,6% com a combinação da abordagem *Wrapper* com J48 como medida de avaliação e PCA com 99% dos atributos. No melhor caso, a classificação média fica em torno de 69% utilizando a abordagem *Wrapper* com 7-NN e PCA 90% e 99%.

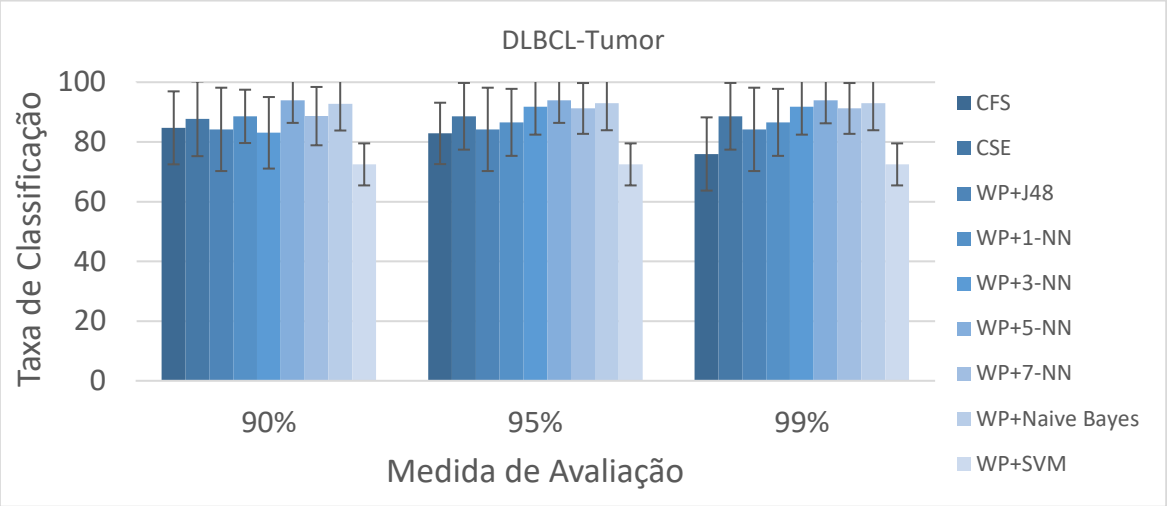
Gráfico 6 - Média (em %) das taxas de acerto para a Seleção + Extração de atributos, na base DLBCL-Outcome



Fonte: Autoria própria

Para a base de dados DBLCL-Outcome no gráfico 6 é possível observar que todos as abordagens obtiveram resultado em média melhor do que quando são utilizados todos os atributos para classificação. Os resultados com a aplicação de diferentes configurações do PCA se mostram equilibrados, onde em um mesmo subconjunto a maior diferença entre o melhor e o pior resultado é de cerca de 5,5%.

Gráfico 7 - Média (em %) das taxas de acerto para a Seleção + Extração de atributos, na base DLBCL-Tumor



Fonte: Autoria própria

Pode-se notar que, quando comparada com a classificação de todos os atributos, as abordagens CFS+PCA99% e as três aplicações do PCA sob o subconjunto Wp+SVM tiveram uma significante diminuição na taxa média de classificação.

Com os resultados apresentados é possível observar que as combinações de abordagens com melhor desempenho são *Wrapper*(7-NN) com o PCA com 95% e 99% dos atributos que apresentou uma média de 93.5%. Além da combinação entre *Wrapper*(5-NN) com o PCA com 90% dos atributos, que apresentou uma média de 95,14% na taxa de acerto.

Devido ao volume de resultados gerados, esses são apresentados no Apêndice A.

5.4 CONSIDERAÇÕES DO CAPÍTULO

Neste Capítulo foram apresentados os resultados da abordagem proposta. Foi comparada a eficácia da redução de dimensionalidade de algumas combinações de seleção e extração de Atributos com os resultados obtidos com todos os atributos.

As combinações de Seleção com Extração de atributos analisadas apresentaram resultados superiores aos obtidos com todos os atributos, porém resultados semelhantes com experimentos utilizando apenas a seleção de atributos.

6 CONCLUSÃO

A seleção e extração de atributos são técnicas presentes no processo KDD utilizadas para reduzir o número de atributos de um problema. Tem grande importância para o processo por remover atributos redundantes ou irrelevantes e transformar uma base com muitos atributos em uma com menos atributos sem perder as características dos dados.

As análises realizadas neste trabalho utilizam as abordagens Filtro e *Wrapper* para a seleção de atributos e o algoritmo PCA para a extração de atributos. Especialmente a análise de componentes principais, escolhidos pela seleção de atributos, possui como limitação o baixo número de atributos de entrada resultantes da etapa de seleção.

Para os experimentos foram utilizadas 5 bases de dados de microarranjos de DNA, em que o número de atributos ultrapassa a casa dos milhares. A execução do experimento se deu em três etapas, na primeira foram consideradas as bases de dados completa (com todos os atributos), na segunda foram aplicadas técnicas de diferentes abordagens de seleção de atributos. Enquanto na terceira etapa, foram aplicados sobre os subconjuntos gerados na segunda etapa a técnica PCA, mantendo 90%, 95% e 99% da quantidade de dados de entrada.

Após a análise dos resultados de classificação dos subconjuntos, foi possível encontrar um aumento significativo na taxa de classificação. Sendo no pior caso aproximadamente 38% de acerto para a base completa e cerca de 50% no pior caso para dados selecionados e extraídos.

6.1 TRABALHOS FUTUROS

Diversos trabalhos podem ser realizados. Uma das propostas é a aplicação da abordagem proposta em outras bases de dados e domínios de problemas com o objetivo de comparar o desempenho das combinações.

Outro ponto a ser investigado é a combinação das abordagens de seleção apresentadas em conjunto com outras formas de extração de atributos.

Pode-se também ser realizada a comparação entre os resultados do trabalho apresentado com outros resultados encontrados na literatura.

REFERÊNCIAS

AHUJA, J.; RATNOO, S. Dimension reduction for microarray data using multi-objective ant colony optimization. International Journal of Computational Systems Engineering, v.3, n. 1-2, p. 58-73, 2017.

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

AROWOLO M.O.; et al. A Comparative Analysis of Feature Extraction Methods for Classifying Colon Cancer Microarray Data. **EAI Endorsed Transactions**, 2017.

ARUNKUMAR C; RAMAKRISHNAN S. Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data. **Future Computing and Informatics Journal**, v. 3, p. 131-142, 2018.

ASH, A. A.; et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. **Nature**, n. 403, p. 503-511, 2000.

AWAD M; KHANNA R. Support vector machines for classification. **Efficient Learning Machines**. Berkeley (CA), p. 39-66, 2015.

AZIZ R; VERMA, C.K; SRIVASTAVA, N. (2016). A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data. **Genomics Data**, v. 8, p. 4-15, 2016.

BOLÓN-CANEDO, V. et al. A review of microarray datasets and applied feature selection methods. **Information Sciences**, v. 282, p. 111-135, 2014.

BÓLON-CANEDO, V; MAROÑO, N. S. Feature selection for high-dimensional data. 1. ed. Switzerland: **Springer International Publishing**, 2015.

BORGES, H. B. **Redução de dimensionalidade em bases de dados de expressão gênica**. 2006. 123 f. Dissertação (Mestrado) – Programa de Pós Graduação em Informática, Pontifica Universidade Católica do Paraná. Curitiba, 2006.

BORGES, H. B; NIEVOLA, J. C. **Comparing the dimensionality reduction methods in gene expression databases.** Expert Systems with Applications, v. 39, n. 12, p. 10780-10795, 2012.

CHANDRASHEKAR G; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, v. 40, p. 16-28, 2014.

CHEN, J.; et al. Feature selection for text classification with Naïve Bayes. **Expert Systems with Applications**, v. 36, p. 5432-5435, 2009.

CHIZI B; MAIMON O. **Dimension reduction and feature selection**. 2. ed. Boston: Springer, 2010.

CHUANG, L.; et al. A hybrid BPSO-CGA approach for gene selection and classification of microarray data. **Journal of Computational Biology**, v. 19, n. 1, p. 68-82, 2012.

CLARKE, B; FOKOUE, E; ZHANG H. H. **Principles and theory for data mining and machine learning**. 1. ed. New York: Springer, 2009.

CUNNINGHAM, J. P; GHANRAMANI, Z. Linear dimensionality reduction: survey, insights, and generalizations. **Journal of Machine Learning Research**, v. 16, p. 2859-2900, 2015.

DASH, R. A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: a case study. **Journal of King Saud University - Computer and Information Sciences**, 2017.

DASH, M; LIU, H. Consistency-based search in feature selection. **Artificial Intelligence,** v. 151, p.155-176, 2003.

DING, S; et al. A survey on feature extraction for pattern recognition. **Artificial Intelligence Review**, v. 37, p. 169-180, 2012.

EBRAHIMPOUR, M. K; et al. Occam's razor in dimension reduction: using reduced row Echelon form for finding linear independent features in high dimensional microarray datasets. **Engineering Applications of Artificial Intelligence**, v. 62, p. 214-221, 2017.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n.3, p. 37, 1996.

FREITAS, A. Data mining and knowledge discovery with evolutionary algorithms. **Springer Science & Business Media**, p. 265, 2002.

FRIEDMAN, J; HASTIE, T; TIBSHIRANI, R. **The elements of statistical learning**. 2 ed. New York: Springer series in statistics, 2001.

GOLDSCHIMIDT, R; PASSOS, E. **Data mining**: um guia prático. 2. ed. Rio de Janeiro: Campus, 2005.

GOLDSCHMIDT, R; PASSOS, E. **Data mining**: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Campus, 2005.

GOLUB T. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. **Science**, v. 286, p. 531-537, 1999.

GUO, S; et al. A L1-regularized feature selection method for local dimension reduction on microarray data. **Computational Biology and Chemestry**, v. 67, p. 92-101, 2017.

GUYON, I; et al. **Feature extraction**: foundations and applications. v. 207, Springer, 2006.

HAN J; KAMBER, M; PEI J. **Data mining concepts and techniques.** 30 ed. Morgan Kaufmann, 2012.

HAND, D. J; SMYTH P; MANNILA H. **Principles of data mining**. Cambridge: The MIT Press, 2001.

HE, S; et al. Robust twin boosting for feature selection from high-dimensional omics data with label noise. **Information Sciences**, v. 291. p. 1–18, 2015.

ISLAM, M. J; et al. Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. **Journal of Convergence Information Technology**, p. 1541-1546, 2007.

JIANG Z; XU R. A novel feature extraction approach for microarray data based on multi-algorithm fusion. **Bioinformation**, v. 11, p. 27-33, 2015.

KANTARDZIC, M. **Data mining:** concepts, models, and algorithms. 2. ed. Wiley-IEEE Xplore, 2011.

KASTRIN, A; PETERLIN, B. Rasch-based high-dimensionality data reduction and class prediction with Applications to microarray gene expression data. **Expert Systems with Applications**, v. 37, p. 5178-5185, 2010.

KEERTHI, S; GILBERT, E. Convergence of a generalized SMO algorithm for SVM classifier design. **Machine Learning**, v. 46, p. 351-360, 2002.

KOHAVI R; JOHN, G. H. The *wrapper* approach. **Feature extraction, construction and selection**, v. 453, p. 33–49, 1998.

KUMAR, M; RATH N. K; RATH S. K. Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier. **Journal of Biomedical Informatics**, v. 60, p. 395-409, 2016.

LATKOWSKI T; OSWSKI S. Gene selection in autism - comparative study. **Neurocomput**, v. 250, p. 37-44, 2017.

LEE, J. A; VERLEYSEN, M. **Nonlinear dimensionality reduction**. New York: Springer, 2007.

LIANG, S.; et al. A review of matched-pairs feature selection methods for gene expression data analysis. **Computational and structural biotechnology journal**, v. 16, p. 88-97, 2018.

LIU, H; MOTODA, H. **Feature selection for knowledge discovery and data mining**. 1. ed. Norwell: Kluwer Academic Publishers, 1998.

LIU, H; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on knowledge and data engineering**, v. 17, n. 4, p. 491-502, 2005.

LLERENA, S. E. **Redução dimensional de dados de alta dimensão e poucas amostras usando Projection Pursuit**. 2013. Tese - Doutorado em Sistemas Dinâmicos, Escola de Engenharia de São Carlos. São Carlos, 2013.

LU H; et al. A Hybrid Feature selection algorithm for gene expression data classification. **Neurocomputing**, v. 256, p. 56-62, 2017.

MAULIK, U; CHAKRABORTY, D. Fuzzy preference based feature selection and semi supervised SVM for cancer classification. **IEEE transactions on nanobioscience**, v. 13, n. 2, p. 152-160, 2014.

MOLINA, L. C; MUÑOZ, L. B; NEBOT, A. Feature selection algorithms: a survey and experimental evaluation. **IEEE International Conference on Data Mining**, 2002.

MOLLAEE, M; MOATTAR M. H. A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. **Biocybernectics and Biomedical Engineering**, v. 36, p. 521-529, 2016.

MONTGOMERY, D. C; RUNGER, G. C. Estatística aplicada e probabilidade para engenheiros. 6. ed. Rio de Janeiro: **LTC**, 2018.

NANNI, L; BRAHNAM, S.; LUMINI A. Combining multiple approaches for gene microarray classification. **Bioinformatics**, v. 28, p. 1151–1157, 2012.

NURFALAH, A; ADIWIJAYA, K; ARDIYANTI, A. Cancer detection based on microarray data classification using PCA and modified back propagation. **Far East Journal of Electronics and Communications**, v. 16, p. 269-281, 2016.

PEREZ-DIEZ A; MORGUN A; SHULZHENKO N. Microarrays for cancer diagnoses and classification. **Advances in Experimental Medicine and Biology**, v. 593, p. 74-85, 2007.

ROSENWALD, A.; et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. **The New England Journal of Medicine**, n. 346, p. 1937-1947, 2002.

SALZBERG, S. L. Book Review: C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. **Machine Learning**, n. 16, p. 235-240, 1994.

SEIJO-PARDO, B; BOLÓN-CANEDO, V; ALONSO-BETANZOS, A. Testing different ensemble configurations for feature selection. **Neural Processing Letters**, v. 46, p. 857-880, 2017.

SHIPP, M. A.; et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. **Nature Medicine**, n. 8, p. 68-74, 2002.

SREEVANI; MURTHY C. A.; CHANDA, B. Generation of compound features based on feature interaction for classification. **Expert Systems with Applications**, v. 108, p. 61-73, out. 2018.

TAN, P; STEINBACH, M; KUMAR, V. **Classification**: basic concepts, decision trees, and model evaluation. Introduction to data mining, v. 1, p. 145-205, 2006.

TAN, S. An effective refinement strategy for KNN text classifier. **Expert Systems with Applications**, v. 30, p. 290-298, 2006.

WANG, Y.; et al. Gene selection from microarray data for cancer classification—a machine learning approach. **Computational biology and chemistry**, v. 29, n. 1, p. 37-46, 2005.

WAZLAWICK, R. Metodologia de pesquisa para ciência da computação. 2. ed. Elsevier Brasil, 2017.

WITTEN, I. H.; FRANK, E; MARK, A. H. **Data mining**: Practical machine learning tools and techniques. 3. ed. Morgan Kaufmann, 2011.

XING, E. P.; JORDAN, M. I.; KARP Richard M. Feature selection for high-dimensional genomic microarray data. **18th International Conference on Machine Learning**, p. 601-608, 2001.

YOU, W; et al. TotalPLS: local dimension reduction for multicategory microarray data. **IEEE Transactions on Human-Machine Systems**, v. 44, p. 125-138, 2014.

APÊNDICE A - Resultados da classificação de subconjuntos obtidos com Seleção e Extração de Atributos

São apresentados os resultados da classificação de cada base de dados após a aplicação de técnicas de Seleção se atributos seguida da Extração de atributos por meio da análise de componentes principais.

Nas Tabelas 18, 19 e 20 são apresentados as taxas de classificação dos subconjuntos com 90%, 95% e 99% dos atributos da base AML-ALL.

Tabela 18 - Resultado em % dos classificadores - Base AML-ALL - PCA90%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS+PCA	100,00±0,00	98,75±3,95	91,43±9,99	91,43±12,05	87,50±13,81	91,79±9,23	93,04±7,36
CSE+PCA	90,54±11,11	94,29±9,99	91,61±11,97	93,21±11,47	93,04±9,98	93,04±9,98	95,71±9,64
Wp+J48+PCA	93,04±7,36	94,64±9,11	91,61±7,24	94,46±7,17	94,46±7,17	94,46±7,17	94,46±7,17
Wp+1NN+PCA	94,29±7,38	90,36±6,69	100,00±0,00	95,89±6,63	94,46±7,17	91,61±9,89	95,71±6,90
Wp+3NN+PCA	80,54±11,92	88,75±11,23	98,57±4,52	97,14±6,02	97,32±5,66	93,04±7,36	94,29±7,38
Wp+5NN+PCA	94,29±7,38	94,29±9,99	93,04±9,98	97,14±6,02	98,57±4,52	95,71±9,64	98,57±4,52
Wp+7NN+PCA	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52
Wp+Naive Bayes+PCA	95,89±6,63	97,14±6,02	97,14±6,02	94,29±9,99	94,29±9,99	91,43±9,99	98,57±4,52
Wp+SVM+PCA	65,36±6,69	65,36±6,69	78,57±17,82	72,86±18,07	66,96±16,95	69,64±17,92	65,36±6,69

Fonte: Autoria própria

Tabela 19 - Resultado em % dos classificadores - Base AML-ALL - PCA95%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS+PCA	100,00±0,00	98,75±3,95	79,64±12,12*	90,54±8,83*	88,93±10,71	87,68±12,04	94,46±7,17
CSE+PCA	90,54±11,11	94,29±9,99	91,61±11,97	93,21±11,47	93,04±9,98	93,04±9,98	95,71±9,64
Wp+J48+PCA	93,04±7,36	94,64±9,11	91,61±7,24	94,46±7,17	94,46±7,17	94,46±7,17	94,46±7,17
Wp+1-NN+PCA	94,29±7,38	90,36±6,69	100,00±0,00	95,89±6,63	94,46±7,17	91,61±9,89	95,71±6,90
Wp+3-NN+PCA	77,86±13,15	88,75±11,23	98,57±4,52 v	**98,57±4,52**	**97,32±5,66**	**91,61±9,89**	**95,71±6,90**
Wp+5-NN+PCA	94,29±7,38	94,29±9,99	93,04±9,98	97,14±6,02	98,57±4,52	95,71±9,64	98,57±4,52
Wp+7-NN+PCA	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52
Wp+Naive Bayes+PCA	95,89±6,63	97,14±6,02	97,14±6,02	94,29±9,99	94,29±9,99	91,43±9,99	98,57±4,52
Wp+SVM+PCA	65,36±6,69	65,36±6,69	78,57±17,82	72,86±18,07	66,96±16,95	69,64±17,92	65,36±6,69

Fonte: Autoria própria

Tabela 20 - Resultado em % dos classificadores - Base AML-ALL - PCA99%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS+PCA	100,00±0,00	98,75±3,95	77,86±13,15*	73,57±10,02	69,64±10,24*	69,64±10,24*	98,75±3,95
CSE+PCA	90,54±11,11	94,29±9,99	91,61±11,97	93,21±11,47	93,04±9,98	93,04±9,98	95,71±9,64
Wp+J48+PCA	93,04±7,36	94,64±9,11	91,61±7,24	94,46±7,17	94,46±7,17	94,46±7,17	94,46±7,17
Wp+1-NN+PCA	94,29±7,38	90,36±6,69	100,00±0,00	95,89±6,63	94,46±7,17	91,61±9,89	95,71±6,90
Wp+3-NN+PCA	77,86±13,15	88,75±11,23	**98,57±4,52**	**98,57±4,52**	**97,32±5,66**	**91,61±9,89**	**95,71±6,90**
Wp+5-NN+PCA	94,29±7,38	94,29±9,99	93,04±9,98	97,14±6,02	98,57±4,52	95,71±9,64	98,57±4,52
Wp+7-NN+PCA	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52	98,57±4,52
Wp+Naive Bayes+PCA	95,89±6,63	97,14±6,02	97,14±6,02	94,29±9,99	94,29±9,99	91,43±9,99	98,57±4,52
Wp+SVM+PCA	65,36±6,69	65,36±6,69	78,57±17,82	72,86±18,07	66,96±16,95	69,64±17,92	65,36±6,69

Fonte: Autoria própria

Nas Tabelas 21, 22 e 23 são apresentados as taxas de classificação dos subconjuntos com 90%, 95% e 99% dos atributos da base DLBCL.

Tabela 21 - Resultado em % dos classificadores - Base DLBCL - PCA90%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	100,00±0,00	100,00±0,00	75,00±18,26*	68,50±13,13*	70,50±13,22*	66,50±12,70*	98,00±6,32
CSE	86,50±16,67	93,50±10,55	91,50±11,07	93,50±10,55	93,50±10,55	96,00±8,43	90,50±17,07
Wp+J48	93,50±10,55	93,50±10,55	87,50±10,87	93,50±10,55	93,50±10,55	93,50±10,55	91,50±11,07
Wp+1-NN	84,00±16,30	93,00±11,35	93,50±10,55	86,50±16,67	91,00±11,74	92,50±16,87	89,00±15,06
Wp+3-NN	86,50±15,10	97,50±7,91	91,00±11,74	97,50±7,91	93,00±11,35	93,00±11,35	86,50±15,10
Wp+5-NN	93,50±10,55	100,00±0,00	93,00±11,35	94,00±9,66	96,00±8,43	94,00±9,66	97,50±7,91
Wp+7-NN	89,00±16,63	92,50±12,08	97,50±7,91	100,00±0,00	100,00±0,00	100,00±0,00	83,00±9,19
Wp+Naive Bayes	93,50±10,55	93,50±10,55	80,00±16,83	86,50±11,80	84,50±14,42	85,00±14,14	91,00±11,74
Wp+SVM	89,50±17,39	96,00±8,43	94,00±9,66	91,50±11,07	93,50±10,55	91,00±11,74	94,00±9,66

Fonte: Autoria própria

Tabela 22 - Resultado em % dos classificadores - Base DLBCL - PCA95%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	100,00±0,00	98,00±6,32	57,50±16,20*	58,00±12,06*	55,00±11,79*	55,00±11,79*	100,00±0,00
CSE	86,50±16,67	90,50±17,07	91,50±11,07	93,50±10,55	93,50±10,55	96,00±8,43	93,50±10,55
Wp+J48	93,50±10,55	91,50±11,07	87,50±10,87	93,50±10,55	93,50±10,55	93,50±10,55	93,50±10,55
Wp+1-NN	84,00±16,30	89,00±15,06	93,50±10,55	86,50±16,67	91,00±11,74	92,50±16,87	93,00±11,35
Wp+3-NN	86,50±15,10	86,50±15,10	91,00±11,74	97,50±7,91	93,00±11,35	93,00±11,35	97,50±7,91
Wp+5-NN	93,50±10,55	97,50±7,91	87,00±11,35	85,00±14,14	89,50±11,17	89,50±11,17	96,00±8,43
Wp+7-NN	89,00±16,63	83,00±9,19	97,50±7,91	100,00±0,00	100,00±0,00	100,00±0,00	92,50±12,08
Wp+Naive Bayes	91,50±11,07	91,00±11,74	91,50±11,07	93,50±10,55	93,50±14,15	95,50±9,56	96,00±8,43
Wp+SVM	89,50±17,39	94,00±9,66	94,00±9,66	91,50±11,07	93,50±10,55	91,00±11,74	96,00±8,43

Fonte: Autoria própria

Tabela 23 - Resultado em % dos classificadores - Base DLBCL - PCA99%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	93,00±16,36	98,00±6,32	59,00±21,96*	51,00±8,76*	53,50±11,56*	55,00±11,79*	100,00±0,00
CSE	86,50±16,67	90,50±17,07	91,50±11,07	93,50±10,55	93,50±10,55	96,00±8,43	93,50±10,55
Wp+J48	93,50±10,55	91,50±11,07	87,50±10,87	93,50±10,55	93,50±10,55	93,50±10,55	93,50±10,55
Wp+1-NN	84,00±16,30	89,00±15,06	93,50±10,55	86,50±16,67	91,00±11,74	92,50±16,87	93,00±11,35
Wp+3-NN	86,50±15,10	86,50±15,10	91,00±11,74	97,50±7,91	93,00±11,35	93,00±11,35	97,50±7,91
Wp+5-NN	93,50±10,55	97,50±7,91	87,00±11,35	85,00±14,14	89,50±11,17	89,50±11,17	96,00±8,43
Wp+7-NN	89,00±16,63	83,00±9,19	97,50±7,91	100,00±0,00	100,00±0,00	100,00±0,00	92,50±12,08
Wp+Naive Bayes	91,50±11,07	91,00±11,74	91,50±11,07	93,50±10,55	93,50±14,15	95,50±9,56	96,00±8,43
Wp+SVM	89,50±17,39	94,00±9,66	94,00±9,66	91,50±11,07	93,50±10,55	91,00±11,74	96,00±8,43

Fonte: Autoria própria

Nas Tabelas 24, 25 e 26 são apresentados as taxas de classificação dos subconjuntos com 90%, 95% e 99% dos atributos da base DLBCL.

Tabela 24 - Resultado em % dos classificadores - Base DLBCL-NIH - PCA90%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	72,92±10,62	66,25±8,88	55,83±8,15*	57,08±8,80	57,08±7,62*	57,50±7,03*	72,50±11,98
CSE	66,25±8,21	66,67±6,80	53,75±11,19*	57,08±6,53*	58,75±6,35	57,08±6,23	65,42±12,43
Wp+J48	64,17±6,27	**75,83±7,30**	50,83±8,74*	51,25±8,57	53,33±10,90	54,17±7,86*	62,92±6,93
Wp+1-NN	59,58±10,03	62,50±6,80	**80,00±8,52**	69,17±7,40	67,92±8,11	69,17±6,27	65,00±5,62
Wp+3-NN	59,58±9,01	56,25±4,91	66,67±8,78	**76,25±8,57**	70,42±12,02	68,33±10,97	65,83±9,58
Wp+5-NN	55,42±9,01	57,50±1,76	55,83±5,62	64,17±8,15	67,08±7,97	63,33±9,98	56,25±2,95
Wp+7-NN	57,92±10,84	63,33±6,15	68,33±8,83	73,75±10,22	**77,92±8,57**	77,50±7,91	65,42±5,22
Wp+Naive Bayes	75,83±8,29	68,75±9,47	50,83±9,17*	52,50±8,83*	59,58±7,10*	57,50±5,12*	66,67±8,33*
Wp+SVM	64,58±9,47	69,17±9,04	55,83±9,86	54,17±12,11	56,25±11,99	55,83±8,83	**74,17±7,03**

Fonte: Autoria própria

Tabela 25 - Resultado em % dos classificadores - Base DLBCL-NIH - PCA95%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	72,92±10,62	66,25±8,88	55,83±8,15*	57,08±8,80	57,08±7,62*	57,50±7,03*	72,50±11,98
CSE	66,25±8,21	66,67±6,80	53,75±11,19*	57,08±6,53*	58,75±6,35	57,08±6,23	65,42±12,43
Wp+J48	64,17±6,27	**75,83±7,30**	50,83±8,74*	51,25±8,57	53,33±10,90	54,17±7,86*	62,92±6,93
Wp+1-NN	59,58±10,03	62,50±6,80	**80,00±8,52**	69,17±7,40	67,92±8,11	69,17±6,27	65,00±5,62
Wp+3-NN	59,58±9,01	56,25±4,91	66,67±8,78	**76,25±8,57**	70,42±12,02	68,33±10,97	65,83±9,58
Wp+5-NN	55,42±9,01	57,50±1,76	55,83±5,62	64,17±8,15	67,08±7,97	63,33±9,98	56,25±2,95
Wp+7-NN	57,92±10,84	63,33±6,15	68,33±8,83	73,75±10,22	**77,92±8,57**	**77,50±7,91**	65,42±5,22
Wp+Naive Bayes	75,83±8,29	68,75±9,47	50,83±9,17*	52,50±8,83*	59,58±7,10*	57,50±5,12*	66,67±8,33*
Wp+SVM	64,58±9,47	69,17±9,04	55,83±9,86	54,17±12,11	56,25±11,99	55,83±8,83	**74,17±7,03**

Fonte: Autoria própria

Tabela 26 - Resultado em % dos classificadores - Base DLBCL-NIH - PCA99%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	66,67±7,35	57,92±10,48	53,75±7,72*	55,00±9,58*	54,58±7,72*	57,92±7,47	70,00±9,17
CSE	62,08±10,29	55,83±8,61	57,92±7,20	55,42±7,36	57,08±10,40	56,25±9,67	62,92±10,10
Wp+J48	57,92±6,93	60,00±11,82	52,92±11,29	53,33±5,49	50,83±5,12	51,67±7,40	62,50±8,56
Wp+1-NN	54,58±11,02	51,67±7,40	**66,25±11,19**	67,08±10,66	62,92±6,65	63,75±9,01	66,25±9,10
Wp+3-NN	61,67±9,58	60,42±9,87	60,83±8,83	65,42±9,63	62,50±9,62	65,83±9,38	67,08±6,65
Wp+5-NN	55,00±8,52	57,50±1,76	55,83±7,91	52,92±9,01	56,67±10,43	61,25±9,01	56,67±3,51
Wp+7-NN	60,00±9,46	52,92±5,57	61,67±8,96	**67,50±9,17**	67,08±8,88	65,83±4,73	**69,58±8,11**
Wp+Naive Bayes	66,25±8,66	66,25±11,19	55,83±8,15*	62,92±8,21	65,83±10,17	65,42±9,22	67,92±10,40
Wp+SVM	64,58±6,87	57,92±6,65	61,25±9,63	61,25±11,29	60,83±9,46	62,92±9,31	**73,75±9,83**

Fonte: Autoria própria

Nas Tabelas 27, 28 e 29 são apresentados as taxas de classificação dos subconjuntos com 90%, 95% e 99% dos atributos da base DLBCL-Outcome.

Tabela 27 - Resultado em % dos classificadores - Base DLBCL-Outcome - PCA90%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	70,00±22,44	73,33±20,12	52,33±25,63	54,00±26,00	47,00±32,83	58,00±25,10	77,00±15,19
CSE	66,67±18,19	72,00±13,81	73,00±23,28	73,33±20,12	75,33±15,33	73,67±17,17	**81,00±12,28**
Wp+J48	53,00±15,98	72,00±20,26	62,33±16,78	**72,33±14,32**	67,00±15,90	63,33±13,97	**71,67±14,51**
Wp+1-NN	52,67±16,98	42,67±10,63	79,33±17,13	70,67±13,86	67,33±21,36	55,00±25,20	53,33±25,04
Wp+3-NN	54,67±13,54	73,67±17,95	74,00±18,31	76,00±11,31	72,00±20,92	70,67±20,78	**80,33±16,29**
Wp+5-NN	80,33±18,82	82,67±15,78	82,00±15,57	85,33±19,06	89,67±14,27	89,67±14,27	84,00±17,55
Wp+7-NN	62,00±24,56	65,33±20,14	70,67±17,76	83,00±22,25	85,00±24,15	81,67±22,84	**80,67±17,34**
Wp+Naive Bayes	69,00±15,32	75,33±19,64	79,33±13,03	65,33±27,85	65,33±24,30	63,33±22,28	**84,33±14,74**
Wp+SVM	48,33±10,57	55,33±7,24	65,67±15,48	43,00±24,52	44,67±25,20	53,00±17,81	44,33±18,60

Fonte: Autoria própria

Tabela 28 - Resultado em % dos classificadores - Base DLBCL-Outcome - PCA95%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	65,33±19,45	76,67±18,53	49,33±24,59	56,67±14,82	47,67±21,08	57,00±24,82	78,67±18,54
CSE	70,00±22,44	72,00±13,81	66,33±21,51	74,00±12,25	67,33±12,15	68,67±15,89	77,67±13,52
Wp+J48	53,00±15,98	72,00±20,26	62,33±16,78	**72,33±14,32**	67,00±15,90	63,33±13,97	**71,67±14,51**
Wp+1-NN	52,67±16,98	42,67±10,63	79,33±17,13	70,67±13,86	67,33±21,36	55,00±25,20	53,33±25,04
Wp+3-NN	58,67±19,26	75,67±19,75	74,00±18,31	77,67±16,48	76,00±13,77	77,67±16,48	78,67±16,79
Wp+5-NN	78,67±17,58	81,00±14,58	84,00±15,70	88,00±15,96	93,00±12,01	91,33±12,09	82,33±18,40
Wp+7-NN	62,00±24,56	65,33±20,14	70,67±17,76	83,00±22,25	**85,00±24,15**	81,67±22,84	**80,67±17,34**
Wp+Naive Bayes	69,00±15,32	75,33±19,64	79,33±13,03	65,33±27,85	65,33±24,30	63,33±22,28	**84,33±14,74**
Wp+SVM	48,33±10,57	55,33±7,24	65,67±15,48	43,00±24,52	44,67±25,20	53,00±17,81	44,33±18,60

Fonte: Autoria própria

Tabela 29 - Resultado em % dos classificadores - Base DLBCL-Outcome - PCA99%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	66,67±21,31	68,00±20,07	54,67±26,54	53,33±28,37	63,00±19,65	63,00±19,65	78,67±18,54
CSE	70,00±22,44	72,00±13,81	66,33±21,51	74,00±12,25	67,33±12,15	68,67±15,89	77,67±13,52
Wp+J48	53,00±15,98	72,00±20,26	62,33±16,78	**72,33±14,32**	67,00±15,90	63,33±13,97	**71,67±14,51**
Wp+1-NN	51,33±17,16	44,33±10,31	75,00±19,64	69,00±18,73	65,67±20,61	62,00±22,51	56,67±16,78
Wp+3-NN	58,33±21,67	75,67±19,75	70,67±17,76	78,00±17,58	68,67±18,47	75,67±19,75	78,67±19,26
Wp+5-NN	78,67±17,58	81,00±14,58	84,67±16,64	90,00±14,05	80,67±24,54	81,00±22,06	82,33±18,40
Wp+7-NN	62,00±26,95	65,33±20,14	72,67±17,55	70,67±24,98	65,67±21,89	73,67±17,95	79,00±17,85
Wp+Naive Bayes	58,67±16,57	**75,33±19,64**	67,67±26,20	65,33±27,85	62,00±25,10	65,67±28,07	**78,67±20,80**
Wp+SVM	48,33±10,57	55,33±7,24	65,67±15,48	43,00±24,52	44,67±25,20	53,00±17,81	44,33±18,60

Fonte: Autoria própria

Nas Tabelas 30, 31 e 32 são apresentados as taxas de classificação dos subconjuntos com 90%, 95% e 99% dos atributos da base DLBCL-Tumor.

Tabela 30 - Resultado em % dos classificadores - Base DLBCL-Tumor - PCA90%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	93,21±9,85	98,75±3,95	70,36±17,05*	75,36±20,46	81,43±17,40	77,86±10,52*	96,07±6,34
CSE	85,54±15,62	80,36±14,60	93,21±11,47	89,29±13,88	86,79±11,24	88,21±9,90	90,89±10,96
Wp+J48	80,36±9,37	80,18±16,74	82,86±10,98	85,89±16,40	85,89±16,40	85,89±16,40	88,39±11,49
Wp+1-NN	84,46±7,76	85,71±9,18	88,04±9,93	90,71±6,45	88,04±13,76	90,89±8,64	92,14±6,80
Wp+3-NN	85,36±14,82	80,36±12,54	81,79±10,71	78,93±16,19	82,86±9,27	84,29±8,37	87,86±12,02
Wp+5-NN	88,21±11,52	90,89±8,64	96,07±6,34	96,07±6,34	94,64±6,94	95,89±6,63	95,89±6,63
Wp+7-NN	85,54±9,72	84,46±10,28	92,32±8,87	90,71±9,33	93,39±6,99	92,14±9,57	81,96±13,64
Wp+Naive Bayes	83,04±13,84	94,82±8,93	**97,32±5,66**	**94,64±6,94**	90,71±9,33	93,21±11,47	**95,89±6,63**
Wp+SVM	73,75±7,34	75,36±3,93	70,00±12,22	66,25±6,32	68,75±6,85	77,86±8,72	75,36±3,93

Fonte: Autoria própria

Tabela 31 - Resultado em % dos classificadores - Base DLBCL-Tumor - PCA95%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	89,29±8,83	98,75±3,95	70,18±10,52*	75,18±15,51	74,11±15,84	75,18±11,68*	**97,32±5,66**
CSE	85,54±15,62	86,96±12,26	89,46±10,05	90,89±10,46	89,64±8,07	86,96±10,75	90,71±11,03
Wp+J48	80,36±9,37	80,18±16,74	82,86±10,98	85,89±16,40	85,89±16,40	85,89±16,40	88,39±11,49
Wp+1-NN	81,79±11,19	82,86±12,88	93,21±7,18	89,11±11,15	83,93±16,13	84,29±10,74	90,71±9,33
Wp+3-NN	92,14±9,00	84,46±7,76	93,39±6,99	93,21±9,85	93,21±9,85	90,36±15,04	95,71±6,90
Wp+5-NN	88,21±11,52	90,89±8,64	96,07±6,34	96,07±6,34	94,64±6,94	95,89±6,63	95,89±6,63
Wp+7-NN	80,54±13,56	88,21±9,90	96,07±6,34	94,64±6,94	93,39±6,99	93,39±6,99	92,32±8,87
Wp+Naive Bayes	89,46±8,77	94,82±8,93	94,46±9,83	91,96±9,68	89,46±10,56	93,21±9,85	**97,32±5,66**
Wp+SVM	73,75±7,34	75,36±3,93	70,00±12,22	66,25±6,32	68,75±6,85	77,86±8,72	75,36±3,93

Fonte: Autoria própria

Tabela 32 - Resultado em % dos classificadores - Base DLBCL-Tumor - PCA99%

Subconjunto	Naive Bayes	J48	1-NN	3-NN	5-NN	7-NN	SVM
CFS	85,54±12,81	98,75±3,95	60,89±15,71*	64,29±15,88*	60,71±16,47*	63,04±16,52	**98,57±4,52**
CSE	85,54±15,62	86,96±12,26	89,46±10,05	90,89±10,46	89,64±8,07	86,96±10,75	90,71±11,03
Wp+J48	80,36±9,37	80,18±16,74	82,86±10,98	85,89±16,40	85,89±16,40	85,89±16,40	88,39±11,49
Wp+1-NN	81,79±11,19	82,86±12,88	93,21±7,18	89,11±11,15	83,93±16,13	84,29±10,74	90,71±9,33
Wp+3-NN	92,14±9,00	84,46±7,76	93,39±6,99	93,21±9,85	93,21±9,85	90,36±15,04	95,71±6,90
Wp+5-NN	84,11±10,70	90,89±8,64	**96,07±6,34**	**97,32±5,66**	**95,89±6,63**	**97,14±9,04**	95,89±6,63
Wp+7-NN	80,54±13,56	88,21±9,90	96,07±6,34	94,64±6,94	93,39±6,99	93,39±6,99	92,32±8,87
Wp+Naive Bayes	89,46±8,77	94,82±8,93	94,46±9,83	91,96±9,68	89,46±10,56	93,21±9,85	**97,32±5,66**
Wp+SVM	73,75±7,34	75,36±3,93	70,00±12,22	66,25±6,32	68,75±6,85	77,86±8,72	75,36±3,93

Fonte: Autoria própria