HELYANE BRONOSKI BORGES

REDUÇÃO DE DIMENSIONALIDADE EM BASES DE DADOS DE EXPRESSÃO GÊNICA

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná, como requisito parcial para obtenção do título de Mestre em Informática.

CURITIBA
2006

HELYANE BRONOSKI BORGES

REDUÇÃO DE DIMENSIONALIDADE EM BASES DE DADOS DE EXPRESSÃO GÊNICA

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná, como requisito parcial para obtenção do título de Mestre em Informática.

Área de Concentração: Descoberta do Conhecimento e Aprendizagem de Máquina

Orientador: Prof. Dr. Júlio Cesar Nievola

CURITIBA

2006

Borges, Helyane Bronoski
Redução de Dimensionalidade de Atributos em Bases de Dados de Expressão Gênica. Curitiba, 2006. 123p.

Dissertação (Mestrado) – Pontifícia Universidade Católica do Paraná. Programa de Pós Graduação em Informática.

1. Mineração de Dados. 2. Seleção de Atributos. 3. Projeção Aleatória. 4. Bioinformática. I. Pontifícia Universidade Católica do Paraná. II. Centro de Ciências Exatas e de Tecnologia. III. Programa de Pós Graduação em Informática.

Esta página deve ser reservada à ata de defesa e termo de aprovação que serão fornecidos pela secretaria após a defesa da dissertação e efetuada as correções solicitadas.

Dedicatória

Aos meus pais pelo sacrifício realizado em meu benefício, pela dedicação e constante apoio.

Agradecimentos

A Deus por ter me guiado e concedido discernimento para fazer as escolhas certas como as que venho fazendo.

Aos meus pais por me fornecerem suporte em todos os momentos de minha vida e por todos os sacrifícios que fizeram em prol de meu beneficio. Espero um dia poder retribuir uma parte de tudo o que vocês fizeram por mim.

Ao meu orientador, Júlio Cesar Nievola, pela confiança, incentivo e paciência com que me acompanhou durante o desenvolvimento desse trabalho.

Aos meus queridos amigos: Andréia, Daniele, Euclides, Edson, Luis Gustavo, Márcio e Richard, pelo auxílio e pela palavra amiga.

E a todos aqueles que de alguma maneira contribuíram para que esse trabalho fosse realizado.

SUMÁRIO

Lista de Abreviaturas e Siglas

ADCA	Adenocarcionama
ALL	*Acute Lymphoblastic Leukemia*
	Leucemia Linfoblática Aguda
AM	Aprendizagem de Máquina
AML	*Acute Myeloid Leukemia*
	Leucemia Mieloide Aguda
CFS	*Correlation-basead Feature Selection*
DNA	*Deoxyribonucleic Acid*
	Ácido Desoxirribonucléico
cDNA	Ácido Desoxirribonucléico Complementar
IA	Inteligência Artificial
KDD	*Knowledge Discovery in Databases*
	Descoberta do Conhecimento em Bases de Dados
k-NN	*k-Nearest Neigboor*
	k Vizinhos mais Próximos
LDGCB	Linfoma Difuso de Grandes Células B
LF	Linfoma Folicular
LSI	*Latent Semantic Analysis*
LVF	*Las Vegas Filter*
LVI	*Las Vegas Incremental Filter*
MPL	Mesotelioma Pleural Maligno
MPSS	*Massively Parallel Signature Sequence Technology*
mRNA	Ácido Ribonucléico Mensageiro
PCA	*Principal Component Analysis*
	Análise de Componentes Principais
PCL	*Prediction by Collective Likelihood of Emerging Patterns*
RNA	*Ribonucleic Acid*
	Ácido Ribonucléico
RT-PCR	*Reverse-Transcription Polymerase Chain Reaction*
SA	Seleção de Atributos
SAGE	*Serial Analysis of Gene Expression*
SBG	*Sequential Backward Generation*
	Geração Seqüencial para trás

SFG	*Sequential Forward Generation*
	Geração Seqüencial para frente
SVM	*Support Vector Machines*
	Máquinas de Vetores Suporte
SGBD	Sistemas Gerenciadores de Banco de Dados
SOM	*Self Organizing Map*

Lista de Equações

Lista de Figuras

Lista de Tabelas

Resumo

O rápido desenvolvimento das pesquisas na área de genoma e proteoma têm contribuído para o crescimento acelerado das bases de dados biológicas, inviabilizando a análise humana sem algum suporte tecnológico.

Uma das características desses tipos de dados é que eles possuem um número grande de atributos (genes) e um número pequeno de amostras, o que compromete o desempenho do algoritmo de mineração de dados. A utilização de métodos de redução de dimensionalidade, tal como a seleção de atributos, permite, além da remoção de dados redundantes e irrelevantes, uma melhor compreensibilidade dos resultados gerados, identificando a influência de cada atributo selecionado e do seu nível de expressão de acordo com o objetivo desejado.

Nesse trabalho é apresentado um estudo comparativo de métodos de redução de dimensionalidade aplicados em cinco bases de expressão gênica. Os métodos aplicados: são: a seleção de atributos e o método de projeção aleatória. Ambos os métodos serão usados como uma etapa de pré-processamento na Mineração de Dados.

A seleção de atributos tem como objetivo descobrir um subconjunto de atributos relevantes para uma tarefa alvo, considerando os atributos originais, e é importante, entre outras coisas, por tornar o processo de aprendizagem mais eficiente. A seleção de atributos é um método de redução de dimensionalidade que obtém resultados promissores quando aplicado em bases de expressão gênica.

O método de projeção aleatória é um método alternativo, pois, além de diminuir o custo computacional quando aplicado, principalmente em conjunto com a seleção de atributos, produz resultados significativos.

Os resultados dos experimentos mostram que a aplicação desses métodos de redução de dimensionalidade produz uma taxa de acerto do classificador maior do que quando aplicado somente o algoritmo de mineração sobre as bases de dados com todos os atributos.

Palavras Chave: Mineração de Dados, Seleção de Atributos, Projeção Aleatória, Bioinformática.

Abstract

The fast development of the research in the genome and proteoma areas has contributed for the accelerated growth of the biological databases, making impracticable the manual analysis without some technological support.

One of the characteristics of these types of data is that they possess a large amount of attributes (genes) and a small number of samples: it compromises the performance of the mining algorithm. Thus the use of methods of dimension reduction, as the selection of attributes, allows not only the removal of redundant and irrelevant data but also a better understanding of the generated results, identifying the influence of each selected attribute and its level of expression in accordance with the desired objective.

In this work a comparative study of methods of reduction of dimension applied in five bases of gene expression is made. The methods of attribute selection and the random projection method were applied. Both methods will be used as a stage of pre-processing in the Data Mining.

The selection of attributes has as objective the discovery of a excellent subset of attributes for the given task, considering the original attributes and it is important, among others things, because the learning process becomes more efficient. It is a method of reduction of dimension that gives good results when applied in bases of gene expression.

The random projection method is an alternative method: besides diminishing the computational cost when applied, mainly together with the selection of attributes, it produces significant results.

The results of the experiments show that the applications of these methods of dimension reduction produce a classifier performance better than when applied only the mining algorithm on the databases with all the attributes.

Keywords: Data Mining, Attributes Selection, Random Projection, Bioinformatics.

1 Introdução

Os dados biológicos estão sendo disponibilizados em quantidades elevadas, fazendo com que os bancos de dados atuais cresçam exponencialmente [BAL01]. Esse fenômeno vem sendo causado pela utilização de técnicas novas e eficientes na análise de seqüências de genoma e proteoma. Nesse contexto, surgiu uma nova área em Ciência da Computação, que é a Biologia Computacional, também denominada Bioinformática.

O emprego de métodos computacionais na Biologia iniciou-se na década de 1980 [ALB97], quando biólogos experimentais, em conjunto com cientistas da computação, físicos e matemáticos, começaram a aplicar esses métodos na modelagem de sistemas biológicos. Durante esse período, ferramentas computacionais foram desenvolvidas por essa comunidade para análise dos dados, utilizando algoritmos convencionais da Ciência da Computação [SOU03].

Na segunda metade da década de 90, com o surgimento dos seqüenciadores automáticos de Ácido Desoxirribonucléico (DNA), houve uma explosão na quantidade de seqüências a serem armazenadas, exigindo recursos computacionais cada vez mais eficientes [SOU03]. Além do armazenamento ocorria, paralelamente, a necessidade de análise desses dados, o que tornava indispensável a utilização de plataformas computacionais eficientes para a interpretação dos resultados obtidos.

Assim nascia a bioinformática. Essa nova ciência envolveria a união de diversas linhas de conhecimento - a engenharia de *software*, a matemática, a estatística, a ciência da computação e a biologia molecular.

A Bioinformática ou Biologia Computacional diz respeito à utilização de técnicas e ferramentas de computação para a resolução de problemas da Biologia [BAL01]. Dentre as diversas áreas da Biologia, aquela em que a aplicação de técnicas computacionais tem se mostrado mais promissora é a Biologia Molecular [SET97]. Nesse contexto, a computação pode ser aplicada na resolução de problemas como comparação de seqüências (DNA, Ácido Ribonucléico – RNA e proteínas), montagem de fragmentos, reconhecimento de genes, identificação e análise da expressão de genes e determinação da estrutura de proteínas [SET97], [BAL01].

A análise da expressão dos genes é de grande interesse para as Ciências Biológicas. Esse tipo de análise pode fornecer informações importantes sobre as funções de uma célula, uma vez que as mudanças na fisiologia de um organismo são geralmente acompanhadas por mudanças nos padrões de expressão dos genes [ALB97].

Diversas técnicas têm sido propostas para obtenção da expressão dos genes entre elas a de microarranjos de DNA. Essas técnicas podem ser utilizadas em estudos de genomas inteiros, da expressão de genes ativos, no ordenamento e seqüenciamento dos genes, na determinação de variantes genéticas, em diagnósticos de doenças e várias outras aplicações.

A tecnologia de microarranjo de DNA tem sido a ferramenta padrão na maioria dos laboratórios de pesquisa genômica. A razão para esta popularidade deve-se ao fato de que os microarranjos permitem que o processo de análise de expressão gênica possa ser realizado em larga escala e em tempo viável. Esta tecnologia possibilita o monitoramento simultâneo de níveis transcricionais de todos os genes ou de uma porção significante de genes de um organismo em um determinado estado.

À medida que os experimentos de microarranjo são realizados e se tornam procedimentos rotineiros, os dados gerados vão sendo acumulados rapidamente, tornando prioritário o desenvolvimento de métodos eficientes que possam explorar o potencial desta tecnologia.

A utilização da Inteligência Artificial bem como a aplicação dos Algoritmos de Aprendizagem de Máquina vem sendo cada vez mais empregada para tratar problemas em Biologia Molecular, pela sua capacidade de aprender automaticamente a partir de grandes volumes de dados e produzir hipóteses úteis [BAL01].

A Inteligência Artificial contempla diversas técnicas entre elas a Mineração de Dados aplicada nesse trabalho. Uma das suas metas está relacionada com a descoberta do conhecimento, a qual constitui um dos principais desafios da área de Aprendizado de Máquina.

As bases de dados de microarranjos possuem características particulares diferentes das bases tradicionais o que requer um tratamento diferencial no processo de mineração. Esses tipos de dados apresentam novos desafios aos algoritmos de Aprendizagem de Máquina. Sendo assim, métodos de redução de dados como a seleção de atributos têm sido aplicada nesses tipos de dados, visto que a maioria dos algoritmos de AM não trabalha bem na presença de um número grande de atributos podendo "atrapalhar" o processo de aprendizagem e o desempenho das predições.

Nesse trabalho são aplicadas técnicas de redução da dimensionalidade dos dados em bases de microarranjos. Duas abordagens diferentes são utilizadas: a Seleção de Atributos e a Projeção Aleatória para a tarefa de classificação.

A seleção de atributos é uma técnica que vem contribuindo para um aumento na aplicação prática de métodos de AM, sobretudo em situações com um grande número de atributos, como é o caso nas bases de dados de expressão gênica, tornando possível o

aprendizado em situações antes impossíveis [LIU02]. Já a projeção aleatória é um método que reduz da dimensionalidade dos dados e tem um custo computacional menor comparado com a seleção de atributos.

1.1 Objetivos

1.1.1 Objetivo Geral

Analisar o efeito de métodos de redução de dimensionalidade, através da seleção de atributos e do método de projeção aleatória, para a tarefa de classificação, em dados de expressão gênica obtida através da técnica de microarranjos.

1.1.2 Objetivos Específicos

Como objetivos específicos têm-se:

- Analisar o comportamento de técnicas de mineração de dados em bases de dados de expressão gênica obtidos pela técnica de microarranjos;

- Analisar o desempenho dos classificadores quando utilizado em bases de dados com todos os atributos e em subconjuntos de atributos gerados por métodos de redução de dimensionalidade;

- Verificar o desempenho dos algoritmos de seleção de atributos, pertencentes à abordagem filtro e *wrapper*, para a área de bioinformática;

- Comparar as abordagens de seleção de atributos;

- Analisar o comportamento do método de projeção aleatória em bases de dados de microarranjos;

- Utilizar os métodos de seleção de atributos em conjunto com o método projeção aleatória;

- Identificar a melhor técnica de redução de dimensionalidade em bases de dados de microarranjos para as bases de dados escolhidas.

1.2 Estrutura do Trabalho

Este trabalho está organizado em 5 capítulos. O capítulo 2 apresenta a fundamentação teórica, descrevendo conceitos importantes para esse trabalho. O capítulo 3 descreve a metodologia aplicada nessa pesquisa. O capítulo 4 apresenta os resultados obtidos. No capítulo 5 é apresentada a conclusão do trabalho e as perspectivas para novas pesquisas.

O trabalho apresenta ainda 3 apêndices com informações adicionais sobre o desenvolvimento do trabalho. No apêndice A são apresentados os genes mais

selecionados pelos métodos de seleção de atributos. No apêndice B são apresentadas as médias de número de atributos selecionados em cada base de dados para cada um dos métodos de seleção. No apêndice C são apresentadas maiores informações sobre a estatística do teste *t* pareado, método usado para avaliar os resultados obtidos.

2 Fundamentação Teórica

O advento da tecnologia de microarranjos de DNA vem proporcionando aos biólogos a possibilidade de medir o nível de expressão de milhares de genes em um único experimento. Com isso, cada vez mais, o volume de dados excede a capacidade de sua análise pelos métodos tradicionais. Para atender a esta necessidade o processo de descoberta do conhecimento (KDD), do inglês *Knowledge Discovery in Databases,* tem sido utilizado, pois permite a extração automática do conhecimento a partir de grandes volumes de dados nas mais diversas áreas de aplicação, entre elas a Biologia Molecular.

2.1 Principais Conceitos de Biologia Molecular

Todos os seres vivos, dos mais simples aos mais complexos, são constituídos por células. A Biologia Molecular retrata o estudo das células e moléculas, blocos básicos utilizados na construção de todas as formas de vida [CAS92]. Em particular, estuda-se o genoma dos organismos, definido como o conjunto de suas informações genéticas. Gregor Mendel, em seus experimentos realizados no século XVII, foi o primeiro a identificar fatores responsáveis pela hereditariedade nos organismos vivos.

Esses fatores foram, posteriormente, denominados de genes, os quais codificam a informação genética. Na busca pela localização dos genes foram identificados os cromossomos, que são estruturas que possuem capacidade de replicação (reprodução) e estão presentes em todas as células. Estudos acerca dos cromossomos, por sua vez, levaram à descoberta de que eles são compostos por moléculas de DNA e que genes são seqüências contíguas de DNA.

2.1.1 DNA, Expressão Gênica e Proteínas

Uma molécula de DNA consiste de duas fitas anti-paralelas entrelaçadas em forma de dupla hélice, conforme pode ser visualizado na Figura 1. Cada fita é composta por uma seqüência de nucleotídeos (bases), que podem ser de quatro tipos: Adenina (A), Guanina (G), Citosina (C) e Timina (T). Cada nucleotídeo de uma fita se liga a outro complementar da segunda, conforme a regra: A - T, T - A, C - G e G – C.

Um fragmento de DNA pode conter diversos genes. A propriedade mais importante dos genes está no fato de que eles codificam proteínas, componente essencial de todo ser vivo. As proteínas possuem diversas funções biológicas [LEW01]. Elas podem ter papel estrutural, como no caso do colágeno presente nos tendões, ou

estar ligadas a atividades regulatórias, como no caso das enzimas, que catalisam diversas reações químicas nas células.

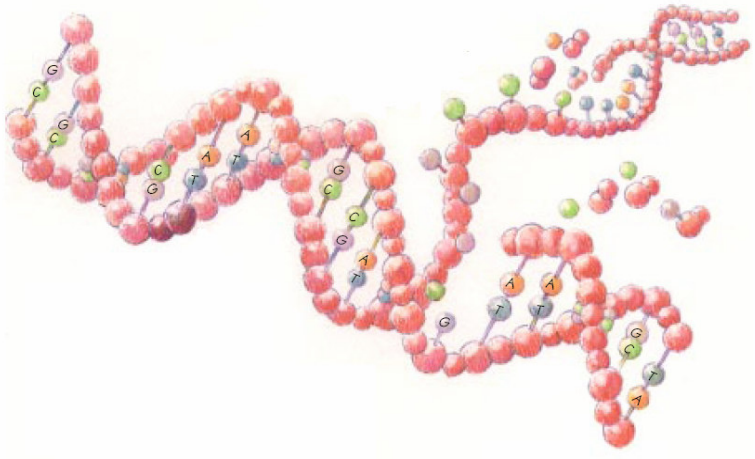

Figura 1: Estrutura de uma molécula de DNA [ALB97].

As proteínas também são seqüências lineares, compostas de conjuntos de aminoácidos. O processo pelo qual as seqüências de nucleotídeos dos genes são interpretadas na produção de proteínas é denominado expressão gênica (Figura 2). A expressão é composta por duas etapas: na primeira, denominada transcrição, um RNA polimerase se liga a uma região do DNA denominada promotora e inicia a síntese de um RNA mensageiro (mRNA). O mRNA é similar ao DNA, com exceção de duas características: é composto por apenas uma fita e possui o nucleotídeo Uracila (U) no lugar do nucleotídeo Timina (T).

Na segunda etapa da expressão, denominada tradução, é realizada a síntese da molécula de proteína, a partir do mRNA. Cada grupo de três nucleotídeos do mRNA representa um aminoácido, constituinte de uma proteína. O código genético consiste no mapeamento desses grupos, também referenciados por códons, nos aminoácidos correspondentes.

Há 64 possíveis combinações de triplas de nucleotídeos, ou seja, 64 códons. Porém, existem apenas 20 aminoácidos. Portanto, muitos deles são mapeados por mais de um códon. Desses 64 códons, 3 são responsáveis por indicar o final da tradução, sendo denominados códons de parada. As diferentes codificações podem ser visualizadas na tabela 1. O primeiro, segundo e terceiro nucleotídeos dos códons são representados, respectivamente, pela coluna mais à esquerda, a primeira linha e a coluna mais à direita da tabela.

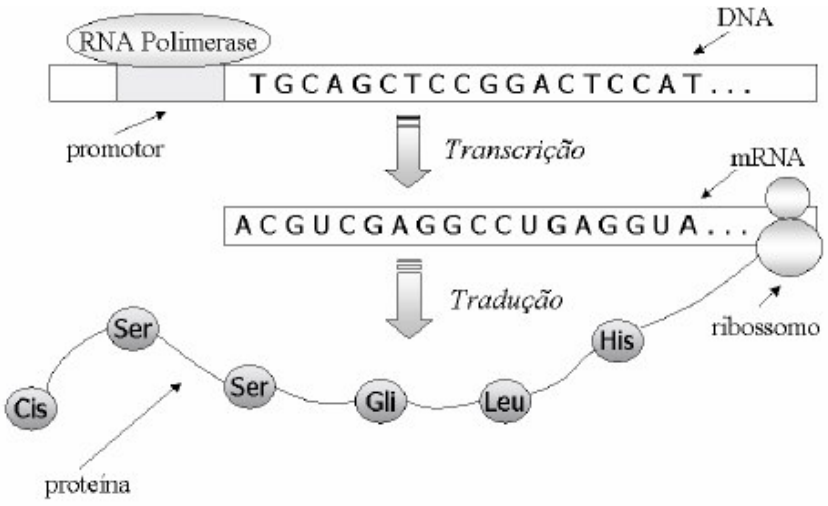

Figura 2: Processo de expressão gênica [ALB97].

Tabela 1: Código Genético [LEW01].

	U	C	A	G	
U	Phelinanina (Phe) Phe Leucina (Leu) Leu	Serina (Ser) Ser Ser Ser	Tirosina (Tir) Tir Parada Parada	Cisteína (Cis) Cis Parada Tritophan (Tri)	**U** **C** **A** **G**
C	Leu Leu Leu Leu	Prolina (Pro) Pro Pro Pro	Histidina (His) His Glutamina (Glu) Glu	Arginina (Arg) Arg Arg Arg	**U** **C** **A** **G**
A	Isoleucina (Iso) Iso Iso Metionina (Met)	Treonina (Tre) Tre Tre Tre	Aspargina (Asp) Asp Lisina (Lis) Lis	Ser Ser Arg Arg	**U** **C** **A** **G**
G	Valina (Val) Val Val Val	Alanina (Ala) Ala Ala Ala	Ácido Áspártico (Aca) Aca Ác. Glutamínico (Acg) Acg	Glicina (Gli) Gli Gli Gli	**U** **C** **A** **G**

Existem algumas diferenças na forma como os procedimentos descritos anteriormente ocorrem em organismos eucariotos (seres vivos complexos, tais como os humanos), que possuem o material genético em um núcleo delimitado por uma membrana, e procariotos (seres unicelulares, como por exemplo, as bactérias), que possuem o material genético difuso em suas células.

2.1.2 Experimentos com Expressão Gênica e a Técnica de Microarranjo

A análise da expressão dos genes é de grande interesse para as Ciências Biológicas. Esse tipo de análise pode fornecer informações importantes sobre as funções

de uma célula, uma vez que as mudanças na fisiologia de um organismo são geralmente acompanhadas por mudanças nos padrões de expressão dos genes [ALB97].

Diversas técnicas têm sido propostas para obtenção da expressão dos genes: MPSS (*Massively Parallel Signature Sequence technology*), SAGE (*Serial Analysis of Gene Expression*), Real-time RT-PCR (*Reverse-Transcription Polymerase Chain Reaction*) e microarranjo de DNA [SOU03]. Muitas dessas técnicas podem ser utilizadas em estudos de genomas inteiros, da expressão de genes ativos, no ordenamento e seqüenciamento dos genes, na determinação de variantes genéticas, em diagnósticos de doenças e várias outras aplicações.

A técnica de microarranjos de DNA tem revolucionado a pesquisa biológica em um curto tempo desde a sua existência [RUB03]. A análise de dados de microarranjos tornou-se uma ferramenta essencial na pesquisa biomédica [SHA03]. A técnica permite um estudo confiável da função e dos padrões de expressão dos genes tornando-se uma ferramenta poderosa na busca de causadores genéticos e tumores [MIL02].

Entre os principais tipos de análise destacam-se [TAM03]:

- Agrupamento: encontra novas classes biológicas ou refina a existência de algumas já conhecidas;

- Classificação: classificação de doenças ou resultados de predição baseados nos padrões de expressão dos genes;

- Seleção de genes: em mineração de dados, este é um processo de seleção de atributos no qual se encontram os atributos fortemente relevantes para uma classe em particular.

No caso de microarranjos de DNA, o principio básico empregado é o seguinte: moléculas de DNA complementar (cDNA[1]) ou oligonucleotídeos[2] correspondentes aos genes cuja expressão deve ser analisada (sondas) são afixadas, de uma maneira ordenada (*arrays*), a um suporte sólido que pode ser uma lâmina (*slide*) de vidro. A miniaturização e automação da criação dessas lâminas com o uso de robô (ou síntese *in situ* de oligonucleotídeos) tornou possível a sua produção com milhares de genes (isto é, uma parte substancial do genoma) representados em poucos centímetros quadrados – *microarrays* [SOU03].

Ainda no caso específico de um microarranjo de cDNA, primeiro as sondas são replicadas um grande número de vezes [DUG99]. Em seguida, um robô fixa essas sondas em determinados pontos (*spots*) da lâmina de vidro. Ao final, a lâmina conterá

---

[1] Moléculas de DNA produzidas a partir de um mRNA, e portanto, sem introns [2]

[2] Seqüências de DNA curtas de 20 a 30 nucleotídeos [2]

milhares de pontos com DNA, colocados lado a lado, cada ponto contendo milhares de sondas de cDNA que foram projetadas para hibridizar com o mRNA de um certo gene.

Em um próximo passo, para medir a abundância relativa dos transcritos correspondentes a uma determinada célula, as suas moléculas de mRNA são também transcritas para moléculas de cDNA. Essa transcrição é necessária porque moléculas de RNA são instáveis, tendendo a degradar rapidamente. Posteriormente, as moléculas de cDNA produzidas são, em geral, marcadas com rótulos fluorescentes verdes (Cy3). Da mesma forma, as moléculas de mRNA da célula de controle ou referência também são separadas e transcritas, sendo que nesse caso são marcadas com rótulos vermelhos (Cy5) [DUG99].

As moléculas de cDNA de ambas as células são então despejadas na lâmina. Depois de um certo tempo, a lâmina é lavada, removendo as moléculas de cDNA que não hibridizaram com as sondas. Em seguida, a lâmina é escaneada, produzindo como resultado uma imagem com as intensidades de todos os pontos (todo o processo é ilustrado na Figura 3). A imagem digital da lâmina é, por fim, processada por meio de métodos computacionais, com o objetivo de calcular a intensidade obtida para cada mRNA.

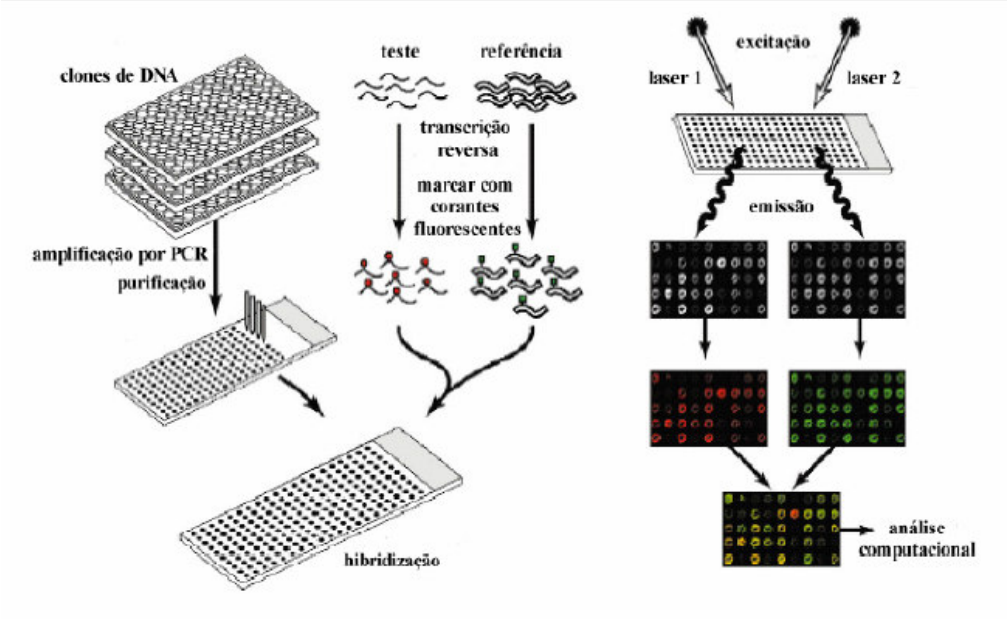

Figura 3: Esquema de um microarranjo de cDNA [DUG99].

Os dados de microarranjos são geralmente representados no formato de uma matriz (mostrada na figura 4). Cada coluna representa um gene e cada linha representa uma amostra com um rótulo c_i. O conteúdo da matriz é o nível de expressão de cada

gene sobre cada condição em que ele foi submetido. Para cada amostra o nível de expressão de todos os genes em estudo é medido, ou seja f_{ij} é a medida do nível de expressão do gene j para a amostra i onde $j = 1,...,N$ e $i = 1,...,M$. O formato do conjunto de dados de microarranjos é igual ao formato de dados convencionais da aprendizagem de máquina e em mineração dos dados, onde um gene pode ser considerado uma característica ou um atributo e cada amostra indica um exemplo ou um ponto de dados [YUL04].

Gene 1	Gene 2	. . .	Gene N	
f_{11}	f_{12}	. . .	f_{1N}	c_1
f_{21}	f_{22}	. . .	f_{1N}	c_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
f_{M1}	f_{M2}	. . .	f_{MN}	c_M

Figura 4: Exemplo de matriz de Expressão Gênica.

Uma característica das bases de dados de microarranjos é que, geralmente, ela é formada por grande quantidade de atributos, que correspondem aos genes, e um pequeno número de amostras. Isso acontece devido à grande quantidade de genes dos quais é medida a expressão, da ordem de milhares, e ao pequeno número de experimentos, dezenas a centenas, realizados devido ao elevado custo do processo.

2.2 Mineração de Dados e Descoberta do Conhecimento

O termo Descoberta de Conhecimento em Base de Dados conhecido também como *Knowledge Discovery in Databases* (KDD) é o processo de descobrir informação implícita e útil em grandes bases de dados [FAY96]. A importância da utilização da técnica de Descoberta de Conhecimento em Base de Dados está relacionada ao crescente aumento no volume de informações que não podem ser recuperadas adequadamente pelas limitações e capacidades de consultas, dos sistemas gerenciadores de banco de dados (SGBD) atuais. A Figura 5 apresenta as etapas do processo de descoberta de conhecimento, dividido em 3 grandes etapas e descritas na seqüência do trabalho.

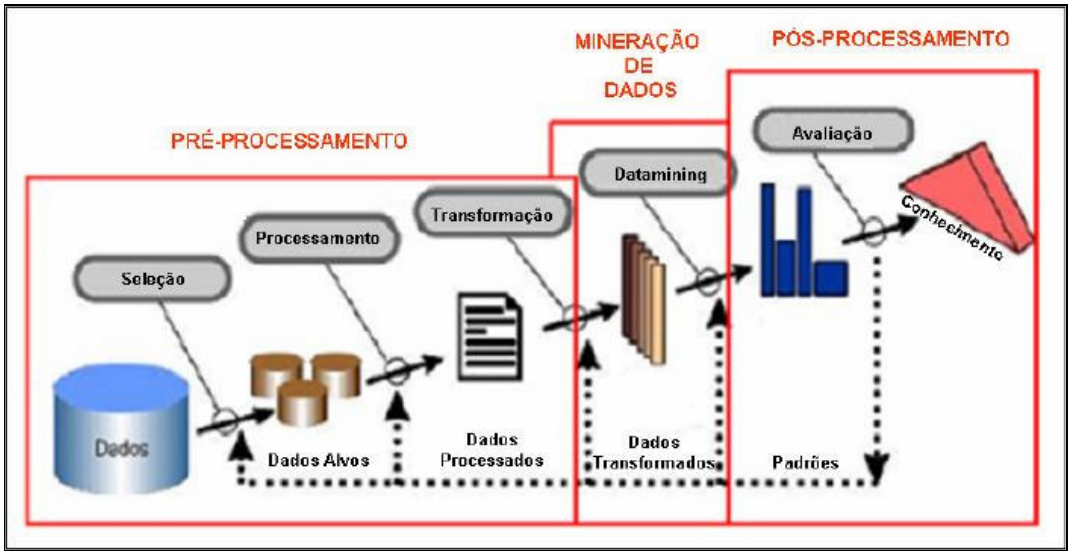

Figura 5: Etapas do Processo de Descoberta de Conhecimento [FAY96].

2.2.1 Etapas do Processo de Descoberta de Conhecimento

Independente da abordagem, o processo KDD é definido por um conjunto de etapas, envolvendo desde o entendimento do domínio da aplicação até a interpretação e consolidação dos resultados. Um exemplo desse processo pode ser observado em três grandes fases como o Pré-Processamento, que contempla a seleção, processamento e transformação dos dados, a Mineração de dados descobrindo padrões, e o Pós-Processamento que contempla os padrões e conhecimento descoberto, como mostra a Figura 5.

O processo de KDD é interativo e iterativo, ou seja, é necessária a intervenção do engenheiro do conhecimento para a execução dos procedimentos como a seleção e limpeza de dados até que obtenha dados expressivos, envolvendo numerosos passos com diversas decisões a serem tomadas pelo especialista e/ou engenheiro. Além disso, cada etapa é fundamental para que os objetivos estabelecidos sejam alcançados, sendo que para isso, estas etapas necessitam ser executadas corretamente [WON02].

a) Definição e Entendimento do Problema

O início do Processo KDD é realizado a partir da definição e entendimento do problema que se deseja resolver. Nesta fase é feita uma análise das atividades, com a finalidade de atingir os objetivos propostos. Assim, aumenta a possibilidade de que os resultados obtidos no processo serão úteis. O problema deve ser definido pelo engenheiro do conhecimento, juntamente com o especialista da área, tendo como

finalidade especificar o problema da melhor forma possível, para chegar a resultados positivos e úteis.

b) Pré-Processamento

Nesta fase, são realizadas operações básicas e de análise dos dados para definir pontos como estrutura das tabelas, valores potenciais para atributos, formatos e tipos de dados. Se houver necessidade serão removidos ruídos, coletadas informações para modelar, definidas estratégias para manusear e tratar campos que não influenciam na solução das perguntas que se deseja responder.

Além dessa etapa ter como objetivo disponibilizar para a etapa seguinte uma base de dados íntegra, consolidada e coerente, é o momento de definir a estratégia para resolver o problema significativo de ausência ou a não disponibilidade de dados.

Dependendo da complexidade, das fontes e do repositório destino dos dados, este procedimento pode variar de uma simples consulta a banco de dados, a um procedimento complexo de migração que inclui conversões de tipos e formatos de dados.

Depois de extraídos dos repositórios, os dados geralmente não estão prontos para mineração. Durante e depois do processo de extração, os dados devem ser formatados e algumas das operações mais comuns nesta fase são:

- Padronização de caracteres: É comum encontrar algoritmos que são sensíveis a certos caracteres especiais ou não diferenciam entre letras maiúsculas ou minúsculas.

Nestes casos, o engenheiro do conhecimento precisa transformar o fluxo de caracteres de entrada em um padrão aceitável pelo algoritmo de mineração:

- Concatenação: É comum que múltiplos campos de dados estejam codificando o valor de um só atributo para mineração.

- Formato de representação: É comum que certos tipos de atributos sejam representados por formatos e fontes de dados diferentes. O analista precisa assegurar que o formato dos dados seja consistente com os formatos assumidos pela ferramenta de mineração;

- Limpeza de caracteres: Alguns caracteres especiais são interpretados de forma errônea ou simplesmente não são aceitos por certas ferramentas. O caso típico é o caractere "$" em valores monetários que devem ser interpretados como números. Estes caracteres devem ser filtrados do fluxo de entrada antes que os dados sejam minerados;

- Limpeza de dados: Freqüentemente a fonte de dados possui campos com valores faltando, valores que não são de interesse, ou simplesmente valores errados.

Estes campos devem ser tratados pelo analista. Ele pode fazer interpolações, entrar códigos especiais nestes campos (por exemplo, não se aplica), ou simplesmente

eliminar os registros com estes campos. A ação deve considerar o tipo de dados e seu impacto no processo de mineração;

- Redução do conjunto de dados: Alguns conjuntos de dados podem ser grandes demais para certos algoritmos de mineração. Neste caso o analista pode repartir o conjunto de dados em conjuntos menores e mais específicos de dados, ou o analista pode criar uma amostra do conjunto de dados antes de minerá-lo.

Ainda no pré-processamento realiza-se a fase de transformação dos dados. Esta operação é freqüentemente necessária para adaptar dados para certas técnicas de mineração de dados. Algumas técnicas, como redes neurais, trabalham essencialmente com valores numéricos. Neste caso, atributos com valores categóricos precisam ser mapeados para números. Algumas das operações mais comuns de transformação de dados são:

- Redução de escala: Alguns algoritmos lidam apenas com escalas nominais (categóricas) ou ordinais. Neste caso, os analistas têm que mapear campos numéricos em campos categóricos;

- Extensão da escala: Outros algoritmos trabalham apenas com escalas numéricas. É necessário então transformar escalas categóricas em campos numéricos.

Uma abordagem comum é transformar um atributo categórico em um conjunto de atributos numéricos;

- Conversão de unidades: É comum que diferentes fontes de dados represente o mesmo atributo em diferentes escalas. O analista deve assegurar que atributos deste tipo são gravados consistentemente no repositório e unidades heterogêneas devem ser convertidas para uma unidade comum, geralmente a que é utilizada localmente. Esta tarefa aparentemente simples pode se tornar complexa;

- Normalização de valores: Algumas técnicas de mineração requerem que valores sejam normalizados para um certo intervalo, geralmente de 0 a 1. O valor mínimo é mapeado para 0 e o valor máximo para 1. Todos os valores no intervalo são então mapeados para o intervalo normalizado. Normalmente é feito com valores numéricos, mas pode ser feito com valores categóricos;

- Adaptação de conjunto de dados: Alguns conjuntos de dados desbalanceados podem afetar os resultados de alguns algoritmos de mineração de dados. O correto será estabelecer a combinação de conjuntos de registro em um conjunto equivalente que representa vários registros de uma só vez.

c) Mineração de Dados

Mineração de dados é a principal etapa do processo KDD, sua finalidade é extrair padrões dos dados. Esta fase é considerada o centro do processo e se preocupa em ajustar modelos ou determinar padrões a partir dos dados observados. Também pode ser vista como uma forma de selecionar, explorar e modelar grandes conjuntos de dados para detectar padrões de comportamento [FAY96].

Nesta etapa, é escolhido o método e são definidos os algoritmos que realizarão a busca pelo conhecimento implícito e útil do banco de dados. Ela utiliza técnicas baseadas em análise estatística e Inteligência Artificial (IA), mais especificamente Aprendizagem de Máquina. Esses relacionamentos representam conhecimento valioso sobre a base de dados e, conseqüentemente, sobre o domínio do mundo real que elas representam [HOL91].

d) Pós-Processamento

Nesta fase são analisados os resultados obtidos na fase de mineração de dados e se observará a necessidade ou não de retornar a qualquer fase anterior. Os padrões identificados, depois de transformados em conhecimento, serão utilizados para explicar os fenômenos observados e para apoiar decisões humanas.

Na etapa de apresentação do conhecimento, as técnicas de visualização e representação de conhecimento são usadas com a finalidade de apresentar ao analista, de forma clara, o conhecimento minerado.

Em geral, a principal meta dessa fase é melhorar a compreensão do conhecimento descoberto, validando-o através de medidas da qualidade da solução e da percepção de um analista de dados. Esses conhecimentos serão consolidados em forma de relatórios demonstrativos com a documentação e explicação das informações relevantes ocorridas em cada etapa do processo de KDD.

Esta apresentação das atividades pode sugerir que exista uma trajetória linear e seqüencial das etapas do processo de KDD. No entanto, isso geralmente não se verifica, uma vez que em cada etapa pode ser identificada a necessidade de retorno para cada uma das etapas anteriores, pois existe uma grande iteratividade e interatividade entre estas fases.

2.2.2 Tarefas do Processo de Descoberta do Conhecimento

Entre as principais tarefas resolvidas pelo processo de KDD estão a associação, a classificação e o agrupamento.

● **Associação:** A tarefa de associação tem como principal objetivo encontrar, a partir de um conjunto de exemplos, um conjunto de regras de associação, ou seja, descobrir quais atributos aparecem freqüentemente associados nesses exemplos.

Esse tipo de tarefa é normalmente aplicado a um tipo especial de dados, denominado “cesta de mercado” (*basket data*), em que cada registro consiste de um conjunto de atributos denominados itens, geralmente binários. Cada tupla corresponde a uma transação, e um item pode assumir um valor verdadeiro ou falso, dependendo se ele está ou não presente na transação.

● **Classificação:** A classificação é uma das tarefas mais comumente resolvidas com técnicas de mineração de dados. Um sistema de classificação é utilizado para prever a classe de um objeto baseado em seus atributos.

Os dados utilizados para resolução desse tipo de tarefa consistem em um conjunto de atributos denominados previsores e um atributo denominado meta, que define a classe a que esse registro pertence. O objetivo dessa tarefa é descobrir um relacionamento entre os atributos previsores e o atributo meta, usando registros cuja classe é conhecida, para que posteriormente esses atributos previsores possam ser utilizados para prever a classe de um registro cuja classe é desconhecida.

Quando trabalha-se com um classificador, os exemplos disponíveis para geração de um modelo de classificação são divididos em dois conjuntos mutuamente exclusivos: um conjunto de treinamento e um conjunto de teste. O conjunto de treinamento fica disponível para o classificador, que analisa as relações entre os atributos previsores e o atributo meta. Os relacionamentos descobertos, a partir desses exemplos, são então utilizados para prever a classe dos registros presentes no conjunto de teste. Para o classificador, o atributo meta do conjunto de teste fica indisponível. Após prever a classe dos exemplos do conjunto de teste, as classes previstas são então comparadas com as classes reais dos exemplos, definidas pelo atributo meta. Se a classe prevista for igual a real, a previsão foi correta, caso contrário a previsão foi incorreta.

Um dos principais objetivos na tarefa de classificação é maximizar a taxa de classificações corretas nos dados de teste, que corresponde à razão entre o número de exemplos corretamente classificados e o número total de exemplos disponíveis no conjunto de testes.

● **Agrupamento:** A tarefa de agrupamento divide os dados em grupos formados por elementos com características semelhantes. Nesse tipo de problema, o sistema deve particionar o conjunto de dados em subconjuntos. Um algoritmo de agrupamento deve ser capaz de maximizar a semelhança entre os elementos de um mesmo grupo e minimizar as semelhanças entre exemplos pertencentes a grupos diferentes.

Normalmente não existe uma resposta correta para um problema de agrupamento. A tarefa de agrupamento possibilita um entendimento inicial dos dados, e na maioria dos casos, após o agrupamento, métodos de classificação ou sumarização são aplicados a fim de obter regras de classificação (que distinguem registros pertencentes a classes diferentes) ou regras de sumarização (que caracterizem cada grupo/classe) [FRE02].

2.2.3 Algoritmos de Classificação

Vários são os algoritmos que podem ser aplicados em bases de dados formadas por exemplos e estes formados por um conjunto de atributos de entrada e um conjunto de atributos de saída (classe).

Vários são os paradigmas de aprendizagem e alguns deles são descritos a seguir.

2.2.3.1 Classificação Baseada no Teorema de Bayes

Os Métodos de Aprendizagem Bayesiana são relevantes para a aprendizagem de máquina, pois há um grande interesse em determinar a melhor hipótese sobre um conjunto de instâncias, a partir dos dados observados [MIT97].

De acordo com Mitchell as quantidades de cada classe de interesse são governadas por uma distribuição probabilística e as decisões para se classificar otimamente podem ser tomadas levando-se em consideração estas probabilidades juntamente com os dados [MIT97].

Verifica-se, então, que o conhecimento *a priori* é necessário para o desenvolvimento deste método e, para cada uma das possíveis hipóteses, pode ser associada uma probabilidade *a priori*, possibilitando, assim, o suporte a mais de uma hipótese através de pesos [GEO95].

O Teorema de *Bayes* permite calcular a melhor hipótese baseada em probabilidades *a priori*. Seja a hipótese h e o conjunto de treinamento D. Entende-se $P(h)$ como sendo a probabilidade inicial de uma hipótese h acontecer antes de se observar qualquer conjunto de treinamento, também conhecida por probabilidade *a priori*. Caso não haja esta informação, admite-se que cada uma das possíveis hipóteses possui a mesma probabilidade. De forma análoga, $P(D)$ é a probabilidade *a priori* do conjunto de treinamento antes de se admitir alguma hipótese para este conjunto. Já $P(D \mid h)$ significa a probabilidade de se observar o conjunto de treinamento admitindo-se a hipótese h, ou seja, é a probabilidade de D dado h.

Nos problemas de Aprendizagem de Máquina, o foco é na $P(h \mid D)$, ou seja, na probabilidade *posteriori* de h dado o conjunto de treinamento D. Ela mede a influência do conjunto de treinamento em contraste com a probabilidade *a priori* [MIT97].

O Teorema de *Bayes* é a base para métodos de aprendizagem bayesiana porque ele proporciona uma maneira para calcular a probabilidade *posteriori* $P(h \mid D)$, da probabilidade a *priori* $P(h)$, juntamente com $P(D)$ e $P(D \mid h)$.A segui, pode-se observar a fórmula do Teorema de *Bayes*.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{1}$$

Algoritmo *Naïve Bayes*

Dentre os métodos de Aprendizado Bayesiana, existe um conhecido por *Naïve Bayes* ou Classificador Bayesiano Ingênuo que, em alguns casos, pode apresentar bons resultados de desempenho, se comparado com outros algoritmos como redes neurais e aprendizagem por árvores de decisão, principalmente quando combinado com alguns métodos de seleção de atributos para a eliminação de redundância de informação. A vantagem deste classificador advém da simplicidade no seu cálculo, pois admite, ingenuamente, independência entre atributos resultando na busca pela classificação que maximiza o produtório da sua fórmula [MIT97]. Todavia, pode ser que haja casos em que esta suposição não retrate a realidade, prejudicando a análise final.

A fórmula do Teorema de Bayes para o Classificador *Naïve Bayes* é:

$$h_{NB} = \arg\max_{h_j \in H} P(h_j) * \prod_i P(a_i \mid h_j) \tag{2}$$

onde a hipótese de *Naïve Bayes* h_{NB} é a que maximiza o valor de um produto entre a probabilidade de ocorrência de uma hipótese h_j (uma entre as possíveis no conjunto de hipóteses H) e um produtório de probabilidades das valorações dos *i*-ésimos atributos dada a hipótese h_j.

Um problema que pode ser solucionado por este classificador é quando a combinação de uma possível valoração de um atributo com uma certa classe não ocorre, ou seja, o contador é igual a zero. Isto faz com que esta probabilidade seja zero e, conseqüentemente, o h_{NB} também o seja. Uma solução para este problema é utilizar o estimador Laplaciano, o qual consiste em iniciar o contador de cada valoração possível de um atributo com o número 1.

Um caso especial do problema citado é quando existem instâncias cuja valoração do atributo é inexistente. Neste caso, o contador de freqüência não é incrementado e a probabilidade se baseia no número de instâncias valoradas ao invés de se basear no número total de instâncias [MIT97].

2.2.3.2 Classificação Baseada em Árvore de Decisão

A aprendizagem por árvore de decisão é um dos métodos mais usados para inferência indutiva. É um método que consegue fazer aproximações de funções com valores discretos e tem a vantagem de ser robusto para manipular dados com ruídos ser capaz de aprender expressões disjuntas [RUS04].

Este modelo classifica as instâncias percorrendo uma árvore a partir do nó raiz até alcançar uma folha. Cada um dos nós testa o valor de um único atributo e, para cada uma de suas valorações, oferece arestas diferentes a serem percorridas na árvore a partir deste nó. Sua vantagem é a estratégia adotada conhecida por "dividir-para-conquistar" o qual divide um problema maior em problemas menores. Assim, sua capacidade de discriminação dos dados provém da divisão do espaço definido pelos atributos em subespaços [MIT97].

Vários algoritmos implementam árvores de decisão entre eles o ID3 [QUI86] e o C4.5 [QUI93].

Algoritmo ID3

O ID3 é um algoritmo recursivo de busca gulosa que procura sobre um conjunto de atributos, aqueles que "melhor" se encaixam nas raízes das sub-árvores a serem construídas. Inicialmente, todos os atributos, menos o classificatório, são reunidos em um conjunto. Em seguida, o "melhor" atributo é escolhido e passa a ser a raiz da sub-árvore em construção. Para cada possível valoração deste atributo, é criada uma aresta até as futuras sub-árvores obtidas com a recursividade deste algoritmo. Os dois únicos critérios de parada são quando não há mais instâncias ou atributos a serem analisados [QUI86].

Dentre um conjunto de atributos, o "melhor", para ser um nó raiz de uma sub-árvore, é aquele que gera a menor sub-árvore cujas folhas são as mais puras possíveis, ou seja, tendem a possuir instâncias de uma única classe. A função utilizada para esta medição é o "Ganho de Informação" [QUI86].

Algoritmo C4.5

O algoritmos C4.5 é uma melhora do ID3, ou seja, além de possuir as mesmas características, ele possui a vantagem de poder lidar com a poda (*prunning*) da árvore

para evitar o sobre-ajustamento, com a ausência de valores, com a valoração numérica de atributos e com a presença de ruídos nos dados [QUI93].

Ao contrário do algoritmo que o originou, que manipula apenas dados nominais, o C4.5 pode manipular também dados numéricos. Contudo, lidar com este tipo de dado não é tão simples, até porque atributos nominais são testados uma única vez em qualquer caminho da raiz até as folhas, enquanto que atributos numéricos podem ser testados mais de uma vez no mesmo percurso caracterizando uma possível desvantagem do C4.5 pois, em alguns casos, pode tornar a árvore difícil de se entender [QUI93].

2.2.3.3 Classificação Baseada na Teoria Estatística de Aprendizagem

Algoritmo SVM

O algoritmo SVM (Máquinas de Vetores Suporte do inglês *Support Vector Machines*) utiliza conceitos de modelos lineares e aprendizagem baseada em instâncias e tem se mostrado eficiente comparado com a grande maioria dos classificadores, em diversas aplicações, entre elas aplicações em Bioinformática [SOU03]. Baseia-se no fato que em altas dimensões do espaço de atributos, todos os problemas tendem a se tornar linearmente separável.

Esses resultados são alcançados pelo emprego dos conceitos da Teoria de Aprendizado Estatístico, introduzida por Vapnik [VAP95] que apresenta diversos limites na capacidade de generalização de um classificador linear. Para tal, dado um conjunto de treinamento E com n pares (x_i, y_i), em que $x_i \in \Re^m$ e $y_i \in \{-1,+1\}$, as SVMs buscam o classificador linear $g(x) = \text{sgn}(W + x + b)$ capaz de separar os dados pertencentes a E com erro mínimo e a margem p máxima de separação entre as classes presentes em E (Figura 6), onde margem é a distância de dois elementos de classes diferentes.

Dada uma função linear $f(x) = W * x + b$, a margem $p(x_i, y_i)$ utilizada para classificar um padrão x_i é fornecida por $y_i f(x_i)$. Ela mede a distância do padrão x_i em relação ao hiperplano separador.

Maximizar p equivale a minimizar a norma de $\|W\|$ [HEA98]. Logo, pode-se manter p fixo e buscar um hiperplano com $\|W\|$ pequeno tal que não haja exemplos de treinamento com margem menor que p [SMO02].

Este é um problema clássico em otimização que pode ser solucionado usando a programação quadrática, para o qual há uma ampla e estabelecida teoria [HEA98]. Pode-se verificar que a determinação do classificador final se dá unicamente em função de

padrões denominados vetores suporte (SVs, - do inglês *Support Vectors*). Esses padrões correspondem aos exemplos de treinamento mais próximo ao hiperplano separador e são considerados os dados mais informativos do conjunto de treinamento [SOU03].

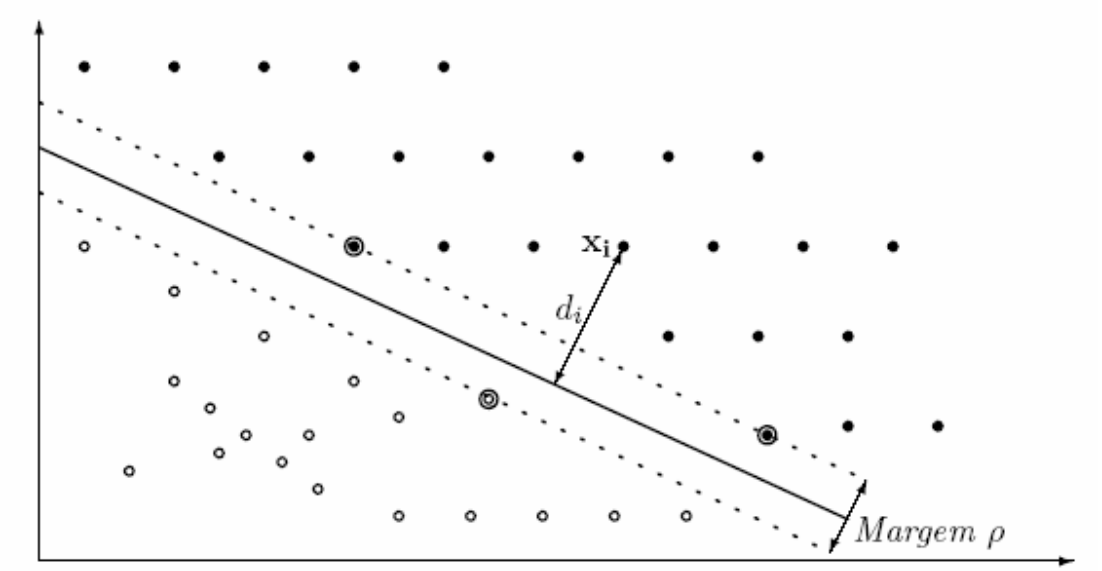

Figura 6: Margem geométrica de um ponto x_i e a margem p do hiperplano de separação ótimo. Os círculos fechados são os exemplos positivos e os círculos abertos são os exemplos negativos. Os círculos que caem sobre as margens (linhas tracejadas) são os vetores suporte para esse conjunto de treinamento. Os vetores suporte são realçados com um círculo mais externo [LIM02].

Há muitos casos em que não é possível dividir satisfatoriamente os dados de treinamento por um hiperplano e uma fronteira não linear é mais adequada.

Para generalizar as SVMs para lidar com essas situações, mapeia-se cada padrão do conjunto de treinamento E para um novo espaço, denominado espaço de características. Uma característica singular desse espaço é que a escolha de uma função de mapeamento ϕ apropriada torna o conjunto de treinamento mapeado linearmente separável.

As SVMs lineares podem então ser utilizadas sobre o conjunto de treinamento mapeado no espaço de características [CRI00]. Para isto, basta aplicar a função de mapeamento ϕ a cada padrão nas Equações listadas para o caso linear.

Por meio desse procedimento, percebe-se que a única informação necessária sobre o mapeamento é uma definição de como o produto interno $\phi(x_i) * \phi(x_j)$ pode ser calculado. Isto é obtido com a introdução do conceito de *Kernel*s, funções que recebem

dois pontos x_i e x_j do espaço de entradas e computam o produto escalar $\phi(x_i)*\phi(x_j)$ no espaço de características [HAY99]. De maneira geral, a função *Kernel* é mais simples que a do mapeamento ϕ. Por este motivo, é comum defini-la sem conhecer-se explicitamente o mapeamento ϕ. Alguns dos *Kernels* mais utilizados são os Polinomiais, os Gaussianos e os Sigmoidais [SOU03].

A escolha da função *Kernel* e seus parâmetros, assim como da constante que impõe um peso diferente para o treinamento em relação a generalização no problema de otimização, têm influência direta no desempenho do classificador gerado por uma SVM [MÜL01]. Essa sensibilidade a escolhas de parâmetros representa uma das deficiências das SVMs. Outra deficiência diz respeito à dificuldade de interpretação do modelo gerado, como no caso das Redes Neurais.

Deve-se destacar também que as SVMs realizam originalmente classificações binárias. Diversas aplicações em Bioinformática, porém, envolvem mais de duas classes. Existem diversos métodos para generalizar as SVMs a problemas multiclasses [SOU03]. Duas abordagens usuais para tal são as decomposições “um-contra-todos” e “todos-contra-todos” [SMO02].

2.2.3.4 Classificação Baseada em Instâncias

Em contraste com os demais métodos de aprendizagem, que criam modelos sobre o conjunto de treinamento para classificar as novas instâncias, a aprendizagem baseada em instância atribui uma classificação a cada elemento do conjunto de treinamento e os armazena para poder classificar as novas instâncias [MIT97]. A generalização oferecida por outros modelos é feita sob demanda neste método, ou seja, à medida que chegam novas instâncias. Eis o porquê de ser denominado método preguiçoso de aprendizagem e da demora no processamento fazendo com que a nova instância deva ser comparada a todas as instâncias já classificadas.

Algoritmo *k*-NN

Um dos algoritmos de aprendizagem baseada em instância é o algoritmo *k*-NN (*K-Nearest Neighbour* ou *k*-vizinhos mais próximos), onde as instâncias são agrupadas conforme a maior proximidade entre elas. O *k*-NN assume que todas as instâncias correspondem a pontos no espaço n-dimensional. A função que mede a distância entre as instâncias é de suma importância para este algoritmo. Um exemplo de função que é muito utilizada, quando o atributo é do tipo numérico, é a distância Euclidiana [MIT97].

Vamos considerar uma instância arbitrária x que é descrita pelo vetor de atributos (*feature vector*): $< a_1(x), a_2(x), ..., a_n(x) >$, onde $a_r(x)$ representa o valor do r-ésimo atributo da instância x.

Então a distância entre duas instâncias x_i e x_j é definida como $d(x_i, x_j)$, onde:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} [a_r(x_i) - a_r(x_j)]^2} \quad (3)$$

2.3 Técnicas de Redução de Dimensionalidade

Existem várias métodos para a redução de dimensionalidade. Neste trabalho serão discutidos e utilizados dois métodos: a seleção de atributos e a projeção aleatória.

2.3.1 Seleção de Atributos

A mineração de dados é hoje o termo associado à busca de conhecimento compreensível, útil e surpreendente em grandes bases de dados, e sua aplicação dispensa a presença de um número significativo de atributos ou mesmo registros presentes nas bases de dados originais, e que em certos casos, se não forem removidos, podem até “atrapalhar” o processo de aprendizagem.

Dada a importância de se reduzir o espaço de dados e atributos, pesquisas na área de seleção de atributos foram iniciadas há muito tempo nas áreas de estatística e reconhecimento de padrões, e só posteriormente passaram a ser tratadas na área de aprendizagem de máquina. Porém, a solução para esse problema não é trivial nem única.

A seleção de atributos tem como objetivo descobrir um subconjunto de atributos relevantes para uma tarefa alvo, considerando os atributos originais, e é importante, entre outras coisas, por tornar o processo de aprendizagem mais eficiente. Atributos redundantes prejudicam o desempenho do algoritmo de aprendizagem tanto na velocidade (devido à dimensionalidade dos dados) quanto na taxa de acerto (devido à presença de informações redundantes que podem confundir o algoritmo, ao invés de auxiliá-lo na busca de um modelo correto para o conhecimento) [KIR92].

Quando a tarefa alvo da seleção de atributos é a classificação, a seleção de atributos normalmente busca minimizar a taxa de erro do classificador, a complexidade do conhecimento gerado por ele, e o número de atributos selecionados para compor a “nova” base.

Em bases de dados de expressão gênica, por exemplo, em bases de câncer, supõe-se que existam muitos dados redundantes e irrelevantes do ponto de vista da

aprendizagem de máquina e até mesmo para áreas como biomédicas, onde busca-se descobrir possíveis causadores do câncer.

A seleção de atributos é um processo em que se escolhe um subconjunto de M características do conjunto original de N características $(M \leq N)$, de modo que o espaço de características seja reduzido de acordo com um determinado critério [LIU03]. Dessa forma, a seleção de atributos garante que os dados que chegam à fase de mineração sejam de boa qualidade [LIU98].

De acordo com Dash e Liu, o método de seleção de atributos consiste de 4 passos principais como mostrado na figura 7 [DAS97].

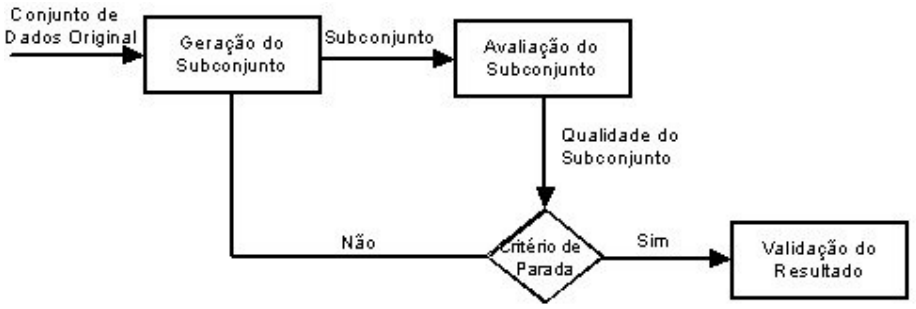

Figura 7: Passos básicos do processo de seleção de atributos [DAS97].

A geração do subconjunto é um procedimento de busca que produz subconjuntos de atributos candidatos para a avaliação baseada em uma estratégia de busca. Cada subconjunto de atributos candidatos é avaliado e comparado com um melhor anterior, segundo um critério de avaliação. Se o novo conjunto se tornar o melhor, ele substitui o melhor anterior. O processo de geração e avaliação de subconjuntos é repetido até satisfazer um dado critério de parada. Então, o melhor subconjunto selecionado usualmente necessita ser avaliado por um conhecimento a priori ou através de diferentes testes em conjuntos de dados reais ou sintéticos [LIU05].

Os algoritmos de seleção de atributos foram desenvolvidos com diferentes critérios de avaliação. A duas principais abordagens são: abordagem filtro e a abordagem *wrapper*. O modelo filtro depende de características gerais dos dados para avaliar e selecionar subconjuntos de características sem envolvimento de um algoritmo de mineração. O modelo *wrapper* requer um algoritmo de mineração pré-determinado e utiliza seu desempenho como critério de avaliação. Ele procura pelo melhor conjunto de características para melhorar o desempenho do algoritmo de mineração, mas tende a ser computacionalmente mais custoso se comparado ao modelo de filtro.

2.3.1.1 Procedimento Geral para a Seleção de Atributos

Geração do Subconjunto

A natureza deste processo é determinada por dois tópicos básicos. Primeiro decide-se o ponto (ou pontos) inicial da busca que passam a influenciar a direção da busca. A busca pode iniciar com um conjunto vazio e sucessivamente adicionar características (para frente), ou iniciar com um conjunto completo e remover sucessivamente as características (para trás), ou iniciar com ambos e adicionar e remover características simultaneamente (bidirecional). A busca também pode ser iniciada com um subconjunto selecionado aleatoriamente para evitar a armadilha de um ótimo local. A seguir, decide-se a estratégia de busca. Para um conjunto de dados com N atributos, existem 2^N subconjuntos candidatos. Este espaço de busca é exponencialmente proibitivo para a busca exaustiva mesmo com um número de atributos moderado. Diferentes estratégias de busca têm sido exploradas: a busca exponencial, a seqüencial e a aleatória [LIU05].

- **Busca Exponencial**

Garante encontrar o resultado ótimo segundo o critério de avaliação utilizado. Tal como a busca exaustiva, faz todas as combinações possíveis antes de retornar um subconjunto de características. A busca exaustiva é completa. No entanto, uma busca não precisa ser exaustiva para garantir que seja completa [LIU05]. Diferentes funções heurísticas podem ser utilizadas para reduzir o espaço de busca sem colocar em risco as chances de encontrar um resultado ótimo. Desta forma, embora a ordem do espaço de busca seja $O(2^N)$, somente um pequeno número de subconjuntos são avaliados.

Alguns exemplos clássicos que utilizam o conceito de busca completa são a busca em largura e a busca em profundidade [LIU98].

- **Busca Seqüencial**

Os algoritmos seqüenciais, como a seleção seqüencial para frente e seqüencial para trás, são eficientes na resolução de problemas de seleção de atributos, mas têm a desvantagem de não levar em conta a interação entre atributos.

A seleção seqüencial para frente inicia a busca pelo melhor subconjunto de atributos com um conjunto vazio de atributos. Inicialmente, subconjuntos de atributos com apenas um atributo são avaliados, e o melhor atributo A^* é selecionado. Esse atributo A^* é então combinado com todos os atributos disponíveis (em pares), e o melhor subconjunto de atributos é selecionado. A busca continua dessa mesma forma, sempre adicionando um atributo por vez ao melhor subconjunto de atributos anteriormente

selecionado, até que não se consiga mais melhorar a qualidade do subconjunto de atributos selecionados.

A seleção seqüencial para trás, ao contrário da seleção seqüencial para frente, inicia a busca por um subconjunto de atributos ótimos com uma solução representando todos os atributos, e a cada iteração um atributo é removido da solução atual, até que não se consiga melhorar a qualidade da solução encontrada.

● **Busca Aleatória**

Inicia com um subconjunto selecionado aleatoriamente e procede de duas formas diferentes. Uma, é seguir a busca seqüencial, que insere aleatoriedade nas abordagens seqüenciais clássicas. Alguns exemplos são subida da montanha (*hill-climbing*) com inicio aleatório e recozimento simulado (*simulated annealing*) [KIR83]. A outra é gerar o próximo subconjunto de maneira completamente aleatória (por exemplo, um subconjunto atual não aumenta ou diminui baseado em qualquer subconjunto anterior seguindo uma regra determinística), também conhecida como algoritmo simulação de monte carlo, como a proposta por Metropolis com seus companheiros. [MET53]. Para todas essas abordagens a utilização de aleatoriedade pode ajudar a escapar de um ótimo local no espaço de busca, e a otimização do subconjunto selecionado depende dos recursos disponíveis.

Cinco operadores podem ser considerados para gerar um estado sucessor: Para Frente, Para Trás, Composta, Aleatória e Ponderada [MOL02]. Todos esses operadores modificam de algum modo o peso w_i do atributo x_i, com $w_i \in \Re$ ou $w_i \in \{0,1\}$. Assumiremos que J é a medida de qualidade do subconjunto a qual será maximizada.

● **Para Frente:** esse operador adiciona atributos à solução atual X', entre os que ainda não foram selecionados. A cada passo, o atributo que possui uma medida de avaliação maior é adicionado à solução. Iniciando com $X' = 0$, o passo seguinte do operador consiste em [MOL02]:

$$X' := X' \cup \{x_i \in X \setminus X' \mid J\left(X' \cup \{x_i\}\right) \acute{e}\ maior\} \tag{4}$$

O critério de parada pode ser: $\left|X'\right| = p'$ (se p' foi fixada com antecedência), o valor de J não é incrementado no último passo j, pois o valor de J_0 é excedido. O custo do operador é $O(n)$.

● **Para Trás:** esse operador remove atributos da solução atual X', entre os atributos que ainda não foram removidos. A cada passo, o atributo que possui uma medida de avaliação melhor é removido da solução. Iniciando com $X' = X$, o passo seguinte do operador consiste em [MOL02]:

$$X' := X' \setminus \{x_i \in X' \mid J(X' \setminus \{x_i\}) \text{ é } maior\} \quad (5)$$

O critério de parada pode ser $|X'| = p'$, o valor de J não é incrementado no último passo j, pois o valor de J_0 é reduzido. O custo do operador é $O(n)$, embora na prática exija mais recurso computacional que o operador para frente.

● **Composta:** a idéia desse operador é simples. Aplica-se f sucessivos passos para frente e b sucessivos passos para trás. Se $f > b$ o resultado é para frente, caso contrário é para trás.

● **Aleatória:** esse grupo inclui os operadores que podem, potencionalmente, gerar outros estados em um único passo. Podem ter componentes aleatórios, porém, são restritos a alguns critérios como o número de atributos e a qualidade da medida de avaliação J a cada passo.

● **Ponderada:** nesse operador o espaço de busca é contínuo, e todos os atributos estão presentes na solução. Um estado sucessor é um estado com pesos diferentes.

Avaliação do Subconjunto

Cada novo subconjunto gerado precisa ser avaliado pelo critério de avaliação. A qualidade de um subconjunto pode ser determinada por algum critério. Um critério de avaliação pode ser, de maneira geral, categorizado dentro de dois tipos de abordagens, a abordagem filtro e a abordagem *wrapper,* discutidas a seguir.

✦ **Abordagem Filtro**

Essa abordagem tenta escolher um subconjunto de atributos independente do algoritmo de classificação, estimando a qualidade dos atributos apenas em relação aos dados. A Figura 8 mostra a seleção de atributos com a abordagem filtro, que faz a seleção usando uma etapa de pré-processamento, baseada nos dados de treinamento [BAL96].

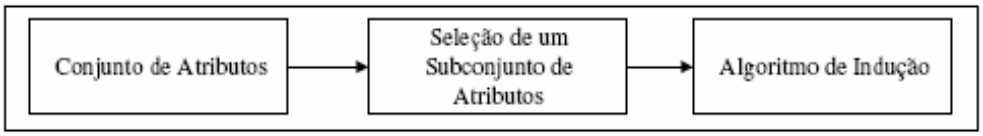

Figura 8: Seleção de Atributos utilizando abordagem Filtro.

Há várias técnicas para avaliar um subconjunto de atributos $J(X')$. Entre as medidas de avaliação destacamos a Medida de Distância, de Informação, de Dependência e de Consistência [LIU05].

● **Medidas de Distância:** são também conhecidas como separabilidade, divergência ou medidas de discriminação. Para o problema de duas classes, um atributo X é preferido quando comparado com outro atributo Y, se X induz uma diferença maior entre a probabilidade condicional de duas classes que Y. Sendo que X e Y são indistinguíveis se a diferença for zero [MOL02].

É suficiente definir uma métrica entre classes e usar como medida:

$$D(\omega_i, \omega_j) = \frac{1}{N_i j_j} \sum_{k_1}^{Ni} \sum_{k_2=k_1+1}^{Ni} d\left(x_{(i,k_1)}, x_{(j,k_2)}\right) \tag{6}$$

$$J = \sum_{i=1}^{m} P(\omega_i) \sum_{j=i+1}^{m} P(\omega_j) D(\omega_i, \omega_j) \tag{7}$$

onde $x(_{i,j})$ é a instância j da classe ω_i, e N_i é o número de instâncias da classe ω_i. A distância mais utilizada D pertence à família Euclidiana. Essas medidas não requerem modelagem de qualquer função de densidade, mas a relação delas para a probabilidade do erro pode ser muito próxima.

● **Medidas de Informação:** tipicamente determina o ganho de informação de um atributo. O ganho da informação de um atributo X é definido como a diferença entre a incerteza anterior e a posterior utilizando X. O atributo X é preferido, ao invés do atributo Y, se o ganho de informação de X for maior que o ganho de informação Y [LIU05].

Observando um exemplo de treinamento e, sem a sua classe associada, podemos computar a probabilidade a posteriori $P(\omega_i \mid e)$ para determinar o ganho de informação da classe e, no que diz respeito a probabilidade *a priori*. Se todas as classes se tornam aproximadamente iguais, então o ganho de informação é mínimo e a incerteza (entropia) é máxima [MOL02].

Muitas medidas podem ser derivadas para fazer uso de $p(e)$ e do conjunto $\{P(\omega_i \mid e) \setminus i = 1,...,n\}$. Por exemplo, usando a entropia de Shannon, (J_{Sha}), tem-se:

$$J_{Sha} = -\int p(e) \sum_{i=1}^{m} P(\omega_i \mid e) \log_2 P(\omega_i \mid e) de \tag{8}$$

As medidas derivam da generalização de entropia de Shannon. A entropia pode ser usada sem conhecimento de densidades como acontece na indução de árvores de decisão, onde o ganho de informação é tipicamente computado independentemente para cada atributo.

● **Medidas de Dependência:** são também conhecidas como medidas de correlação ou medidas de similaridade. Essa medida tem a habilidade de predizer o valor de uma variável de um valor para outro. Na seleção de atributos para classificação, o

atributo melhor avaliado é aquele que prediz melhor a classe. Um atributo X é preferido, ao invés do atributo Y, se a associação entre X e a classe C for mais alta que a associação da característica entre Y e C.

Um dos algoritmos que utiliza essa medida é o CFS (*Correlation-basead Feature Selection*) [HAL99]. Esse algoritmo avalia a importância de um subconjunto de atributos baseado na habilidade preditiva individual de cada atributo e o grau de correlação entre eles. A equação de mérito, neste caso, é dada por:

$$Merit_S = \frac{k\,\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \tag{9}$$

em que $Merit_S$ é o "mérito" do subconjunto S de atributos contendo k atributos, $\bar{r}_{cf}$ é a média de atributos para a correlação da classe $(f \in S)$, e $\bar{r}_{ff}$ é a média de atributos para a correlação do atributo.

A diferença entre os algoritmos de filtro normal e o CFS é que enquanto os algoritmos de filtro normal fornecem um escore para cada atributo independentemente, o CFS apresenta uma recompensa heurística do subconjunto de atributos e informa o melhor subconjunto encontrado.

- **Medidas de Consistência:** é diferente da medida anterior devido à forte dependência em relação a informação da classe e a utilização do bias para a seleção de um subconjunto com poucos atributos [LIU96]. Essa medida tenta encontrar um número mínimo de atributos que separa a classe tão consistentemente quanto o conjunto inteiro de atributos pode separar. Uma inconsistência é definida quando duas instâncias são iguais, ou seja, possuem os mesmos atributos, mas com rótulos diferentes para a classe. Como exemplo, uma inconsistência em X' e S é definida com duas instâncias em S que são iguais quando considerado somente a característica X' e que pertence a classes diferentes [MOL02]. O propósito é encontrar um subconjunto mínimo de atributos que conduzem à inexistência de inconsistências. A quantidade de inconsistência de uma instância $A \in S$ é definida como:

$$ICx'(A) = X'(A) - \max_k X'_k(A) \tag{10}$$

onde $X'(A)$ é o número de instâncias em S igual a A usando somente os atributos em X', e $X'_k(A)$ é o número de instâncias em S da classe k igual a A usando somente atributos em X'. A taxa de inconsistência da amostra do subconjunto de atributos S é:

$$IR(X') = \frac{\sum_{A \in S} IC_X\ (A)}{|S|} \tag{11}$$

Esta é uma medida monotônica: $X_1 \subset X_2 \Rightarrow IR(X_1) \geq IR(X_2)$

Uma possível medida de avaliação seria:

$$J(X') = \frac{1}{IR(X') + 1} \tag{12}$$

Esta medida está contida na faixa [0,1] e pode ser avaliada em tempo $O(|S|)$ usando uma tabela *hash*.

✦ **Abordagem *Wrapper***

A abordagem *wrapper* define um subconjunto ótimo de soluções de acordo com uma base de dados e algoritmo de indução particular, levando em conta a tendência (bias) indutiva do algoritmo e sua interação com o conjunto de treinamento. A Figura 9 esquematiza um algoritmo de seleção de atributos utilizando a abordagem *wrapper*.

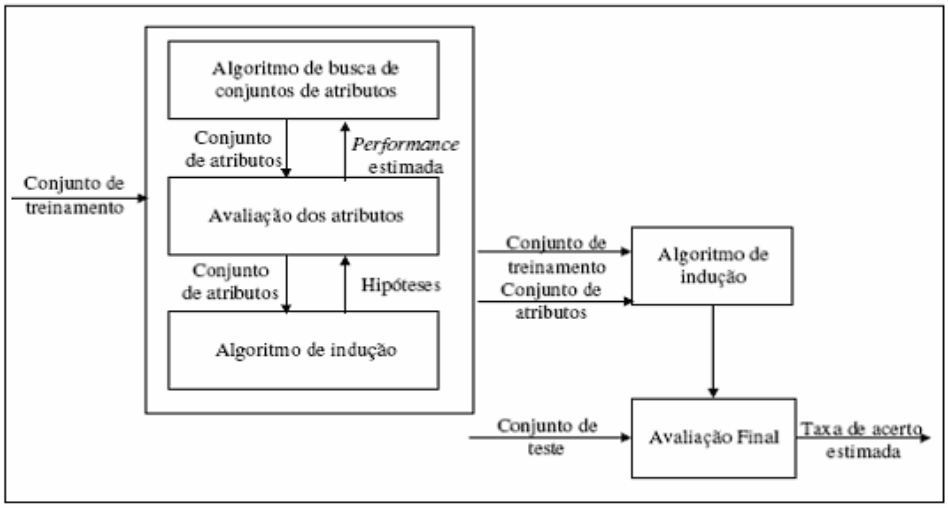

Figura 9: Seleção de Atributos utilizando abordagem *wrapper* [KOH98].

Usualmente esta abordagem revela um desempenho superior quando encontra conjuntos de atributos melhores agrupados para o algoritmo de mineração pré-determinado, mas tende a ser mais custoso computacionalmente e pode não ser adequado para outros algoritmos de mineração. Por exemplo, na tarefa de classificação, a precisão preditiva é amplamente utilizada como uma medida primária. Ela pode ser utilizada como um critério dependente para seleção de características. Como os atributos

são selecionadas por um classificador que posteriormente utilizará estas características selecionadas na predição de classes de instâncias não vistas, a precisão é normalmente alta, mas estimar a precisão de cada subconjunto de característica tem um custo computacional alto.

Critério de Parada

O critério de parada estabelece quando um processo de seleção de atributos deve ser parado. Pode ser feito quando a busca termina, ou quando o objetivo foi alcançado, em que o objetivo pode ser um número especificado (número máximo de atributos ou número máximo das iterações), ou quando um subconjunto suficientemente bom é encontrado (por exemplo, um subconjunto poderia ser suficientemente bom se a razão do erro da classificação é menor que a razão de erro permitida para uma dada tarefa).

Validação dos Resultados

Uma forma para validação do resultado é medir diretamente o resultado utilizando um conhecimento *a priori* dos dados. Se os atributos relevantes são conhecidos previamente é possível comparar o conjunto de atributos conhecido com os atributos selecionados. O conhecimento dos atributos irrelevantes ou redundantes também pode ser útil. Contudo, em uma aplicação real, como no caso onde se trabalha com dados de expressão gênica, não se tem um conhecimento prévio dos dados. Portanto é preciso confiar em alguns métodos indiretos, monitorando a mudança do desempenho da mineração com a mudança de atributos. Por exemplo, utilizando a razão do erro do classificador como um indicador de desempenho, para um subconjunto de atributos selecionado, pode-se conduzir o experimento “antes e depois” para comparar a razão do erro do classificador treinado no conjunto completo de atributos e no subconjunto de atributos selecionado.

2.3.1.2 Algoritmos de Seleção de Atributos

Existe uma vasta quantidade de algoritmos de seleção de algoritmos de características. A tabela 2 mostra as características em comum e suas diferenças baseada nas estratégias de busca e critérios de avaliação de alguns algoritmos de seleção de atributos.

Tabela 2: Características de alguns Algoritmos de Seleção de Atributos

Algoritmo	Geração do Subconjunto	Geração de Sucessores	Medida de Avaliação	Referência
LVF	Aleatória	Aleatória	Consistência	[LIU96]
LVI	Aleatória	Aleatória	Consistência	[LIU98]
CFS	Seqüencial	Para Frente	Dependência	[HAL99]
Relief	Aleatória	Ponderada	Distância	[KIR92]
SBG	Seqüencial	Para Trás	qualquer	[DEV82]
SFG	Seqüencial	Para Frente	qualquer	[DEV82]
Focus	Exponencial	Para Frente	Consistência	[ALM91]

2.3.2 Método de Projeção Aleatória

O método de projeção aleatória surgiu recentemente como um poderoso método para redução de dimensionalidade. O método é mais barato computacionalmente comparado com outros métodos de redução da dimensionalidade.

Em muitas aplicações de mineração de dados, alguns métodos de redução de dimensionalidade são restringidos devido a alta dimensão dos dados. A projeção aleatória é um método que pode ser aplicado em vários tipos de dados como texto, imagem, áudio, entre outros. [LIN03].

Alguns trabalhos utilizando este método já foram feitos e tiveram bons resultados experimentais [BIN01]. Papadimitriou e seus colegas usaram o método de projeção aleatória na fase de pré-processamento em dados textuais antes de usar o LSI (*Latent Semantic Analysis*) [PAP98]. Kaski apresentou resultados experimentais usando a projeção aleatória no sistema WEBSOM [KAS98]. Kurimo aplicou o método de projeção aleatória para indexar documentos auditivos, antes de usar LSI e SOM [KUR99].

2.3.2.1 Descrição do Método

A idéia do método é extremamente simples: dada uma matriz X, a dimensionalidade dos dados pode ser reduzida pela projeção de uma matriz formada por valores aleatórios [LIN03]: $A_{[n\times k]} = X_{[n\times m]} * R_{[m\times k]}$, onde k representa a quantidade de colunas da matriz reduzida.

O método de projeção aleatória é motivada pelo Teorema de Johnson-Lindenstrauss [LIN03]:

Teorema de Johnson-Lindenstrauss:

Para qualquer $0<\varepsilon<1$ e qualquer inteiro n, sendo k um inteiro positivo, tal que: $k \geq 4\left(\varepsilon^2/2-\varepsilon^3/3\right)^{-k} \ln n$.

Então para qualquer conjunto W de n pontos em R^m, há uma função de mapeamento $f: R^m \rightarrow R^k$ tal que para todo $u, v \in W$, $(1-\varepsilon)\| u-v \|^2 \leq \| f(u)-f(v) \|^2 \leq (1+\varepsilon)\| u-v \|^2$.

Analisando o teorema, o desenvolvimento da equação permite deduzir que um conjunto de n pontos em um espaço Euclidiano de alta dimensionalidade pode ser definido como $O\left(\log n/\varepsilon^2\right)$ no subespaço dimensional tal que as distâncias entre os pontos são aproximadamente preservados [LIN03].

Tipicamente, os elementos em R são distribuições Gaussianas, onde uma distribuição Gaussiana é definida por: $G(\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, onde μ é a média e σ é o desvio padrão da distribuição.

Achlioptas propôs duas distribuições [ACH01]:

$$r_{i,j}=\begin{cases}+1 \ com \ \ probabilidade \ \ 1/2\\ -1 \ com \ \ probabilidade \ \ 1/2\end{cases} \tag{13}$$

ou

$$r_{i,j}=\sqrt{3} \ * \begin{cases}+1 \ com \ \ probabilidade \ \ 1/6\\ 0 \ com \ \ probabilidade \ \ 2/3\\ -1 \ com \ \ probabilidade \ \ 1/6\end{cases} \tag{14}$$

Essas distribuições reduzem o tempo computacional para o cálculo de $X * R$ [LIN03]. Com esse método os dados originais de dimensão m são projetados em um subconjunto $k(k << d)$ [BIN01]. Dessa maneira a matriz original $X_{n\times m}$ é projetada pela matriz $R_{m\times k}$ obtendo a matriz reduzida $A_{n\times k}$.

2.4 Trabalhos Relacionados

Existe uma série de estudos que aplicam técnicas de AM não supervisionadas e supervisionadas em dados de expressão gênica. Apesar de neste trabalho não seja aplicada nenhuma das técnicas de aprendizagem de máquina não supervisionada alguns trabalhos pertencente a essa abordagem são relatados.

Quando se utiliza aprendizagem de máquina não supervionada em dados de expressão gênica, estudos investigam diferentes técnicas de análise de dados, medidas de similaridades e base de dados. Entre as técnicas de análise de dados empregadas

nesses estudos estão: redes SOM, agrupamento hierárquico, análise de componentes principais, k-médias e CLICK [TAM99], [EIS98], [WEN98], [GOL99], [ALI00].

Com relação aos dados utilizados, a grande maioria dos trabalhos publicados utiliza níveis de expressão da levedura (*Saccharomyces cerevisiae*) [EIS98], [TAM99]. Uma das razões para essa preferência é que esse organismo possui todos os seus genes conhecidos e com uma boa parte de suas funções descobertas. Existem também estudos com dados de ratos [WEN98] e de humanos [TAM99].

Técnicas de AM não supervisionadas também vêm sendo aplicadas à descoberta de novas classes de doenças. Com o objetivo de analisar dados de tumores, Golub e seus colegas usaram redes SOM com dois aglomerados para agrupar automaticamente 38 exemplos de dois tipos de leucemia bastante relacionadas, com base na expressão gênica de 6817 genes - Leucemia Mielóide Aguda (AML - do inglês *Acute Myeloid Leukemia*) e Leucemia Linfoblástica Aguda (ALL - do inglês *Acute Lymphoblastic Leukemia*). O conjunto de dados era formado por 11 amostras do tipo AML e 27 do tipo ALL. Os resultados obtidos mostram que a rede SOM foi eficiente, embora não tenha obtido uma precisão de 100%, em descobrir automaticamente as duas categorias de leucemia (tumores), conforme apresentado na figura 10(a) [GOL99].

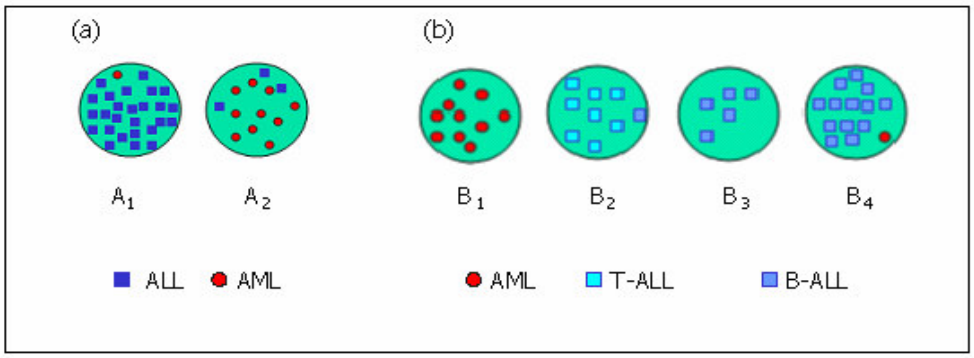

Figura 10: Descoberta da Classe ALL e AML [GOL99].

Nesse mesmo trabalho [GOL99], os autores estenderam a descoberta de classes por meio de uma busca por subclasses mais refinadas. Para isto, foi utilizada uma rede SOM para dividir os exemplos em quatro aglomerados (B1 a B4). Subseqüentemente, obtiveram dados de imuno-fenótipos das amostras, nos quais observaram que os quatro aglomerados correspondiam a AML, ALL - linhagem T, ALL - linhagem B e ALL - linhagem B, respectivamente - Figura 10(b). Portanto, a abordagem de descoberta de classes usando redes SOM automaticamente descobriu as diferenças entre AML e ALL, como também entre as células ALL dos tipos B e T. Essas são as distinções mais

importantes entre as leucemias agudas, ambas em termos da biologia quanto do tratamento clínico. Ou seja, a rede SOM conseguiu dividir os padrões em quatro aglomerados, encontrando uma outra categorização biológica importante.

Um outro exemplo de descoberta de classes utilizando técnicas de agrupamento pode ser encontrado em [ALI00]. Nesse trabalho, o linfoma difuso de grandes células B (DLBCL, do inglês *Diffuse Large B-cell Lymphoma*) foi estudado usando 96 amostras de linfócitos, sendo 72 de células normais e 24 de células malignas, cada amostra contendo 4026 genes. Por meio da aplicação da técnica UPGMA a essas amostras, os autores mostraram que há uma diversidade na expressão gênica entre tumores de pacientes com DLBCL.

Foram identificadas duas formas moleculares distintas de DLBCL, que tinham padrões de expressão gênica indicativa de estágios diferentes da diferenciação da célula B. De fato, esses dois grupos estão correlacionados com a taxa de sobrevivência dos pacientes, portanto confirmando que os aglomerados gerados são biologicamente significativos.

Rosenwald e outros autores usaram a técnica de microarranjos para classificar a condição de sobrevivência, do paciente com Linfoma Difuso de Grandes Células B, após a quimioterapia. O conjunto é formado por 240 amostras de pacientes e 7399 genes. A análise do perfil da expressão do gene integrou duas técnicas complementares à predição do resultado. Na primeira técnica, usou-se agrupamento hierárquico para identificar os subgrupos que diferiram com respeito à expressão de centenas dos genes.

Na segunda técnica, usou-se dados clínicos e da expressão gênica para identificar os genes individuais que predisseram o resultado e combinaram estas variáveis em um modelo multi-variável. Este modelo incorporou diferenças, nos níveis da expressão do gene dos subgrupos do Linfoma Difuso de Grande Célula B, que influenciaram o resultado, bem como as diferenças na expressão do gene que foram associadas com a probabilidade de sobrevivência [ROS02].

As técnicas supervisionadas representam uma alternativa poderosa que pode ser aplicada se existe informação prévia sobre a classe dos genes. Por exemplo, as técnicas de Aprendizagem de Máquina Supervisionadas podem ser utilizadas para as tarefas de predição e classificação.

Um exemplo de aplicação da aprendizagem supervisionada pode ser encontrado em [SHI02]. Os autores utilizaram o algoritmo de voto ponderado para classificar os dois tipos de Linfoma: o Linfoma Difuso de Grandes Células B (LDGCB) e Linfoma Folicular (LF), analisando 6817 genes, obtidos de 77 pacientes (58 pertencentes ao grupo de LDGCB e 19 pertencentes ao grupo de LF). Com essa técnica obtiveram 91% de acerto.

O sucesso da classificação na distinção entre os dois tipos de tumores motivou os autores avaliar a condição de sobrevivência dos pacientes com LDGCB (58 pacientes). O conjunto foi dividido em 32 pertencentes ao grupo de pacientes curados e 26 não curados. Para a classificação desse conjunto utilizaram dois algoritmos: SVM (*Support Vector Machines*) e *k*-NN (*k-Nearest Neighbour*) e obtiveram uma taxa de acerto de 72% e 68% respectivamente.

Gordon e seus companheiros obtiveram uma alta taxa de precisão quando classificaram o câncer de pulmão em adenocarcionama (ADCA) e mesotelioma pleural maligno (MPM) de 95% e 99% respectivamente. O conjunto de dados foi obtido através do nível de expressão de um pequeno número de genes [GOR02].

O câncer do Sistema Nervoso Central foi estudado por Pomeroy e seus colegas que analisaram o resultado do tratamento do câncer em 60 pacientes [POM02]. No estudo, os autores utilizaram a aprendizagem não-supervisionada, utilizando rede SOM, análise de componentes principais e a aprendizagem supervisionada usando o algoritmo *k*-NN.

Uma questão fundamental, tanto para técnicas supervisionadas, quanto para não supervisionadas (embora mais nas primeiras) é que dados de microarranjos apresentam novos desafios aos algoritmos de AM, geralmente desenvolvidos para lidar com um grande número de amostras (exemplos) com relativamente poucos atributos ou características.

Assim, métodos de seleção de atributos são utilizados em experimentos de microarranjos. Lau e Schultz utilizaram métodos de seleção de atributos em duas bases de dados de expressão gênica [LAU03]. A primeira base é a mesma estuda por Golub juntamente com outros autores que distingue os dois tipos de leucemia: AML e ALL [GOL99]. A segunda base de dados corresponde ao nível de expressão gênica da célula. As amostras eram classificadas em normal e tumor celular. Os experimentos foram divididos em etapas. Na primeira etapa os autores fizeram uma redução preliminar dos atributos da base, utilizando o método *Random Forest*. Na segunda etapa fizeram a seleção de atributos, utilizando os seguintes métodos de seleção: busca dispersa, algoritmos genéticos, seleção seqüencial para frente e eliminação para trás. E na terceira etapa a classificação dos subconjuntos, o qual utilizaram o procedimento de validação cruzada e os algoritmos: Árvore de Decisão, k-NN, SVM e Redes Neurais [LAU03].

Os resultados não são conclusivos, pois em alguns casos os subconjuntos de atributos que foram gerados a partir de método *Random Forest* obtiveram melhores resultados dos que foram gerados apenas por métodos de seleção atributos. Sendo assim não pode-se identificar qual o método utilizado deu melhor resultado.

Liu e seus companheiros [LIU02] fizeram um estudo comparativo usando seis métodos heurísticos de seleção de atributos e quatro métodos de classificação em duas bases de dados, sendo uma delas de dados de expressão gênica. A base é formada por 327 amostras de ALL (*Acute Lymphoblastic Leukemia*) obtida pela técnica de microarranjos contendo 12558 genes. Este conjunto de dados contém os subtipos de ALL incluindo *T-cell* (T-ALL), E2A-PBX1, TEL-AML1, MLL, BCR-ABL e *Hyperdiploid* (Hiperdip>50). Um sistema de árvore de decisão estruturada (ver figura 11), proposto por médicos, foi usado para classificar estas amostras.

Quando uma amostra é dada, regras são aplicadas primeiramente para classificar o subtipo T-ALL. Se esta é classificada como T-ALL, então o processo é terminado, caso contrário, o processo continua com outra amostra do subtipo de ALL. O processo de classificação baseia-se nesta árvore de decisão que pode ser terminada com o nível 6 onde o subtipo *Hyperdip*>50 e OTHERS são determinados.

Os dados foram divididos em 215 amostras de treinamento e 112 amostras de teste. Além disso, os autores [LIU02] fizeram a classificação pelos níveis do subtipo de ALL. Para a seleção de atributos os autores utilizaram os seguintes métodos: entropia, estatística x^2, estatística *t*, CFS, MTI. Para a classificação foram usados cinco classificadores: k-NN, C4.5, *Naïve Bayes* e SVM e o PCL (*Prediction by Collective Likelihood of Emerging Patterns*) desenvolvido pelos autores.

Para os dados de treinamento do nível 1, o método CFS selecionou apenas 1 gene (38319-at) do total de 12558 genes. Usando método de entropia foram selecionados 13 genes. O número de genes selecionados pelo método *all-*x^2 foi de 1309.

Os algoritmos de classificação SVM, NB, *k*-NN e PCL tiveram 100% de precisão sobre as amostras de teste para todos os seis grupos de genes selecionados. No nível dois, o método CFS selecionou 1 gene (33355-at), o método de entropia 8 genes, o método *all* x^2 827 genes. Todos os cinco classificadores tiveram precisão de 100% de acerto sobre as amostras de teste. Nos níveis 3, 4, 5 e 6 o resultado não foi tão preciso quanto nos níveis anteriores, porém ainda foi considerado muito bom. O método de entropia foi o que apresentou melhor resultado sobre as amostras de teste.

Wang e seus colegas utilizaram duas abordagens de seleção de atributos: a filtro e *wrapper*. Na abordagem filtro foi utilizado principalmente o algoritmo CFS, e outros 4 métodos: Estatística x^2, Ganho de Informação, *Relief-F* e simetricamente incerto em conjunto com a busca *best first*. Esses métodos forma aplicados em dois conjuntos de dados ALL/AML e DLBCL [WAN05].

Após a seleção os subconjuntos foram submetidos aos classificadores: árvore de decisão, através dos algoritmos C4.5, *Naïve Bayes* e SVM.

Ambos os conjuntos de dados tiveram bons resultados. Analisando o conjunto de dados ALL/AML os autores perceberam que o gene “Zyxin” foi selecionado em várias execuções e descobriram que este é um gene que está envolvido na distinção dos dois tipos de leucemia. Na base DLBCL alguns genes também foram selecionados com mais freqüência entre eles GENE3330X, GENE3328X, porém esses genes ainda não são conhecidos, ou seja, não possuem sua anotação gênica.

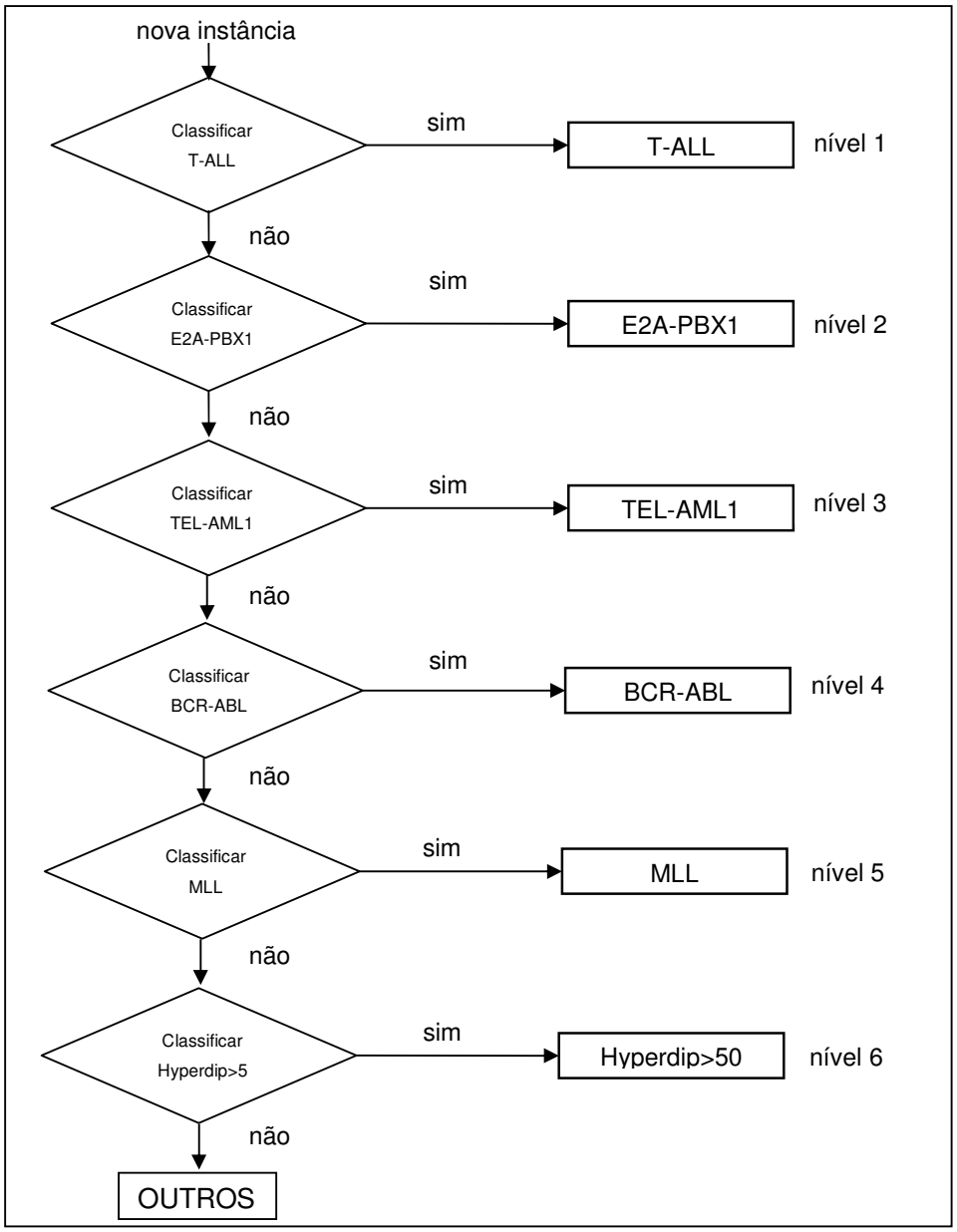

Figura 11: Sistema de Árvore Estruturada para a Predição dos seis subtipos de amostras de ALL [LIU02].

Borges e Nievola, utilizando a mesma base de dados utilizada por Alizadeh em seus experimentos, aplicaram algoritmos de seleção de atributos buscando melhorar o desempenho dos algoritmos de classificação [BOR05a], [BOR05b].

O conjunto de dados utilizado possui 47 exemplos, sendo que 24 deles pertencem ao grupo do centro germinativo da célula B, enquanto 23 pertencem ao grupo da ativação da célula B. Cada exemplo é descrito por 4026 genes, todos com valor numérico, além do atributo meta. O objetivo é identificar a qual classe cada uma das amostras está relacionada: *germinal* ou *activated*.

Para aplicar os métodos de seleção de atributos nesse trabalho, os autores dividiram em dois blocos principais: busca dos subconjuntos de atributos e avaliação dos subconjuntos encontrados.

Como métodos de buscas foram utilizados a busca seqüencial e busca aleatória. Para avaliação dos subconjuntos usou-se duas abordagens principais: a abordagem filtro e a abordagem *wrapper*. Como medidas de avaliação da abordagem filtro utilizou-se a medida de dependência e a medida de consistência. Como a abordagem *wrapper* utiliza o próprio algoritmo de mineração para fazer a avaliação dos subconjuntos, usou-se os seguintes algoritmos: *Naïve Bayes*, Rede Bayesiana, C4.5, Tabela de Decisão e *k*-NN (para os valores de k=1,k=3, k=5 e k=7)

Nesses experimentos pode-se notar uma grande variação no número de atributos selecionados conforme o método de busca e avaliação do subconjunto utilizado. Comparando os resultados obtidos, fica claro que a seleção de atributos melhorou os resultados da classificação em praticamente todas as situações. Em particular, o uso do método de avaliação *wrapper* produziu melhores resultados de maneira constante. Tal resultado foi evidente a partir do momento em que se usa o classificador de maneira consistente com o algoritmo embutido utilizado durante a seleção de atributos. Também observou que o gene identificado na base como GENE3330X foi selecionado em quase todos os procedimentos de seleção. Sendo assim, fica clara a importância desse gene na classificação dessa anomalia.

3 Metodologia

Para o desenvolvimento do trabalho serão seguidas as etapas do processo de descoberta do conhecimento, baseado nas três etapas principais: pré-processamento, mineração de dados e pós-processamento.

Para a execução de uma parcela dos experimentos foi utilizado o *software Weka* versão 3.4. O *Weka* (*Waikato Environment for Knowledge Analysis*) é um pacote formado por um conjunto de implementações de algoritmos de diversas técnicas de Mineração de Dados [WIT05], [SCU04] e [SCU06]. Desenvolvido na Universidade de *Waikato* na Nova Zelândia, ele está implementado na linguagem Java, que tem como principal característica ser portável; desta forma pode ser executável nas mais variadas plataformas, além de, ser um *software* de domínio público.

Basicamente o processo geral a ser seguido pode ser visualizado na figura 12. Esse processo será dividido em 4 partes principais. Na primeira parte se concentrará na classificação dos conjuntos de dados utilizando todos os atributos. Na segunda parte será feita a seleção de atributos e após a classificação dos subconjuntos. Na terceira parte será executado método de projeção aleatória e em seqüência os algoritmos de classificação e na quarta parte do processo serão utilizados os dois métodos em conjunto (projeção aleatória e seleção de atributos) e após isto os subconjuntos serão submetidos aos algoritmos de classificação.

Cada parte detalhada do processo será descrita nas seções a seguir.

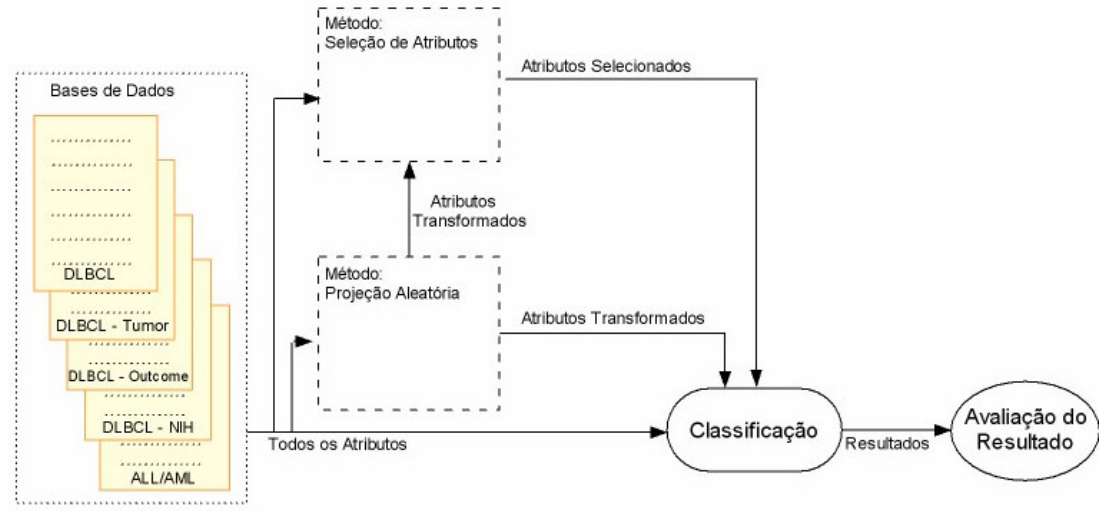

Figura 12: Passos Gerais Executados nos Experimentos.

3.1 Consolidação dos Dados e Descrição dos Conjuntos de Dados

Primeiramente foram selecionados os 5 conjuntos de dados que seriam estudados, quatro deles sobre o estudo do linfoma e um sobre leucemia. Os dados utilizados no experimento foram extraídos do repositório de análise de dados biomédicos *Kent Ridge*[3] e disponibilizados no formato "arff" do *software weka*. Além disso, esse repositório disponibiliza os dados brutos das bases de dados.

Foi feita uma análise detalhada em cada base onde se verificou a quantidade de atributos, de amostras, número de classes e a divisão do número de amostras, conforme a classe pertencente (Ver Figura 13).

- **Base de Dados DLBCL**

Essa base é constituída de dados de expressão gênica, obtida através da técnica de microarranjos, no qual os autores estudaram o câncer Linfoma Difuso de Grandes Células B (LDGCB). Este é o subtipo de linfoma não-*Hodgkin* mais comum que constituem um grupo de cânceres (tumores malignos). Foram identificadas duas formas distintas de células de LDGCB, as quais tiveram padrões de expressão gênica indicada por dois estágios diferentes na diferenciação da célula B. Um tipo de expressão gênica caracteriza o centro germinativo da célula B e o segundo tipo de expressão gênica é a ativação da célula B [ALI00].

O conjunto de dados é formado por 47 exemplos, sendo que 24 deles pertencem ao grupo do centro germinativo da célula B, enquanto 23 pertencem ao grupo da ativação da célula B. Cada exemplo é descrito por 4026 genes, todos com valor numérico, além do atributo meta.

- **Base de Dados DLBCL - Tumor**

Esse conjunto é formado por dois tipos de Linfoma: o Linfoma Difuso de Grandes Células B (LDGCB) e Linfoma Folicular (LF). O conjunto possui 7129 atributos e 77 amostras de pacientes, sendo 58 pertencentes ao grupo de LDGCB e 19 pertencentes ao grupo de LF [SHI02].

- **Base de Dados DLBCL - Outcome**

Esse conjunto de dados estuda o resultado do tratamento após 5 anos dos pacientes com LDGCB. O objetivo nessa base é identificar quais pacientes tiveram sucesso no tratamento e foram curados e os que não foram curados. A base é composta por 58 amostras, 32 pertencentes o grupo dos pacientes curados e 26 ao grupo dos pacientes não curados e descrita através de 7129 atributos [SHI02].

- **Base de Dados DLBCL - NIH**

---

[3] http://sdmc.i2r.a-star.edu.sg/rp/ - Acessado em: 25/04/2006

O objetivo dessa base é analisar a condição de sobrevivência dos pacientes com Linfoma Difuso de Grandes Células B após a quimioterapia [ROS02]. O conjunto é formado por 240 amostras de pacientes (102 vivos e 138 não vivos) e 7399 atributos (genes).

- **Base de Dados ALL/AML**

Essa base analisa dois tipos de leucemia aguda [GOL99]. O conjunto possui 7129 atributos (genes) e 72 amostras divididas em 47 pertencentes ao tipo de leucemia ALL (leucemia linfoblática aguda) e 25 pertencentes a AML (leucemia mielóide aguda).

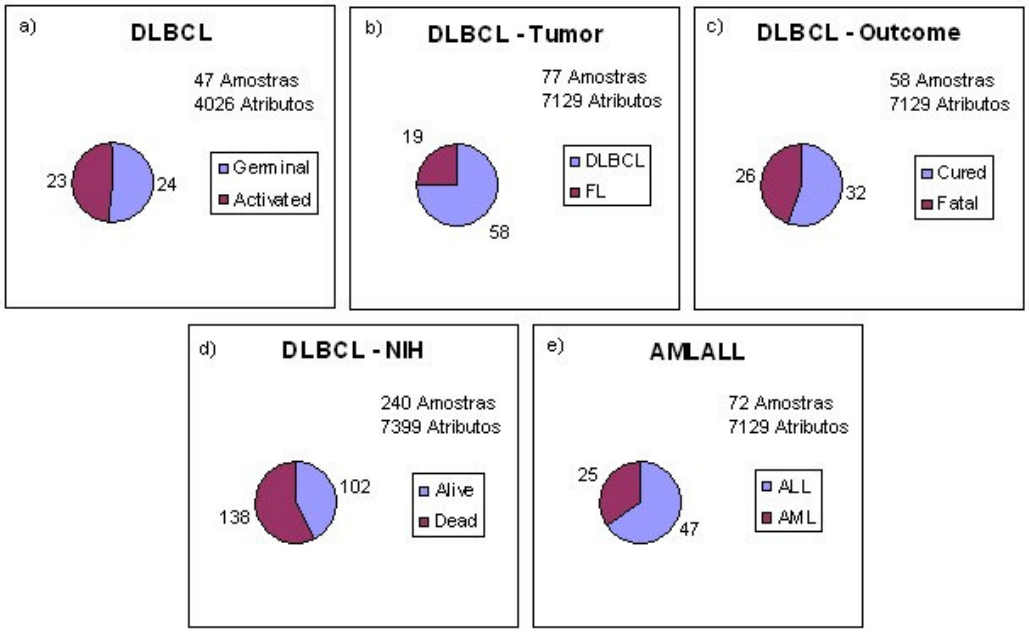

Figura 13: Representação da Divisão das Classes das Bases de Dados.

3.2 Pré-Processamento

Na fase de pré-processamento foi feita a redução da dimensionalidade dos dados utilizando dois métodos diferentes: a seleção de atributos e a projeção aleatória. Primeiramente foi executado separadamente cada um dos métodos e depois foram executados os dois métodos em conjunto, primeiramente o método de projeção aleatória e na saída deste a seleção de atributos. Para a execução dos métodos de redução da dimensionalidade usou-se a base de dados completa.

3.2.1 Execução da Seleção de Atributos

Para a execução da seleção de atributos foram utilizados dois tipos de busca: busca seqüencial e busca aleatória e medidas de avaliação pertencente a abordagem filtro e a abordagem *wrapper*. Para a abordagem filtro foram escolhidas as medidas

conhecidas como dependência e consistência. Para a abordagem *wrapper*, como esta abordagem utiliza o algoritmo de mineração como avaliação do subconjunto, quatro algoritmos de classificação foram usados: *Naïve Bayes*, indução de árvore de decisão, executada pelo algoritmo J48 que é a versão C4.5 no *Weka*, SVM e k-NN. Todos esses algoritmos foram executados considerando seus parâmetros configurados com os valores padrões.

Para o classificador SVM os principais parâmetros utilizados foram: c (complexidade do modelo) com valor igual a 1.0 e o *epsilon* (erros de arredondamento) com valor igual a 1.0e-12. Para o classificador *k*-NN utilizou-se valores de *k*=1, *k*=3, *k*=5 e *k*=7.

A figura 14 mostra como foi feita a seleção de atributos, através da combinação dos métodos de busca e das medidas de avaliação dos subconjuntos gerados.

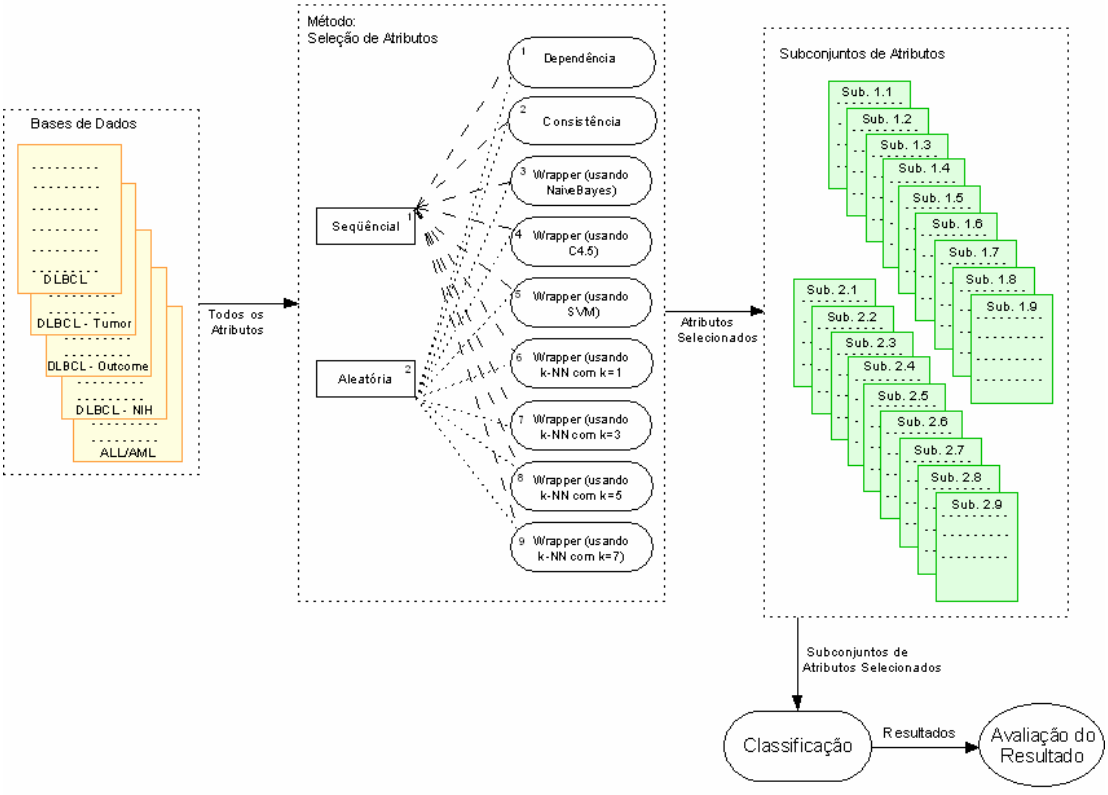

Figura 14: Passos para execução da Seleção de Atributos.

A busca seqüencial aplicada nos experimentos faz uma busca “gulosa” para frente iniciando com nenhum atributo. Como critério de parada, a busca pára quando a adição de um atributo resulta na redução da avaliação, ou seja, o seu erro aumenta.

A busca Genética usada nesses experimentos faz uma busca usando um algoritmo genético simples [GOL89]. A avaliação da solução encontrada (função de *fitness* ou função de aptidão) foi baseada no número de instâncias classificadas corretamente. A população inicial foi criada de acordo com uma distribuição aleatória uniforme. O conjunto dos parâmetros foi estabelecido *a priori*, com o tamanho da população e do número de gerações que estão sendo definidos como 20, e a probabilidade de operadores do cruzamento e da mutação definida como 0.6 e 0.033, respectivamente.

3.2.2 Execução do Método de Projeção Aleatória

A figura 15 mostra os passos seguidos para a execução do método.

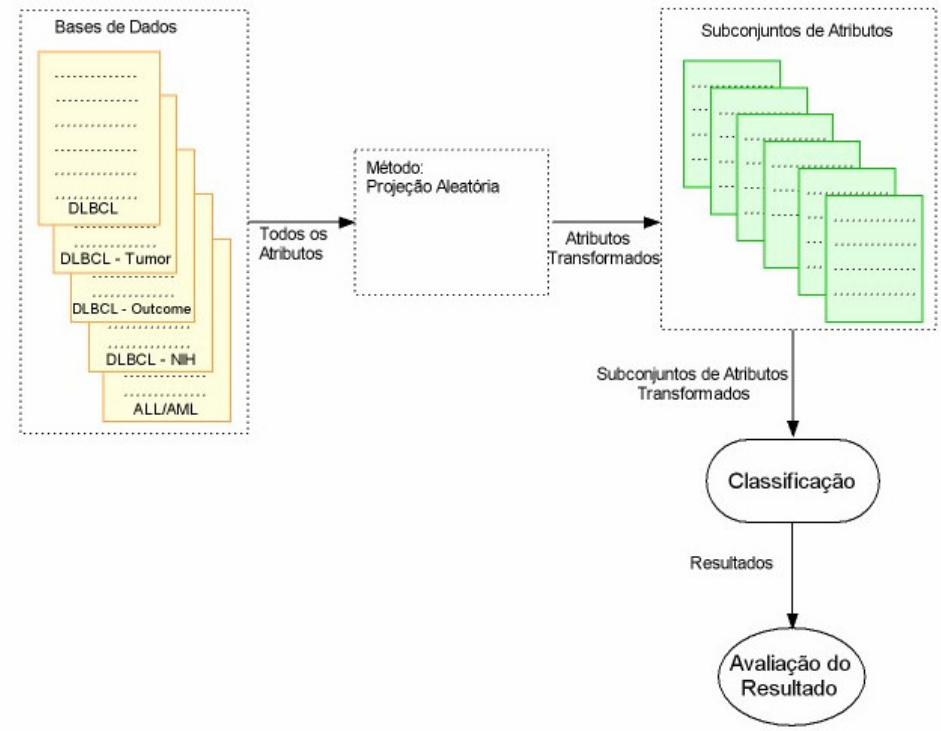

Figura 15: Passos para a Execução do Método Projeção Aleatória.

A dimensão do novo conjunto de atributos foi definida de forma aleatória, porém igual para as cinco bases. Para a formação desse novo conjunto dois critérios foram escolhidos: um utilizando um número fixo de atributos e o outro a porcentagem de atributos. Para o número de atributos foram escolhidos os seguintes valores: 10, 15, 30, 45, 71 atributos. Já para a porcentagem de atributos foram escolhidos: 3%, 10%, 20%,

25% e 50%. Dessa forma será possível comparar os resultados entre as bases através do percentual e também de um número fixo de atributos independentemente do tamanho da base. Sendo assim, 10 novos conjuntos foram gerados.

Para um mesmo número de atributos, o método foi executado 10 vezes variando a semente da geração da matriz aleatória. Foram escolhidos 10 valores aleatoriamente que variaram de 5 a 65. A distribuição utilizada para o cálculo da matriz aleatória está descrita na equação 15. Fazendo uma combinação entre o número de atributos e o número da semente aleatória pode-se observar que 100 novos subconjuntos foram gerados para cada base.

3.2.3 Utilização Conjunta do Método de Projeção Aleatória e da Seleção de Atributos

Os 100 subconjuntos de atributos gerados pelo método projeção aleatória também foram submetidos aos algoritmos de seleção de atributos descritos na subseção 3.2.1. Dessa forma para cada um dos subconjuntos foram gerados 18 novos subconjuntos, totalizando 1800 subconjuntos combinados pelas duas abordagens. Essa utilização conjunta dos métodos pode ser visualizada na figura 16.

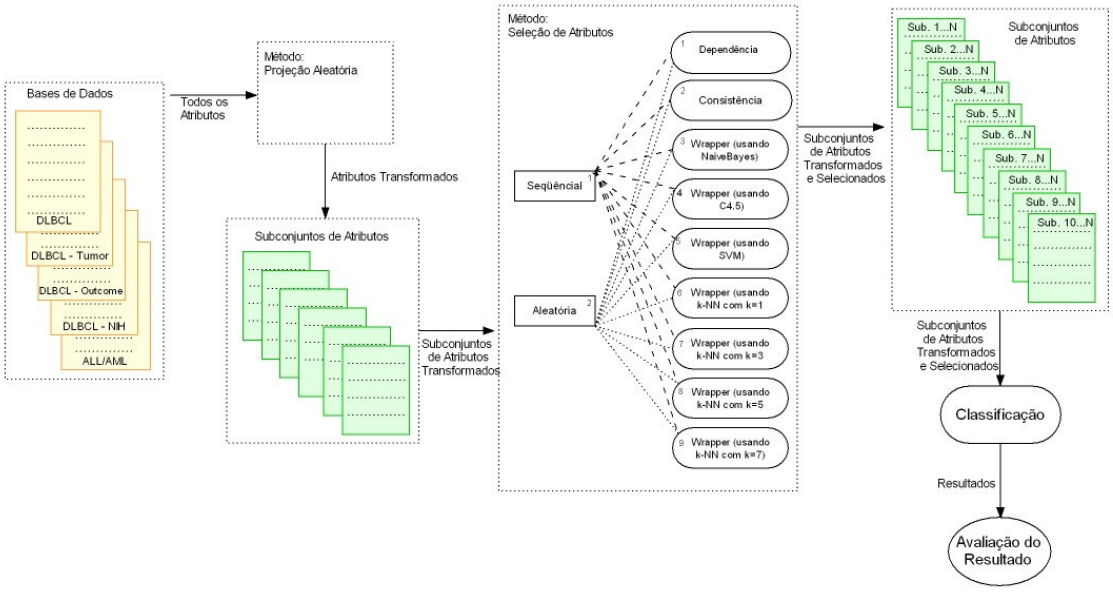

Figura 16: Utilização conjunta dos Métodos de Redução de Dimensionalidade.

3.3 Mineração de Dados

Um dos objetivos do trabalho é a redução da quantidade de atributos em bases de microarranjos, porém é importante lembrar que, se a redução de dimensionalidade for excessiva, o classificador pode ter seu poder de discriminação reduzido. Por isso, é importante analisar a variação do comportamento do classificador com base na quantidade total de atributos, de forma que seja possível estimar a dimensionalidade ideal do conjunto de dados para determinado classificador baseado na sua taxa de acerto.

Sendo assim, os cinco conjunto de dados originais foram submetidos a 4 classificadores: *Naïve Bayes*, C4.5, SVM e *k*-NN (para *k*=1, *k*=3, *k*=5 e *k*=7) como forma de ter um parâmetro de taxa de acerto do classificador para comparar com os outros subconjuntos gerados por métodos de redução de dimensionalidade.

Esses algoritmos foram escolhidos por pertenceram a paradigmas diferentes de aprendizagem e também pelo fato de que são algoritmos tradicionalmente usados e que produzem bons resultados.

O algoritmo *Naïve Bayes* tem base probabilística que está baseada na utilização do teorema de *Bayes*, além de fornecer bons resultados de desempenho e rápida execução, sendo um comparativo para outros algoritmos.

O algoritmo C4.5 é um algoritmo que produz um modelo, em forma de árvore, que pode ser facilmente interpretado por especialistas da área, e é um algoritmo que possui sua execução rápida e produz bons resultados.

O SVM tem boa capacidade de generalização e robustez diante de dados de grande dimensão com os dados de expressão gênica, embora seja um algoritmo que tem um custo computacional superior aos demais devido ao tipo de base que está se utilizando.

O *k*-NN é um algoritmo simples, diferente dos demais métodos de aprendizagem. Embora seja um algoritmo *lazy* produz bons resultados.

Os resultados foram avaliados por meio de estimativa de média de acerto de cada classificador empregando validação cruzada estratificada com 10 partições (*10 fold cross-validation*).

A validação cruzada estratificada consiste primeiramente, em dividir aleatoriamente a base de dados original em *k* partições iguais ou aproximadamente iguais, tanto no número de exemplos, quanto na distribuição das classes. Assim, cada partição tem uma distribuição de classes igual ou aproximadamente igual à distribuição de classes da base de dados completa. Para esse procedimento seleciona-se *k*-1 partições para formar a base de treinamento e, em seguida, computa-se a taxa de acerto

para a partição que ficou de fora. Este passo se repete até que todas as *k* partições sejam usadas *k*-1 vezes como base de treinamento e uma vez como base de teste. Ao final, a taxa de acerto é obtida pela média aritmética das taxas de acerto obtidas (nos dados de teste) para as *k* partições.

3.4 Pós-Processamento

Como essa etapa visa a avaliação dos resultados obtidos ela será discutida no capítulo 4, onde os resultados serão apresentados.

4 Resultados

Foram obtidos resultados da execução dos algoritmos sobre as bases de dados originais, sobre os subconjuntos de atributos gerados através dos métodos de seleção de atributos, do método de projeção aleatória e da utilização conjunta de ambos os métodos.

4.1 Resultado dos Classificadores nas Bases de Dados com todos os Atributos

Com objetivo de estimar o subconjunto ideal baseado na taxa de acerto dos classificadores, foram executados os algoritmos primeiramente sobre as bases de dados com todos os atributos para futuras comparações com os métodos de redução da dimensionalidade.

Como já citado anteriormente, as cinco base de dados analisadas foram submetidas aos 4 classificadores: *Naïve Bayes*, C4.5, SVM e *k*-NN (para *k*=1, *k*=3, *k*=5 e *k*=7), gerando 7 resultados.

A figura 17 mostra os resultados dos classificadores aplicados sobre as bases de dados quando todos os atributos foram usados. A barra indica a média dos classificadores utilizados e as linhas representam o limite inferior e o limite superior através do desvio padrão. A figura 17a mostra os resultados dos classificadores da base de dados DLBCL. A figura 17b e 17c mostra o resultado das bases de dados DLBCL-*Tumor* e DLBCL-*Outcome*. A figura 17d mostra o resultado da base de dados DLBCL-NIH e a figura 17e o resultado da base de dados ALL/AML.

Para uma melhor precisão dos resultados utilizou-se o método de validação cruzada estratificada. Além disso, para todos os resultados foram feitas análises estatísticas para avaliar a significância dos resultados através do teste de hipótese com *p-value* igual a 0.05, o qual estabelece o erro tolerado e, ao mesmo tempo, define a região de rejeição da hipótese nula.

O termo estatisticamente significante utilizado nesse trabalho refere-se ao resultado de uma comparação de algoritmos utilizando a estatística do teste-t pareado, onde o objetivo fundamental é avaliar o comportamento das diferenças observadas em cada elemento [SCU06] e [WIT05]. Maiores informações sobre a estatística do teste-t pareado podem ser encontradas no Apêndice C.

Uma característica a ser notada nas figuras 17c e 17d é que os resultados são bastantes inferiores comparados com os resultados das outras bases. Esse resultado pode ser devido ao tipo de dados que está se analisando. Nessas duas bases o resultado não é dependente somente da interação de alguns genes e sim do tipo de tratamento, estágio da doença, características do paciente (como idade, sexo, etc.) entre outros

fatores, porém esses dados não estão disponíveis nas bases de dados, o que leva a ser um dos motivos da baixa taxa de acerto.

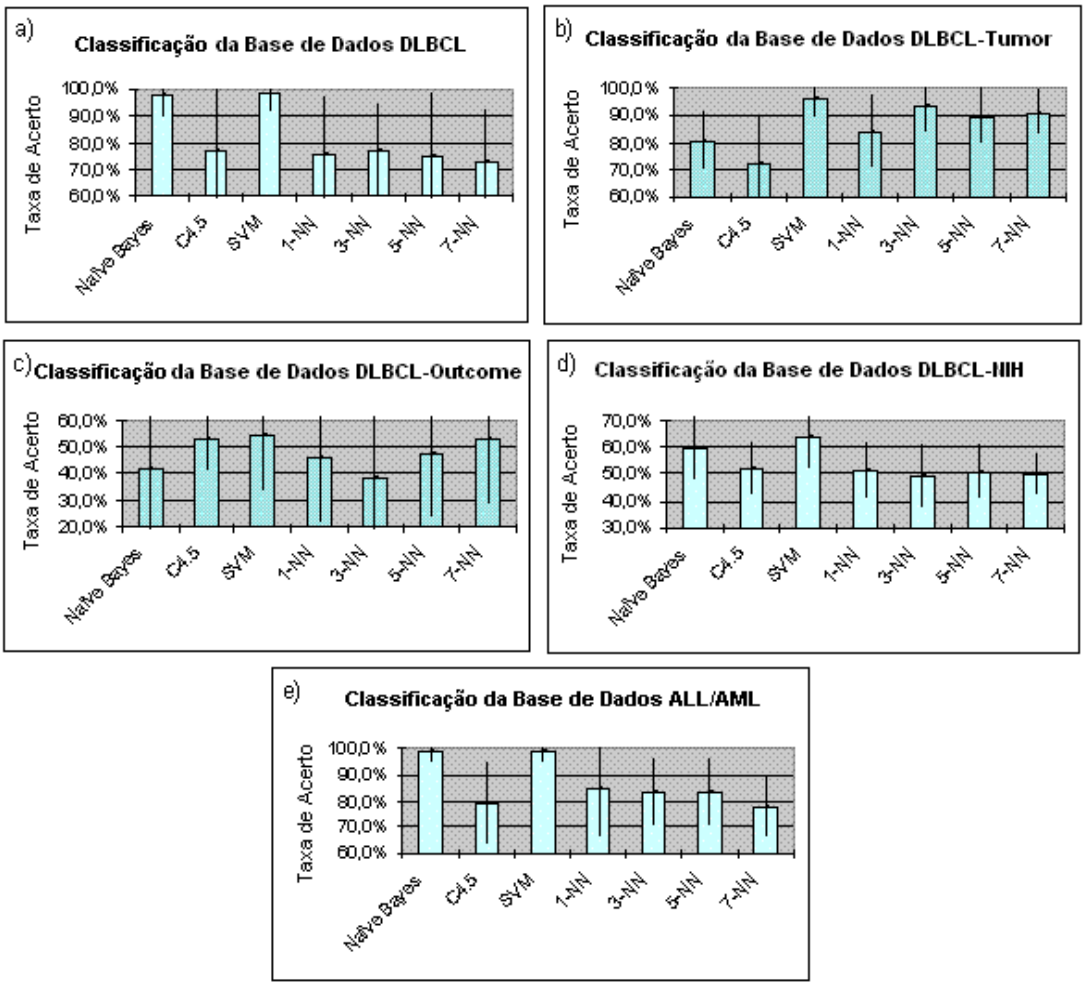

Figura 17: Resultado da Classificação das Bases de Dados com todos os Atributos (observar que as escalas são diferentes no eixo da taxa de acerto).

Na tabela 3, são mostrados os resultados dos algoritmos de classificação de cada base de dados. A anotação "*" indica que um resultado é estatisticamente pior que o resultado do algoritmo base *Naïve Bayes* e os destacados em negrito indicam que os resultados são estatisticamente melhor comparados com o algoritmo de base.

Se definirmos o algoritmo *Naïve Bayes* como algoritmo padrão a ser comparado com a taxa de acerto dos outros algoritmos, observa-se que os algoritmos 1-NN, 3-NN e 7-NN, nos subconjuntos de atributos da base de dados DLBCL tiveram resultados estatisticamente piores. Apesar do algoritmo 3-NN ter resultado superior ao 5-NN e

possuir uma taxa de acerto igual ao algoritmo C4.5, ele é considerado estatisticamente pior. Já nos subconjuntos de atributos da base de dados DLBCL-Tumor o algoritmo SVM apresentou resultados estatisticamente melhores em relação ao algoritmo de base. Embora, o algoritmo C4.5 possuir uma taxa de acerto baixa ele é considerado equivalente ao algoritmo de base.

Tabela 3: Resultados da Classificação das Bases de Dados com todos os Atributos.

	Algoritmos						
Base de Dados	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DCLCL	97,50 ± 7,91	77,00 ± 23,71	98,00 ± 6,32	75,50 ± 21,27*	77,00 ± 17,51*	75,00 ± 23,69	73,00 ± 18,74*
DLBCL-Tumor	80,54 ± 10,70	72,50 ± 16,15	**96,07 ± 6,34**	84,11 ± 13,56	93,21 ± 9,85	89,82 ± 9,93	91,07 ± 8,54
DLBCL-Outcome	42,00 ± 24,81	53,33 ± 11,55	54,33 ± 20,73	45,67 ± 24,65	38,67 ± 24,61	47,67 ± 23,83	53,00 ± 24,47
DLBCL-NIH	59,58 ± 11,96	52,08 ± 9,87	63,75 ± 11,46	51,25 ± 10,22	49,17 ± 11,59	50,83 ± 9,78	50,00 ± 7,61
ALL/AML	98,57 ± 4,52	78,93 ± 15,63*	98,57 ± 4,52	84,64 ± 18,14	83,39 ± 12,88*	83,39 ± 12,88*	77,86 ± 11,30*

Nas bases de dados DLBCL-Outcome e DLBCL-NIH, como já comentado anteriormente, sua taxas de acerto foram baixas comparada com as outras bases. Nos subconjuntos de atributos gerados dessas bases de dados as taxas de acerto dos classificadores também foram baixa, porém esses resultados são considerados estatisticamente equivalentes.

Analisando os subconjuntos de atributos da base de dados ALL/AML estatisticamente, além do algoritmo 7-NN, que possui dentre todos os outros algoritmos taxa de acerto mais baixa, os algoritmos C4.5, 3-NN e 5-NN também possuem resultados piores.

4.2 Resultado da Seleção de Atributos sobre as Bases de Dados

4.2.1 Seleção de Atributos

Como citado no capitulo anterior foram utilizadas duas abordagens para a seleção de atributos: filtro e *wrapper*. A tabela 4 mostra o critério de busca e a medida de avaliação utilizada para a geração de cada subconjunto, a fim de compreender os resultados obtidos por cada método.

É possível observar na tabela 5 a grande variação no número de atributos selecionados por cada método. O método de busca seqüencial (subconjuntos 1.1 a 1.9) causou uma grande redução no número de atributos selecionados comparado com o método de busca aleatório. Esta é uma das características do algoritmo, pois ele adiciona um atributo de cada vez e dessa forma tende a encontrar um subconjunto de atributos pequeno. Para a busca aleatória usou-se um algoritmo genético o qual teve um grande

número de atributos selecionados. Isso acontece devido à aleatoriedade na escolha dos atributos, que não privilegia um tamanho específico de subconjunto.

Tabela 4: Algoritmos de Seleção de Atributos.

Critério de Busca	Medida de Avaliação	Subconjunto de Atributos Gerados pelo Método de Seleção
Seqüencial	Dependência	Subconjunto 1.1
	Consistência	Subconjunto 1.2
	Wrapper (usando *Naïve Bayes*)	Subconjunto 1.3
	Wrapper (usando C4.5)	Subconjunto 1.4
	Wrapper (usando SVM)	Subconjunto 1.5
	Wrapper (usando 1-NN)	Subconjunto 1.6
	Wrapper (usando 3-NN)	Subconjunto 1.7
	Wrapper (usando 5-NN)	Subconjunto 1.8
	Wrapper (usando 7-NN)	Subconjunto 1.9
Aleatória	Dependência	Subconjunto 2.1
	Consistência	Subconjunto 2.2
	Wrapper (usando *Naïve Bayes*)	Subconjunto 2.3
	Wrapper (usando C4.5)	Subconjunto 2.4
	Wrapper (usando SVM)	Subconjunto 2.5
	Wrapper (usando 1-NN)	Subconjunto 2.6
	Wrapper (usando 3-NN)	Subconjunto 2.7
	Wrapper (usando 5-NN)	Subconjunto 2.8
	Wrapper (usando 7-NN)	Subconjunto 2.9

Tabela 5: Seleção de Atributos nas Bases de Dados.

		Nº. Atributos Selecionados em cada Subconjunto de cada Base de Dados				
	Subconjunto	DLBCL	DLDCL Tumor	DLBCL Outcome	DLBCL NIH	ALL/AML
Sequêncial	1.1	33	64	35	45	51
	1.2	3	4	6	14	3
	1.3	3	5	5	6	4
	1.4	1	2	4	12	1
	1.5	3	5	9	11	4
	1.6	3	7	1	1	4
	1.7	2	4	8	1	7
	1.8	3	2	6	2	3
	1.9	4	3	6	2	3
Aleatória	2.1	1769	3073	2232	725	2252
	2.2	305	383	1908	301	162
	2.3	1391	2770	1174	2743	3567
	2.4	892	1333	2841	3806	2411
	2.5	1469	3498	1887	2617	1829
	2.6	2035	3657	3430	3162	3168
	2.7	1411	3496	1026	2501	2058
	2.8	1058	2807	3568	382	3207
	2.9	1940	3690	2180	1731	3722

A figura 18 mostra com clareza esta diferença no número de atributos selecionados por cada método em que a quantidade de atributos é mostrada na escala logarítmica.

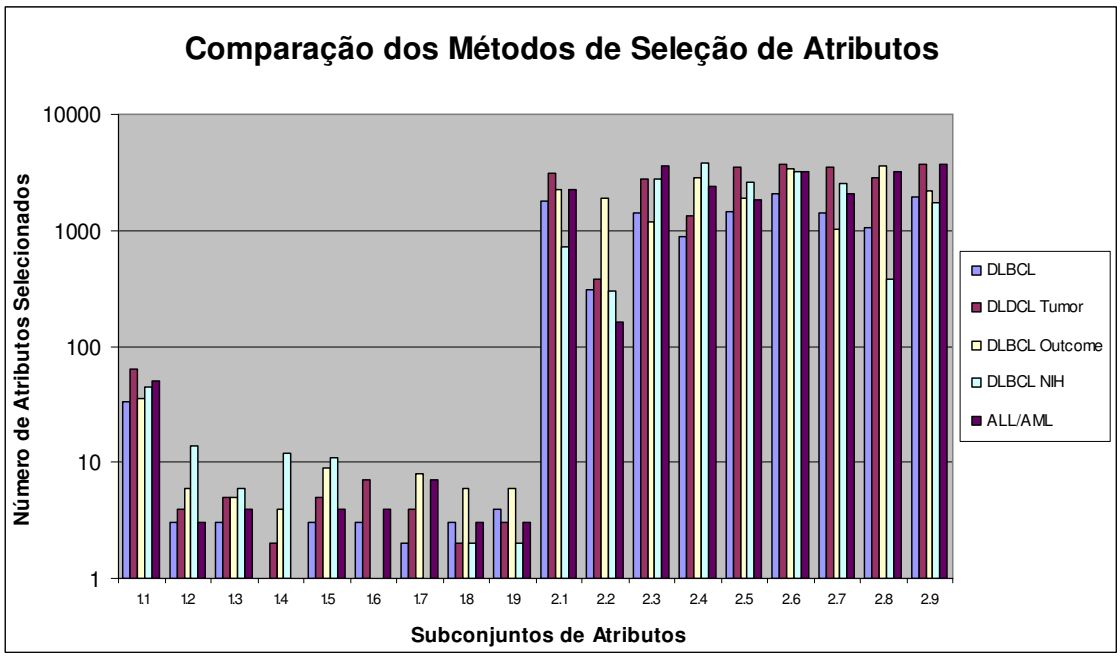

Figura 18: Comparação dos métodos de seleção de atributos de acordo com o número de atributos selecionados em cada Base de Dados.

4.2.2 Classificação dos Subconjuntos de Atributos

4.2.2.1 Abordagem Filtro

Para cada subconjunto de atributos foram aplicados os algoritmos de classificação. A figura 19 mostra a taxa de acerto dos classificadores nos subconjuntos de atributos de cada base de dados utilizando a abordagem filtro. A figura 19a refere-se aos subconjuntos de atributos da base de dados DLBCL, a figura 19b aos subconjuntos de atributos da base de dados DLBCL-Tumor, a figura 19c aos subconjuntos da base de dados DLBCL-Outcome, a figura 19d aos subconjuntos da base de dados DLBCL-NIH e a figura 19e aos subconjuntos da base de dados ALL/AML.

Na tabela 6, os resultados mostram que o método de busca seqüencial e as medidas de avaliação dependência e consistência foram equivalentes em comparação com o algoritmo *Naïve Bayes* utilizado como padrão dos experimentos. No subconjunto de atributos que foi gerado através da busca aleatória e da medida de avaliação dependência dois resultados foram considerados estatisticamente piores: o 1-NN e o 7-NN. Já com a medida de avaliação consistência os resultados estatisticamente piores

foram nos algoritmos 1-NN e 3-NN. Se observarmos os resultados da base de dados original, nota-se que esses três algoritmos de classificação foram os que apresentaram os resultados piores nessa base de dados.

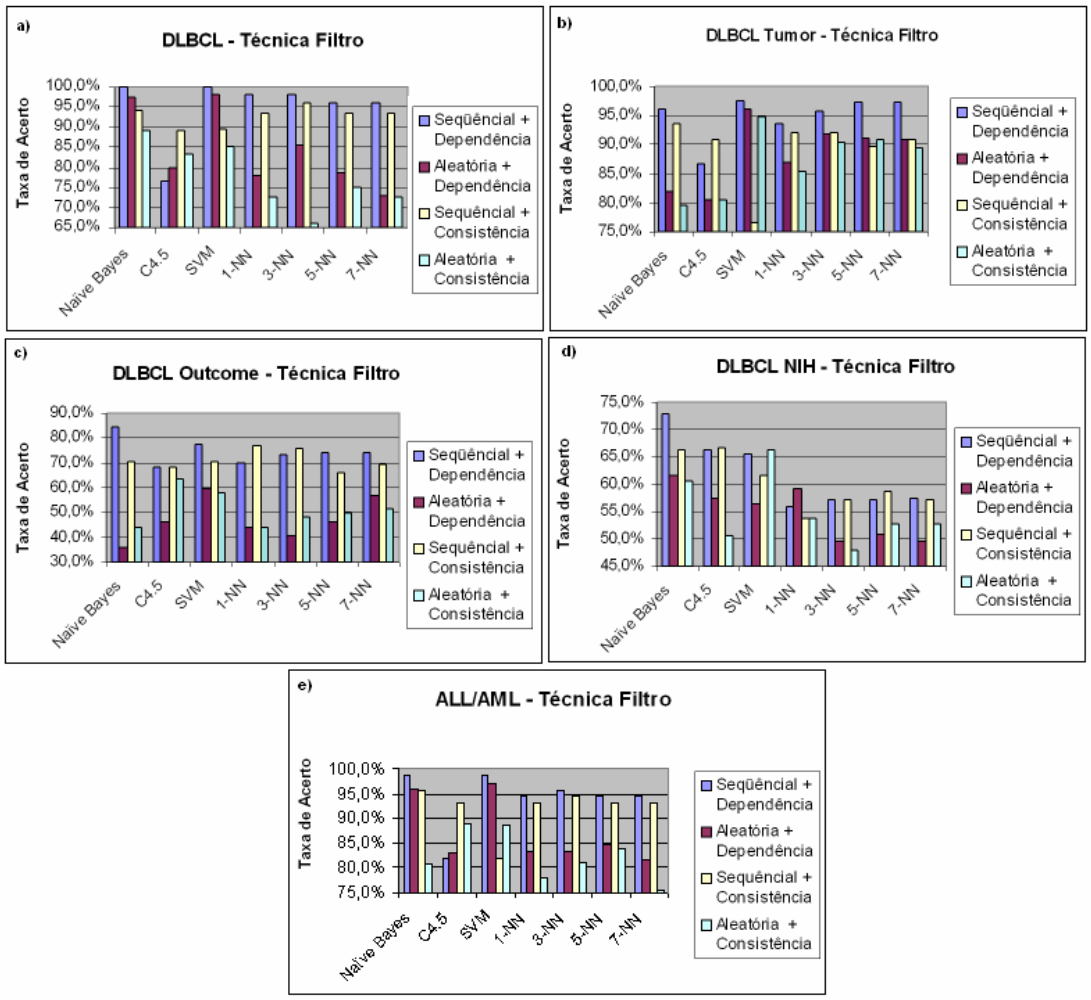

Figura 19: Resultado da Classificação dos Subconjuntos de Atributos utilizando a Abordagem Filtro.

Tabela 6: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados DLBCL após a Seleção de Atributos.

	Algoritmos						
Subconjunto	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	100,00 ± 0,00	76,50 ± 30,00	100,00 ± 0,00	98,00 ± 6,32	98,00 ± 6,32	96,00 ± 8,43	96,00 ± 8,43
1.2	94,00 ± 9,66	89,00 ± 11,74	89,50 ± 11,17	93,50 ± 10,55	96,00 ± 8,43	93,50 ± 10,55	93,50 ± 10,55
2.1	97,50 ± 7,91	80,00 ± 16,83	98,00 ± 6,32	78,00 ± 15,31*	85,50 ± 13,83	78,50 ± 22,37	73,00 ± 18,74*
2.2	89,00 ± 15,06	83,00 ± 13,17	85,00 ± 14,14	72,50 ± 8,90*	66,00 ± 16,96*	75,00 ± 15,63	72,50 ± 19,90

A tabela 7 mostra os resultados dos subconjuntos de atributos da base de dados DLBCL-Tumor onde a maioria dos resultados são estatisticamente equivalentes comparados com o algoritmo *Naïve Bayes*. Apenas dois resultados são considerados piores, o C4.5 no método que se utilizou a busca seqüencial e medida de avaliação dependência, e o SVM no método que se utilizou a busca seqüencial e medida de avaliação consistência.

Na tabela 8 são apresentados os resultados da aplicação da técnica filtro nos subconjuntos de atributos da base de dados DLBCL-Outcome. Apenas dois desses resultados são considerados piores comparados com o algoritmo *Naïve Bayes*, o C4.5 e o 7-NN. Ambos os resultados foram piores na execução do método de busca seqüencial e da medida de avaliação dependência.

Tabela 7: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados DLBCL-Tumor após a Seleção de Atributos utilizando a abordagem filtro.

	Algoritmos						
Subconjunto	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	96,07 ± 3,34	86,79 ± 12,26*	97,50 ± 5,27	93,57 ± 9,00	95,89 ± 6,63	97,32 ± 5,66	97,32 ± 5,66
1.2	93,57 ± 6,80	90,89 ± 10,96	76,61 ± 5,48*	91,96 ± 14,80	92,14 ± 9,00	89,64 ± 9,99	90,89 ± 8,64
2.1	81,96 ± 11,86	80,36 ± 13,86	96,07 ± 6,34	86,79 ± 10,75	91,79 ± 11,88	91,07 ± 8,54	90,89 ± 8,64
2.2	80,54 ± 13,56	79,46 ± 11,73	94,82 ± 6,71	85,36 ± 10,23	90,36 ± 13,45	90,89 ± 10,96	89,46 ± 10,56

Tabela 8: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados Conjunto DLBCL-Outcome após a Seleção de Atributos utilizando a abordagem filtro.

	Algoritmos						
Subconjunto	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	84,00 ± 13,59	68,33 ± 17,23*	77,33 ± 14,38	70,00 ± 28,50	73,00 ± 27,60	73,67 ± 20,45	73,67 ± 17,17*
1.2	70,33 ± 25,07	68,33 ± 15,34	70,33 ± 20,58	77,00 ± 17,10	75,67 ± 12,28	65,67 ± 15,48	69,00 ± 13,15
2.1	35,67 ± 26,11	46,33 ± 20,21	59,33 ± 22,54	44,00 ± 27,25	40,33 ± 28,26	46,00 ± 24,23	56,33 ± 27,86
2.2	44,00 ± 23,03	63,67 ± 20,69	57,67 ± 21,14	44,00 ± 27,25	47,67 ± 23,83	49,33 ± 21,93	51,33 ± 22,67

Na tabela 9 observa-se que o algoritmo k-NN apresentou piores resultados. O algoritmo *k*-NN com *k*=1 apresentou piores resultados em dois métodos de seleção de atributos, ambos na execução do método de busca seqüencial juntamente com as duas medidas de avaliação da abordagem filtro. O algoritmo 3-NN apresentou piores resultados no método de busca seqüencial juntamente com a medida de avaliação e no método de busca aleatória com a medida de avaliação consistência. No algoritmo 5-NN os piores resultados foram na execução do método de busca seqüencial com a medida de avaliação dependência e no método de busca aleatória com a medida de avaliação dependência. Por fim, o algoritmo 7-NN apresentou piores resultados no método de busca seqüencial e aleatória, mas apenas na medida de avaliação dependência.

Já a tabela 10 mostra os resultados obtidos nos experimentos realizados com a base de dados ALL/AML. Nessa tabela quase todos os resultados dos subconjuntos de atributos são estatisticamente equivalentes, apenas três desses resultados são considerados piores comparados com o algoritmo *Naïve Bayes*, o algoritmo C4.5 no método de busca seqüencial com a medida de avaliação dependência, o algoritmo 7-NN no método de busca aleatória com a medida de avaliação dependência e o algoritmo SVM no método de busca seqüencial com a medida de avaliação consistência.

Tabela 9: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados Conjunto DLBCL – NIH após a Seleção de Atributos utilizando a abordagem filtro.

	Algoritmos						
Subconjunto	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	72,92 ± 10,62	66,25 ± 8,88	65,42 ± 8,80	55,83 ± 8,15*	57,08 ± 8,80	57,08 ± 7,62*	57,50 ± 7,03*
1.2	66,25 ± 8,21	66,67 ± 6,80	61,67 ± 11,08	53,75 ± 11,19*	57,08 ± 6,53*	58,75 ± 6,35	57,08 ± 6,23
2.1	61,67 ± 11,42	57,50 ± 9,98	56,25 ± 7,67	59,17 ± 7,81	49,58 ± 9,10	50,83 ± 10,90	49,58 ± 9,31*
2.2	60,42 ± 9,05	50,42 ± 10,66	66,25 ± 12,02	53,75 ± 8,88	47,92 ± 11,83*	52,50 ± 11,98	52,50 ± 6,27

Tabela 10: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados ALL/AML após a Seleção de Atributos utilizando a abordagem filtro.

	Algoritmos						
Subconjunto	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NNN	7-NN
1.1	98,57 ± 4,52	81,96 ± 11,56*	98,57 ± 4,52	94,46 ± 9,83	95,71 ± 9,64	94,46 ± 9,83	94,46 ± 9,83
1.2	95,71 ± 9,64	93,04 ± 9,98	81,96 ± 11,56*	93,04 ± 12,04	94,64 ± 11,33	93,04 ± 12,04	93,04 ± 9,98
2.1	95,89 ± 6,63	83,04 ± 11,33	97,14 ± 6,02	83,39 ± 19,27	83,39 ± 12,88	84,64 ± 14,27	81,79 ± 11,93*
2.2	80,54 ± 13,69	89,11 ± 10,66	88,75 ± 13,10	78,04 ± 18,06	80,89 ± 10,72	83,75 ± 13,48	75,54 ± 12,43

Analisando cada método separadamente é possível observar os resultados das execuções de cada base de dados.

Na figura 20 é apresentado o resultado referente à busca seqüencial e a medida de avaliação dependência dos subconjuntos de atributos de cada base de dados. Observa-se que nas duas bases de dados (DLBCL-Outcome e DLBCL-NIH), onde há falta de dados referentes a sobrevida dos pacientes, a performance dos classificadores foi baixa comparada como as outras bases. Isso pode ser notado também nas figuras 21, 22 e 23 em que são mostrados os outros métodos de seleção de atributos utilizados em cada base de dados.

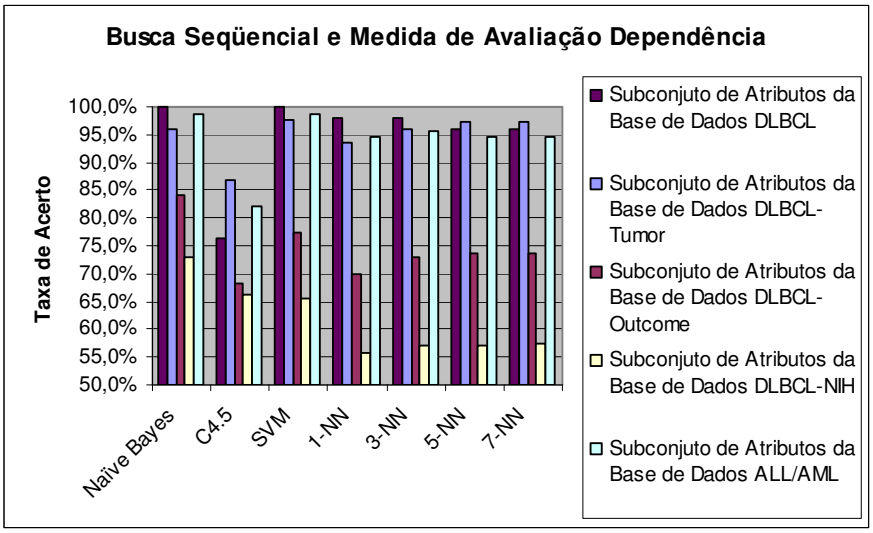

Figura 20: Taxa de acerto dos classificadores sobre os subconjuntos de atributos das bases de dados em que se utilizou a busca seqüencial e a medida de avaliação dependência.

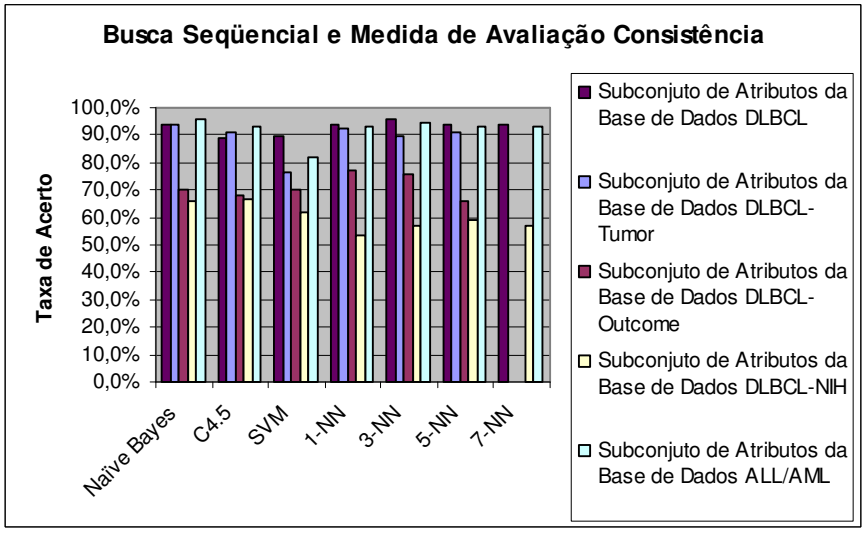

Figura 21: Taxa de acerto dos classificadores sobre os subconjuntos de atributos das bases de dados em que se utilizou a busca seqüencial e a medida de avaliação consistência.

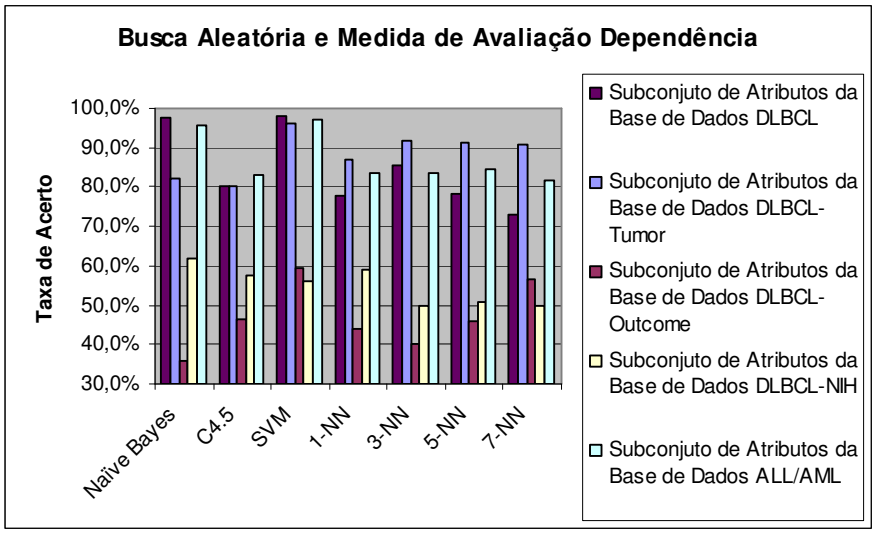

Figura 22: Taxa de acerto dos classificadores sobre os subconjuntos de atributos das bases de dados em que se utilizou a busca aleatória e a medida de avaliação dependência.

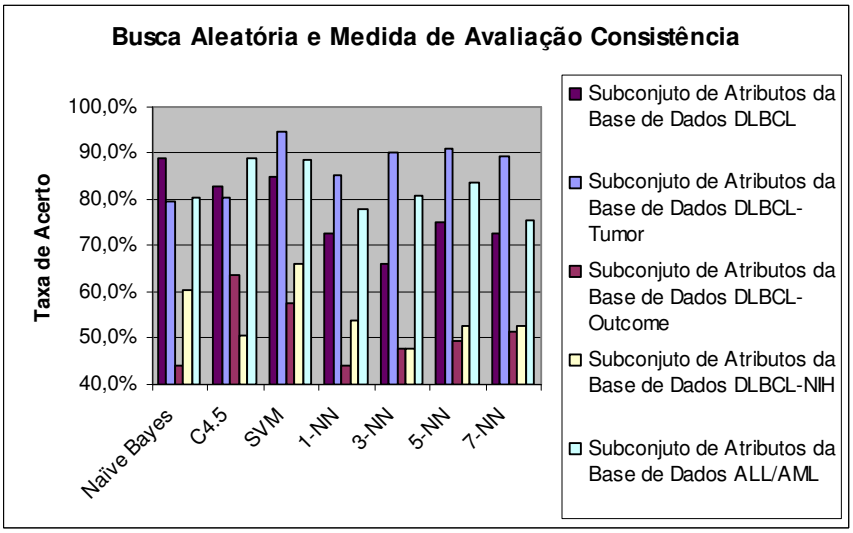

Figura 23: Taxa de acerto dos classificadores sobre os subconjuntos de atributos das bases de dados em que se utilizou a busca aleatória e a medida de avaliação consistência.

Calculando a média do resultado de todas as execuções, tabela 11, é possível analisar que no subconjunto 1.1 apenas o algoritmo SVM teve resultado estatisticamente equivalente ao algoritmo *Naïve Bayes*. Já no subconjunto 2.2 esse algoritmo teve resultado estatisticamente pior comparado com o algoritmo base *Naïve Bayes.* Nos subconjuntos 2.1 e 2.2 o resultado do algoritmo SVM foi estatisticamente melhor que o resultado dos demais algoritmos.

Analisando esses resultados, é possível observar também que o método de busca seqüencial e a medida de avaliação dependência apresentaram melhores resultados. Esse resultado fica mais visível quando se faz à média dos resultados de todas as execuções dessa abordagem (figura 24).

Tabela 11: Média das execuções dos métodos de seleção de atributos, utilizando a abordagem filtro, nos cinco subconjuntos de atributos das bases de dados.

	Algoritmos						
Subconjunto	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	90,3 ± 11,60	75,9 ± 8,72*	87,8 ± 15,57	82,4 ± 18,52*	83,9 ± 18,17*	83,7 ± 17,76*	83,8 ± 17,60*
1.2	84,0 ± 14,38	81,6 ± 12,94	76 ± 10,66	81,9 ± 17,16	83,1 ± 16,67	80,1 ± 16,58	80,7 ± 16,69
2.1	74,5 ± 26,5	69,5 ± 16,53	**81,4 ± 21,56**	70,3 ± 18,16	70,1 ± 23,41	70,2 ± 20,46	70,3 ± 17,23
2.2	70,7 ± 18,21	73,3 ± 15,90	**78,5 ± 15,79**	66,7 ± 17,26	66,6 ± 19,19	70,3 ± 18,60	68,3 ± 16,22

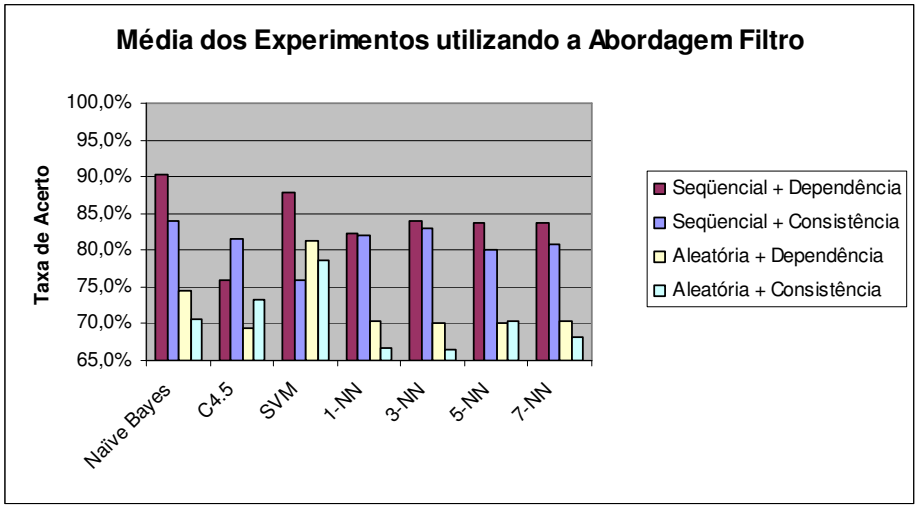

Figura 24: Média das execuções dos métodos de seleção de atributos, utilizando a abordagem filtro, nos cinco subconjuntos de atributos das bases de dados.

4.2.2.2 Abordagem *Wrapper*

Utilizando a abordagem *wrapper* obtiveram-se os seguintes resultados sobre os subconjuntos de dados conforme mostra a figura 25.

As tabelas 12, 13, 14, 15 e 16 mostram os resultados obtidos pelo método de busca seqüencial (1) e método de busca aleatória (2) em cada uma das bases de dados.

A tabela 12 apresenta os resultados obtidos na base de dados DLBCL, onde se observa que os experimentos que se utilizou o método de busca 1, o resultado do

algoritmo C4.5 foi pior comparado com o algoritmo *Naïve Bayes* e o algoritmo 7-NN teve resultado estatisticamente melhor. Já no método de busca 2, apenas o algoritmo SVM teve resultado estatisticamente equivalente ao algoritmo base *Naïve Bayes*, o restante dos resultados foram estatisticamente piores.

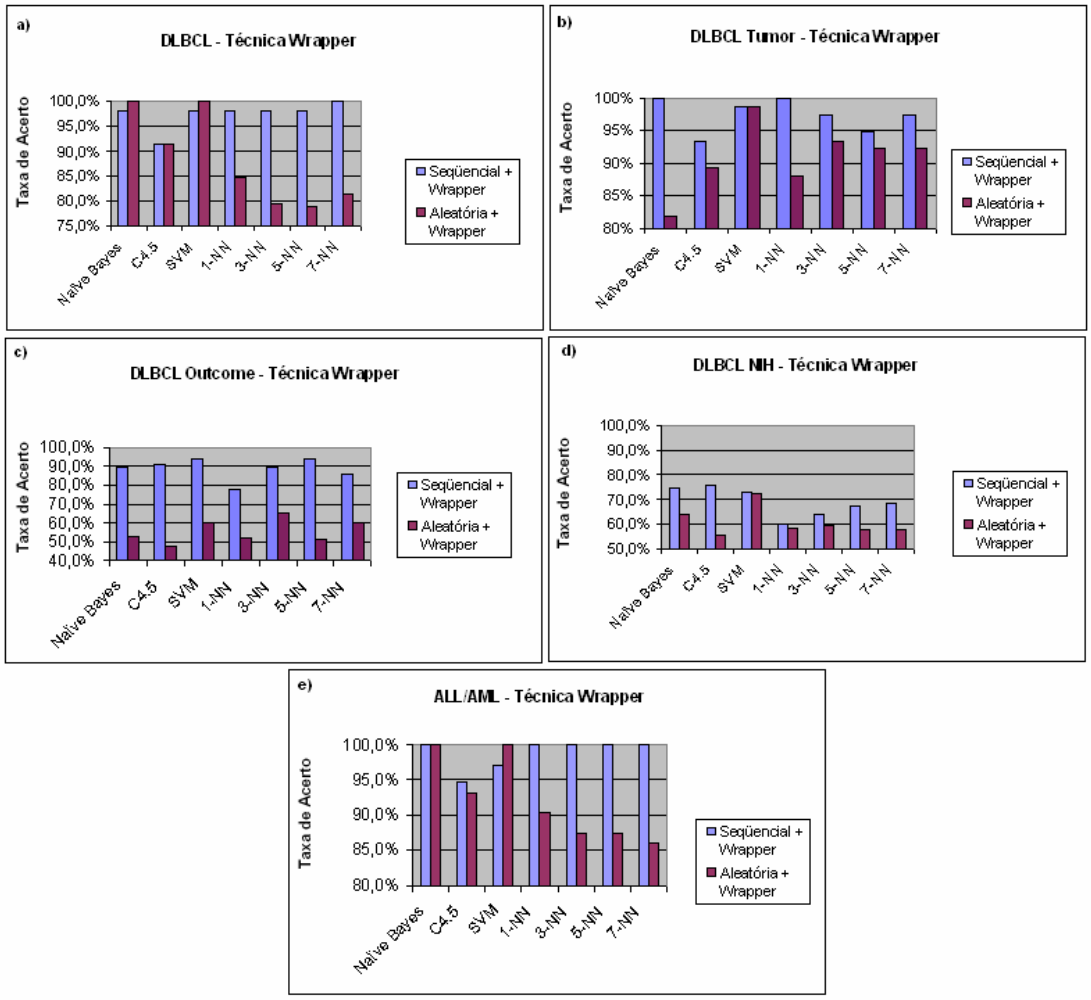

Figura 25: Resultado da Classificação dos Subconjuntos de Atributos utilizando a Abordagem *Wrapper.*

Tabela 12: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados DLBCL após a seleção de Atributos utilizando a Abordagem *Wrapper.*

	Algoritmos						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	98 ± 6,32	91,5 ± 11,07*	98,0 ± 6,32	98 ± 6,32	98 ± 6,32	98 ± 6,32	**100 ± 0**
2	100 ± 0	91,5 ± 11,07*	100 ± 0	84,5 ± 14,42*	79,5 ± 16,41*	79 ± 20,25*	81,5 ± 17,65*

A tabela 13 mostra os resultados obtidos na base de dados DLBCL-Tumor. No método de busca 1 somente o resultado do algoritmo 1-NN foi estatisticamente equivalente ao *Naïve Bayes*. Os outros resultados obtidos pelos demais algoritmos foram considerados estatisticamente piores. No método de busca 2 os resultados avaliados mostram que o algoritmo SVM apresentou melhor resultado estatisticamente e o 1-NN o pior resultado estatisticamente.

Tabela 13: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados DLBCL-Tumor após a seleção de Atributos utilizando a Abordagem *Wrapper.*

	Algoritmos						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	100 ± 0	93,4 ± 11,36*	98,75 ± 3,95*	100 ± 0	97,3 ± 5,66*	94,8 ± 6,71*	97,3 ± 5,66*
2	81,9 ± 10,7	89,3 ± 12,57	**98,75 ± 3,95**	88,04 ± 9,93*	93,2 ± 9,85	92,3 ± 8,87	92,3 ± 8,87

Na tabela 14 são apresentados os resultados obtidos da base de dados DLBCL-*Outcome*. No método de busca 1 o algoritmo 1-NN teve resultado estatisticamente pior comparado com o algoritmo base. Já no método de busca 2 o algoritmo 3-NN se destacou com o melhor resultado avaliado estatisticamente.

A tabela 15 mostra os resultados obtidos com os experimentos realizados na base de dado DLBCL-NIH. No método de busca 1 os resultados dos algoritmos SVM, e k-NN, para k=1, k=3,k=5 e k=7) tiveram os piores resultados estatisticamente. Já no método de busca 2 o algoritmo SVM teve o melhor resultado estatisticamente.

A tabela 16 mostra os resultados obtidos na base de dados ALL/AML. Observa-se que estatisticamente os algoritmos C4.5 e SVM tiveram os piores resultados no método de busca 1. Já no método de busca 2 seus resultados foram estatisticamente equivalente ao algoritmo base *Naïve Bayes*. Os algoritmos 1-NN, 3-NN, 5-NN e 7-NN tiveram os piores resultados estatisticamente.

Tabela 14: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados DLBCL-*Outcome* após a seleção de Atributos utilizando a Abordagem *Wrapper.*

	Algoritmos						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	89,67 ± 11,91	90,67 ± 13,68	94 ± 13,5	77,67 ± 11,0*	89,67 ± 11,91	93,33 ± 11,65	86 ± 13,41
2	52,67 ± 21,76	48 ± 19,7	59,67 ± 21,69	52,33 ± 24,95	**65,33 ± 19,45**	51 ± 19,69	60 ± 21,6

Tabela 15: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados DLBCL-NIH após a seleção de Atributos utilizando a Abordagem *Wrapper.*

	Algoritmos						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	75,00 ± 8,33	75,83 ± 7,30	73,3 ± 4,89*	60,42 ± 9,47*	64,17 ± 9,04*	67,08 ± 7,97*	68,33 ± 8,83*
2	64,17 ± 10,24	55,42 ± 7,87*	**72,5 ± 6,86**	58,75 ± 8,88*	59,58 ± 7,36*	57,92 ± 8,21*	55,42 ± 11,79*

Na figura 26 é feito um comparativo da busca seqüencial em cada base de dados. Observa-se que na base de dados DLBCL-NIH os resultados foram muito baixos independentes do algoritmo de classificação.

Na busca aleatória (figura 27) os resultados são ainda mais baixos comparados com a busca seqüencial.

Tabela 16: Resultado da Classificação dos Subconjuntos de Atributos da Base de Dados ALL/AML após a seleção de Atributos utilizando a Abordagem *Wrapper.*

	Algoritmos						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	100,00 ± 0	94,64 ± 9,11*	97,14 ± 6,02*	100 ± 0	100 ± 0	100 ± 0	100 ± 0
2	100,00 ± 0	93,21 ± 7,18	100 ± 0	90,36 ± 14,68*	87,5 ± 14,19*	87,5 ± 14,19*	86,07 ± 13,49*

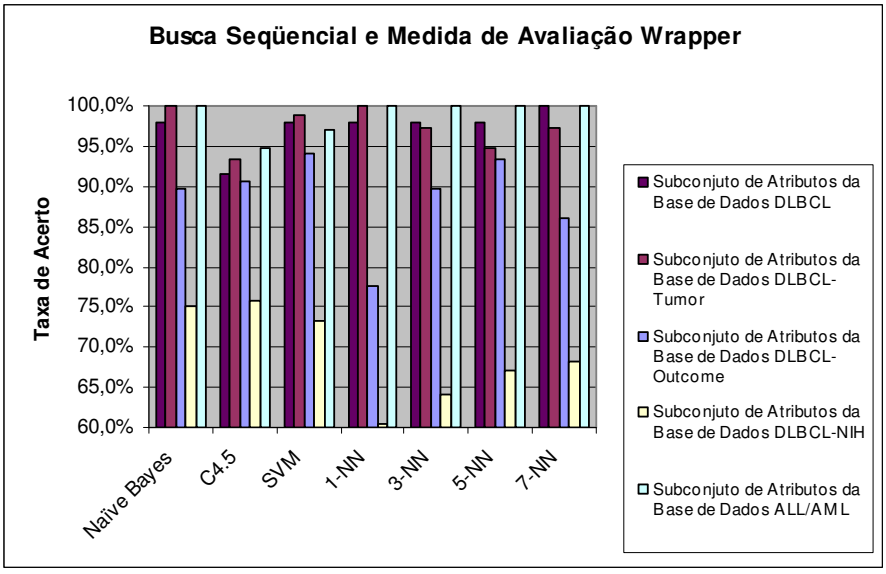

Figura 26: Taxa de acerto dos classificadores sobre os subconjuntos de atributos das bases de dados em que se utilizou a busca seqüencial e a medida de avaliação de cada classificador (*Wrapper*).

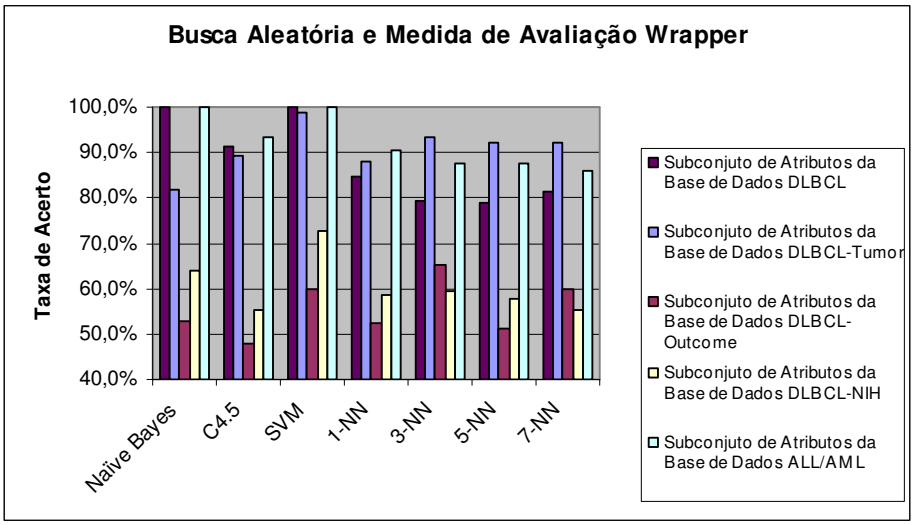

Figura 27: Taxa de acerto dos classificadores sobre os subconjuntos de atributos das bases de dados em que se utilizou a busca aleatória e a medida de avaliação de cada classificador (*Wrapper*).

Calculando a média de todas as execuções dos experimentos (tabela 17), observa-se que no método de busca 1 os algoritmos C4.5 e o 1-NN apresentaram estatisticamente os piores resultados comparado com algoritmo base. No método de busca 2, o algoritmo SVM apresentou estatisticamente o melhor resultado e o algoritmo 5-NN o pior resultado.

Comparando os dois tipos de busca utilizados, seqüencial e aleatória, observa-se que a busca seqüencial teve resultados melhores na abordagem *wrapper* e isso fica visível quando aplicado à média dos resultados sobre as execuções de subconjunto dessa abordagem (figura 28).

Tabela 17: Média das execuções dos métodos de seleção de atributos, utilizando a abordagem *wrapper*, nos cinco subconjuntos de atributos das bases de dados.

	Algoritmos						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	93 ± 10,69	89,2 ± 7,64*	92,2 ± 10,74	87,2 ± 17,69*	89,8 ± 14,87	90,6 ± 13,43	90,3 ± 13,59
2	79,7 ± 21,22	75,5 ± 21,91	**86,2 ± 18,91**	74,8 ± 17,85	77 ± 14,31	73,5 ± 18,23*	75,1 ± 16,37

As figuras 29, 30, 31, 32 e 33 fornecem uma comparação da taxa de acerto dos classificadores quando usados na base de dados original (com todos os atributos) e quando aplicados nos melhores e piores casos.

O uso da seleção de atributos melhorou a taxa de acerto do classificador em todos os subconjuntos de atributos. A sua taxa de acerto é claramente superior nos melhores casos, independente do algoritmo de classificação que está sendo usado. Nos piores casos, os resultados são muito próximos de quando aplicados na base de dados onde todos os atributos são usados. Dessa forma, concluímos que o uso da seleção de atributos é um pré-processamento aconselhável para dados de bioinformática.

.

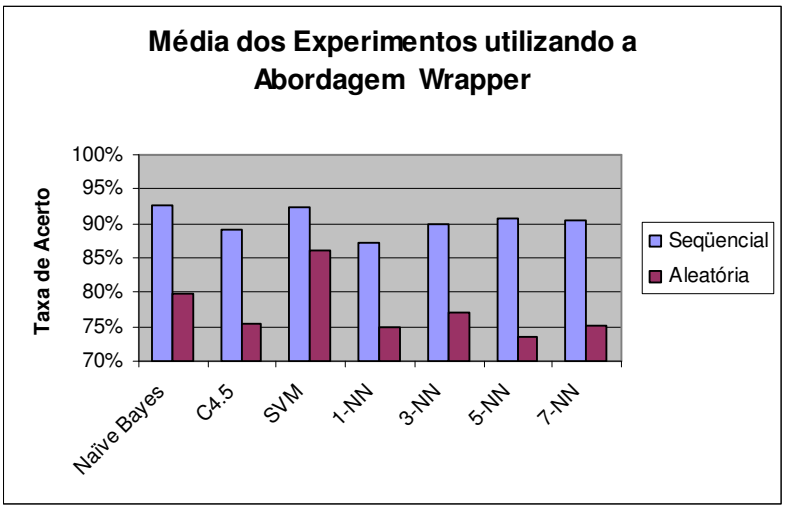

Figura 28: Média das execuções dos métodos de seleção de atributos, utilizando a abordagem *wrapper*, nos cinco subconjuntos de atributos das bases de dados.

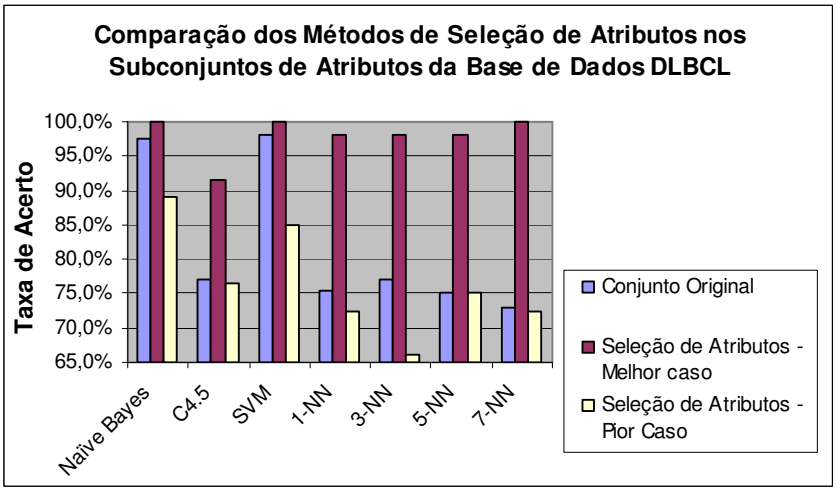

Figura 29: Comparação dos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL no Melhor Caso e no Pior Caso.

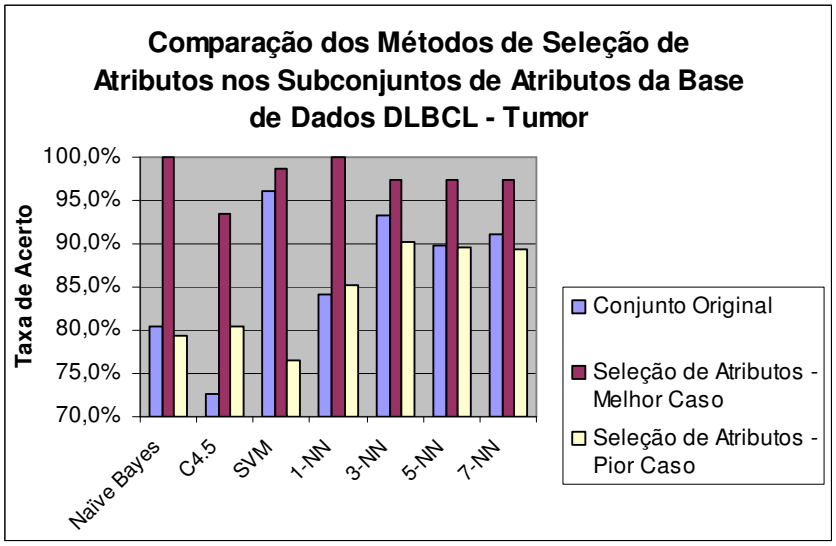

Figura 30: Comparação dos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base DLBCL - Tumor no Melhor Caso e no Pior Caso.

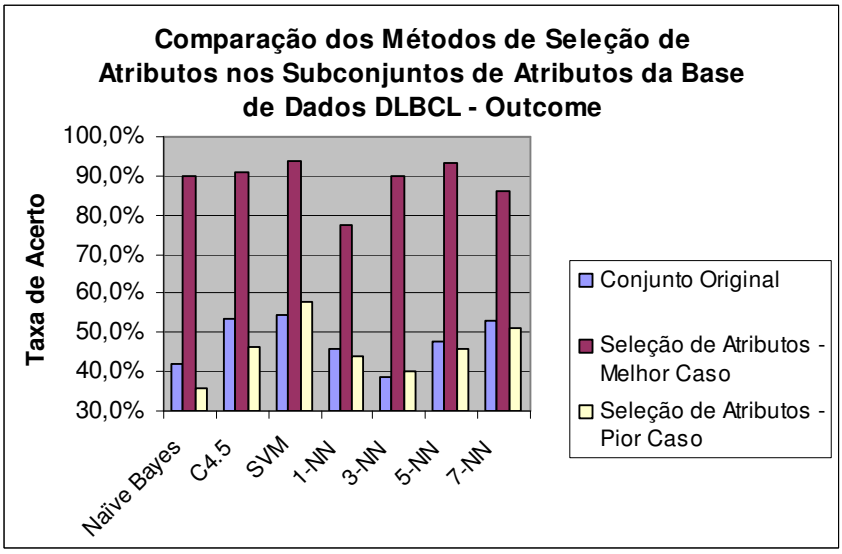

Figura 31: Comparação dos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base DLBCL - Outcome no Melhor Caso e no Pior Caso.

Uma análise mais detalhada dos atributos selecionados (genes) revela que alguns deles forma selecionados na maioria das vezes. Essa sugestão pode ser usada para procurar a sua implicação na doença.

Alguns desses genes mais selecionados são mostrados na tabelas que estão no apêndice A

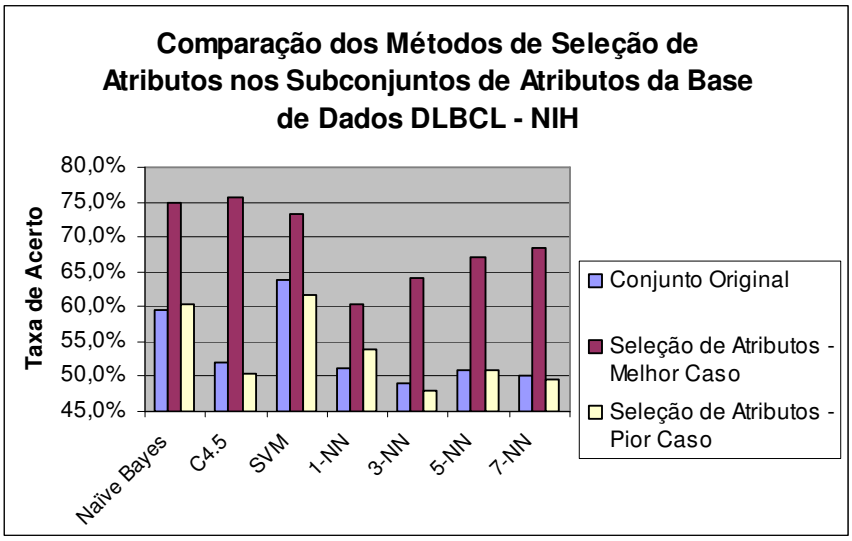

Figura 32: Comparação dos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base DLBCL - NIH no Melhor Caso e no Pior Caso.

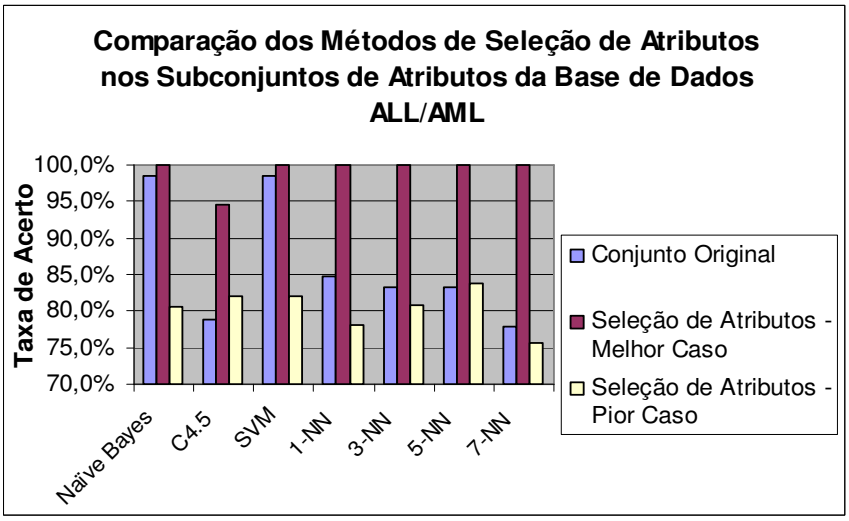

Figura 33: Comparação dos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base ALL/AML no Melhor Caso e no Pior Caso.

4.3 Resultado do Método de Projeção Aleatória sobre as Bases de Dados

Como já descrito na metodologia foi aplicado o método de projeção aleatória nas cinco bases de dados. Para cada número de atributos escolhido executou-se dez vezes o método de projeção aleatória para obter um melhor resultado. Após todas as execuções, tirou-se a média do resultado de cada subconjunto gerado e estes são os resultados apresentados.

Os resultados de cada uma das bases de dados estão divididos em duas tabelas, uma delas mostra os resultados obtidos quando se utilizou um número fixo de atributos para a formação do novo subconjunto de atributos e a outra quando se utilizou uma porcentagem de atributos.

As tabelas 18 e 19 mostram os resultados dos subconjuntos de atributos da base de dados DLBCL. As tabelas 20 e 21 os resultados dos subconjuntos de atributos da base de dados DLBCL-Tumor. As tabelas 22 e 23 os resultados dos subconjuntos de atributos da base de dados DLBCL-Outcome. A tabela 24 e 25 os resultados dos subconjuntos de atributos da base de dados DLBCL-NIH e as tabelas 26 e 27 os resultados dos subconjuntos de atributos da base de dados ALL/AML.

No subconjunto de atributos formado por 10 atributos, estatisticamente os resultados dos algoritmos C4.5, SVM e *k*-NN para *k*=7 são equivalentes ao resultado do algoritmo base *Naïve Bayes*. Para os subconjuntos com 15 atributos e 30 atributos o algoritmo SVM teve resultado melhor comparado com o algoritmo *Naïve Bayes* e o algoritmo 1-NN resultado pior. O algoritmo *k*-NN, para todos os valores de *k* utilizado, teve resultados estatisticamente inferiores comparados com o algoritmo *Naïve Bayes* nos experimentos que se utilizou um subconjunto formado por 45 atributos e 71 atributos.

Tabela 18: Resultado do Método de Projeção Aleatória na Base de Dados DLBCL quando utilizado um número fixo de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 atributos	73,35 ± 9,92	65,25± 8,09	76,40 ± 8,92	66,65 ± 10,59*	66,90 ± 9,79*	68,35 ± 11,36*	69,90 ± 10,13
15 atributos	71,70 ± 10,58	69,35 ± 12,71	**79,70 ± 9,95**	68,70 ± 9,20*	69,80 ± 10,93	74,25 ± 8,87	71,05 ± 11,97
30 atributos	75,25 ± 8,52	70,45 ± 12,73	82,00 ± 8.49	69,95 ± 9,39*	73,20 ± 7,77	71,55 ± 9,69	76,00 ± 11,27
45 atributos	83,40 ± 5,64	76,10 ± 11,54	87,20 ± 4,67	70,15 ± 4,56*	75,15 ± 5,85*	75,50 ± 7,80*	76,75 ± 5,57*
71 atributos	86,25 ± 6,52	75,05 ± 10,92*	86,15 ± 4,66	73,75 ± 7,42*	75,50 ± 8,71*	78,55 ± 7,71*	79,30 ± 8,26*

Analisando estatisticamente a tabela 19, observa-se que apenas o algoritmo SVM teve resultados equivalentes ao algoritmo *Naïve Bayes* em todos os experimentos. Os resultados obtidos pelos outros algoritmos são piores estatisticamente.

Se compararmos o resultado da base de dados original como o resultado do método de projeção aleatória, observaremos o método de projeção aleatória apresentou melhores resultado em várias execuções. Na figura 34, onde é mostrada a comparação entre a base com todos os atributos e a aplicação do método de projeção aleatória utilizando um número fixo de atributos, apenas na execução do conjunto com 10 e 30 atributos, o resultado do conjunto de dados com todos os atributos foi melhor. Já quando se aplicou a projeção aleatória utilizando uma porcentagem de atributos, figura 35, além

de se obter melhores resultados, tanto na comparação entre o número fixo de atributos como na comparação da base de dados com todos os atributos.

Tabela 19: Resultado do Método de Projeção Aleatória na Base de Dados DLBCL quando utilizado a porcentagem de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% atributos	90,35 ± 4,12	79,70 ± 5,08*	90,60 ± 5,61	76,00 ± 5,57*	79,10 ± 4,50*	80,80 ± 6,93*	80,70 ± 9,25*
10% atributos	94,05 ± 2,85	72,60 ± 6,99*	92,95 ± 2,31	81,00 ± 5,33*	86,35 ± 3,50*	86,95 ± 3,00*	84,00 ± 5,88*
20% atributos	94,00 ± 2,22	72,05 ± 7,54*	93,70 ± 3,57	80,70 ± 4,14*	86,55 ± 3,95*	85,85 ± 1,89*	87,80 ± 3,17*
25% atributos	94,35 ± 3,16	73,25 ± 10,59*	93,20 ± 3,24	80,30 ± 3,50*	85,50 ± 2,52*	86,95 ± 2,15*	86,25 ± 3,33*
50% atributos	94,80 ± 1,87	78,90 ± 12,33*	94,35 ± 3,06	82,00 ± 2,25*	85,45 ± 2,19*	87,35 ± 1,72*	87,00 ± 2,48*

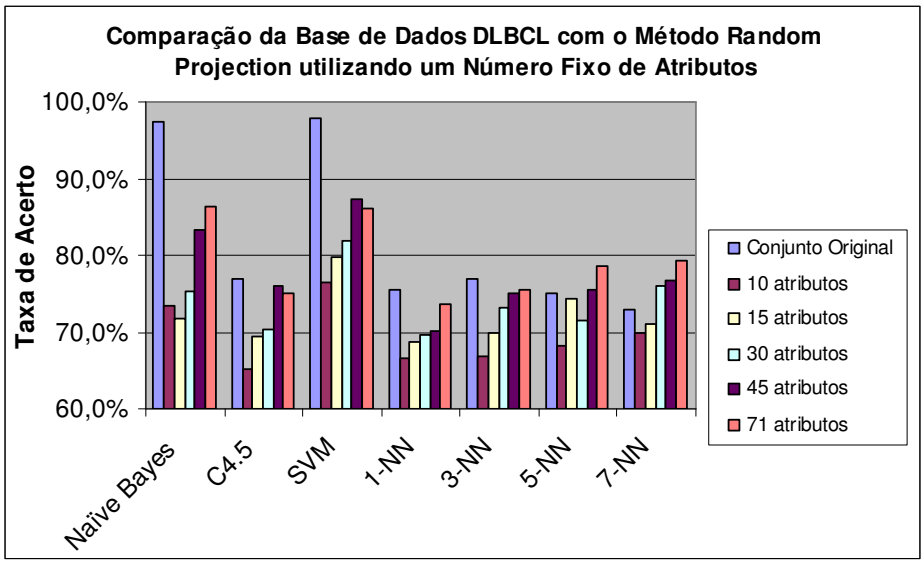

Figura 34: Comparação do Resultado da Base de Dados DLBCL com o Método Projeção Aleatória utilizando um número fixo de atributos para a formação do subconjunto de atributos.

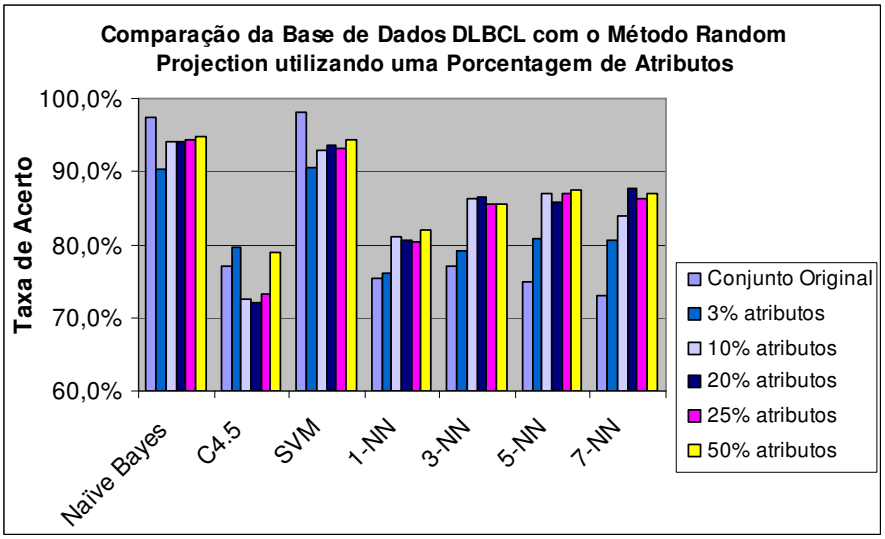

Figura 35: Comparação do Resultado da Base de Dados DLBCL com o Método de Projeção Aleatória utilizando uma porcentagem de atributos para a formação do subconjunto de atributos.

Analisando os resultados obtidos pelo método de projeção aleatória, na base de dados DLBCL-Tumor, os resultados dos algoritmos nos experimentos com 10 e 30 atributos são considerados equivalentes ao algoritmo base *Naïve Bayes*. No subconjunto formado por 15 atributos o algoritmo 7-NN teve resultado estatisticamente melhor que aos demais algoritmos. Já nos experimentos com 45 e 71 atributos o algoritmo SVM apresentou estatisticamente melhores resultados e os algoritmos C4.5 e 3-NN os piores resultados comparados com o algoritmo *Naïve Bayes* (tabela 20).

Tabela 20: Resultado do Método de Projeção Aleatória na Base de Dados DLBCL - Tumor quando utilizado um número fixo de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 atributos	81,27 ± 5,87	79,56 ± 6,42	78,45 ± 3,79	79,05 ± 8,09	81,86 ± 5,36	83,02 ± 5,25	82,25 ± 6,49
15 atributos	82,20 ± 5,19	79,07 ± 5,42	81,95 ± 5,47	81,36 ± 6,69	82,77 ± 6,18	82,63 ± 5,01	**82,91 ± 6,79**
30 atributos	86,02 ± 5,93	82,28 ± 4,82	87,89 ± 4,47	83,53 ± 5,43	82,41 ± 5,47	84,00 ± 5,64	85,77 ± 4,21
45 atributos	88,36 ± 4,00	82,27 ± 4,60*	**92,61 ± 3,77**	86,77 ± 4,11	85,29 ± 4,44*	87,12 ± 4,72	87,38 ± 4,57
71 atributos	88,59 ± 3,04	83,57 ± 5,14*	**94,62 ± 2,53**	87,55 ± 3,10	85,54 ± 3,75*	88,07 ± 3,98	88,59 ± 4,28

Nos experimentos feitos em que se utilizou uma porcentagem para a formação dos subconjuntos de atributos, o algoritmo SVM teve resultados estatisticamente melhor, em todos os casos, comparado com o algoritmo base *Naïve Bayes*. O 7-NN também teve resultados estatisticamente melhores em quase todos os experimentos. Além disso, observa-se estatisticamente que o algoritmo 3-NN teve resultados piores em três

experimentos realizados, nos subconjuntos com 10% de atributos, 20% e 50% e o algoritmo C4.5 no subconjunto com 25% de atributos (tabela 21).

Tabela 21: Resultado do Método de Projeção Aleatória na Base de Dados DLBCL - Tumor quando utilizado a porcentagem de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% atributos	88,55 ± 1,97	83,36 ± 5,66*	**95,87 ± 1,40**	87,84 ± 2,30	86,87 ± 1,79	88,53 ± 1,78	**90,32 ± 2,23**
10% atributos	88,57 ± 2,31	85,84 ± 6,30	**96,11 ± 1,27**	87,73 ± 2,21	86,28 ± 1,77*	90,37 ± 1,62	89,98 ± 1,93
20% atributos	88,84 ± 1,69	85,21 ± 6,27	**96,79 ± 1,12**	87,86 ± 1,57	86,89 ± 1,51*	90,48 ± 1,72	**91,98 ± 2,33**
25% atributos	87,93 ± 3,10	82,86 ± 6,86*	**96,50 ± 1,29**	87,98 ± 2,10	87,41 ± 2,34	90,12 ± 1,24	**91,52 ± 2,18**
50% atributos	88,77 ± 1,41	86,61 ± 4,46	**96,84 ± 0,99**	88,17 ± 1,38	86,51 ± 0,93*	**90,97 ± 1,05**	**91,13 ± 1,55**

Fazendo também uma comparação entre a base de dados original e os subconjuntos de atributos gerados pelo método de projeção aleatória, observa-se que em alguns casos a taxa de acerto do classificador foi melhor no conjunto original comparado com os subconjuntos formados pelo um número fixo de atributos, isso pode ser observado na figura 36. Na figura 37 e mostrada a comparação entre a utilização do método de projeção aleatória utilizando uma porcentagem de atributos e a base de dados original. Nota-se que em praticamente todos os casos que se utilizou o método de projeção aleatória o resultado dos algoritmos de classificação foram superiores.

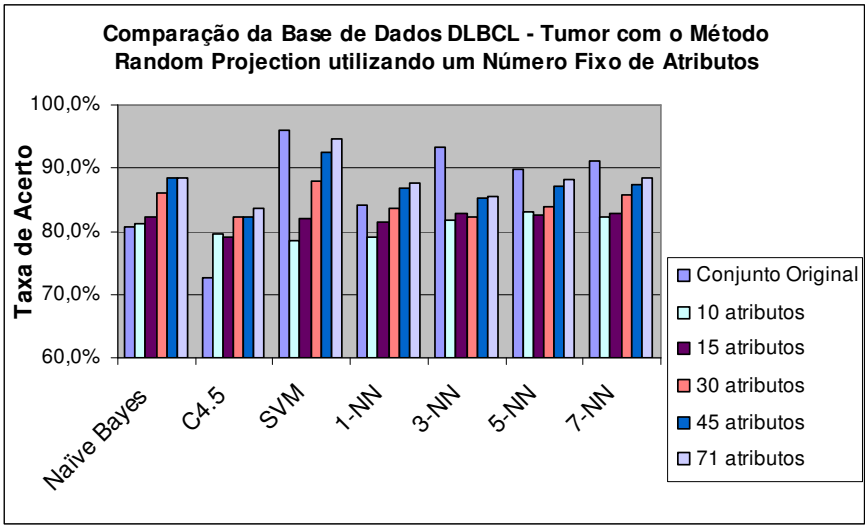

Figura 36: Comparação do Resultado da Base de Dados DLBCL - Tumor com o Método Projeção Aleatória utilizando um número fixo de Atributos para a formação do subconjunto de atributos.

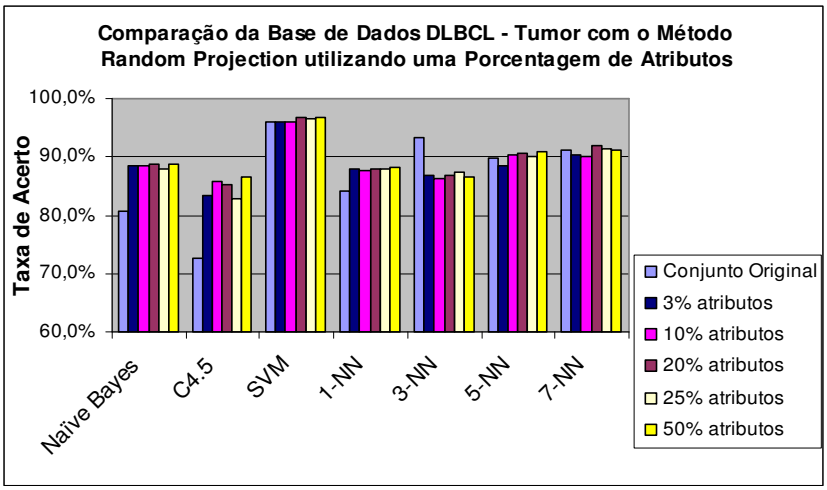

Figura 37: Comparação do Resultado da Base de Dados DLBCL - Tumor com o Método de Projeção Aleatória utilizando uma porcentagem de atributos para a formação do subconjunto de atributos.

Analisando os resultados da base de dados DLBCL-Outcome quando se utilizou o método de projeção aleatória tanto com um número fixo de atributos (tabela 22) como também a porcentagem de atributos (tabela 23), observa-se que estatisticamente quase todos os resultados dos algoritmos foram considerados melhores que o algoritmo base *Naïve Bayes*.

Tabela 22: Resultado do Método de Projeção Aleatória na Base de Dados DLBCL - Outcome quando utilizado um número fixo de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 atributos	45,50 ± 7,34	**53,00 ± 7,41**	**52,80 ± 6,24**	**53,77 ± 9,40**	**55,07 ± 6,66**	**53,47 ± 6,12**	**52,37± 7,41**
15 atributos	45,53 ± 8,11	51,57 ± 5,38	50,63 ± 4,28	**53,90 ± 7,14**	**53,33 ± 4,69**	**52,07 ± 5,57**	**52,47 ± 7,10**
30 atributos	43,37 ± 7,89	47,57 ± 6,28	**48,47 ± 9,77**	**56,30 ± 8,14**	**54,27 ± 5,59**	**51,07 ± 6,58**	48,33 ± 7,11
45 atributos	41,77 ± 4,52	43,07 ± 5,34	**47,07 ± 6,30**	**52,27 ± 4,51**	**54,13 ± 6,54**	**52,43 ± 7,07**	**51,43 ± 7,15**
71 atributos	40,50 ± 6,15	**48,97 ± 7,61**	**51,57 ± 5,88**	**52,07 ± 6,50**	**55,00 ± 4,99**	**51,70 ± 7,76**	**51,00 ± 4,83**

Tabela 23: Resultado do Método de Projeção Aleatória na Base de Dados DLBCL - Outcome quando utilizado a porcentagem de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% atributos	39,73 ± 6,53	45,70 ± 8,59	**52,77 ± 7,15**	**57,10 ± 5,45**	**55,54 ± 4,89**	**55,37 ± 6,99**	**52,23 ± 5,66**
10% atributos	38,80 ± 4,63	**48,37 ± 9,33**	**51,00 ± 2,81**	**54,50 ± 1,71**	**56,57 ± 4,80**	**53,14 ± 3,86**	**49,83 ± 4,54**
20% atributos	39,07 ± 3,56	**48,97 ± 8,46**	**52,00 ± 1,61**	**53,70 ± 2,68**	**57,90 ± 3,04**	**54,24 ± 2,87**	**51,57 ± 2,47**
25% atributos	37,63 ± 4,46	46,83 ± 14,91	**52,30 ± 2,74**	**53,10 ± 2,90**	**57,57 ± 2,34**	**54,47 ± 3,92**	**50,10 ± 3,14**
50% atributos	36,60 ± 2,92	**47,70 ± 8,85**	**52,00 ± 2,25**	**51,63 ± 2,10**	**56,73 ± 2,52**	**52,23 ± 2,73**	**50,50 ± 3,3**

Também é possível observar (figuras 38 e 39) que quase todos os resultados dos algoritmos quando aplicado sobre o método de projeção aleatória foram melhores que os resultados obtidos na base de dados original, apesar de ambos os resultados serem relativamente baixos.

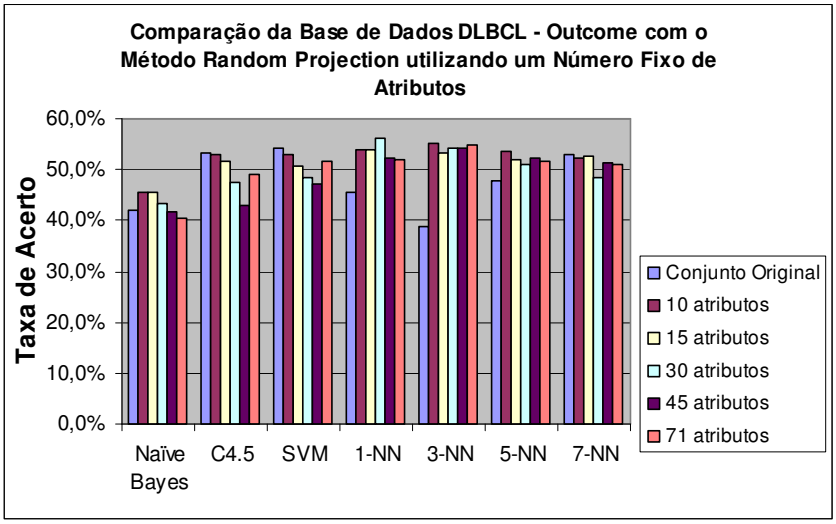

Figura 38: Comparação do Resultado da Base de Dados DLBCL - Outcome com o Método de Projeção Aleatória utilizando um número fixo de atributos para a formação do subconjunto de atributos.

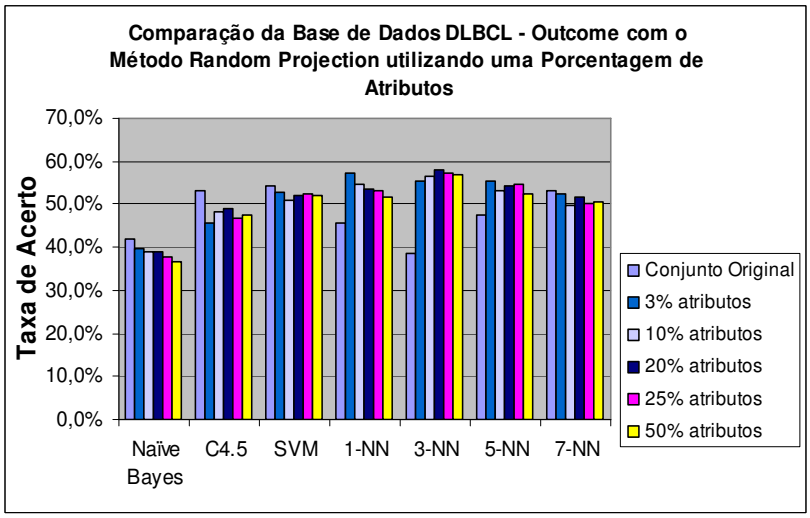

Figura 39: Comparação do Resultado da Base de Dados DLBCL – Outcome com o Método de Projeção Aleatória utilizando uma porcentagem de atributos para a formação do subconjunto de atributos.

Analisando os resultados do método de projeção aleatória na base DLBCL-NIH quando se utilizou um número fixo de atributos (tabela 24) observa-se que, estatisticamente todos os resultados obtidos no subconjunto com 10 atributos são estatisticamente equivalentes ao algoritmo base *Naïve Bayes*. Os resultados obtidos no subconjunto com 15 atributos os algoritmos 3-NN e 5-NN são considerados estatisticamente piores. Da mesma forma isso acontece no subconjunto com 30 atributos com o algoritmo 7-NN, no subconjunto com 45 atributos com o algoritmo 5-NN e no subconjunto com 71 atributos com os algoritmos C4.5 e 1-NN.

Quando se utilizou uma porcentagem de atributos (tabela 25) observa-se que, o algoritmo C4.5 e 1-NN tiveram seus resultados estatisticamente piores comparado com algoritmo *Naïve Bayes* em todos os casos. Os algoritmos 5-NN e 7-NN também apresentaram resultados estatisticamente piores em alguns experimentos. Já o algoritmo SVM apresentou resultados estatisticamente melhores quando se aplicou nos subconjuntos com 25% e 50% de atributos.

Tabela 24: Resultado do Método de Projeção Aleatória na Base de Dados DLBCL - NIH quando utilizado um número fixo de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 atributos	55,63 ± 2,06	56,67 ± 2,17	56,84 ± 0,97	54,17 ± 3,35	54,17 ± 3,29	54,83 ± 3,06	56,34 ± 2,82
15 atributos	56,54 ± 2,77	56,50 ± 2,85	56,54 ± 1,18	54,67 ± 3,93	53,71 ± 2,35*	53,38 ± 2,16*	54,04 ± 3,72
30 atributos	55,63 ± 2,52	54,54 ± 3,29	55,87 ± 4,09	52,58 ± 4,20	54,38 ± 3,30	53,29 ± 3,19	53,12 ± 2,95*
45 atributos	56,67 ± 3,25	54,42 ± 1,77	55,96 ± 4,42	54,50 ± 1,98	54,42 ± 1,67	54,42 ± 2,93*	56,12 ± 2,76
71 atributos	58,54 ± 2,65	52,67 ± 3,46*	56,58 ± 5,67	55,21 ± 2,73*	56,67 ± 3,22	56,38 ± 2,48	56,00 ± 2,76

Tabela 25: Resultado do Método de Projeção Aleatória na Base de Dados DLBCL – NIH quando utilizado a porcentagem de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% atributos	59,83 ± 1,98	55,33 ± 4,61*	59,88 ± 3,43	56,71 ± 2,45*	58,71 ± 1,91	57,21 ± 1,57*	58,79 ± 1,81
10% atributos	59,79 ± 1,76	52,08 ± 2,61*	62,29 ± 3,96	56,50 ± 1,67*	59,54 ± 3,20	59,00 ± 1,79	58,21 ± 2,31
20% atributos	60,50 ± 2,46	54,87 ± 4,31*	62,50 ± 3,90	56,63 ± 1,22*	60,33 ± 2,37	59,00 ± 2,21*	57,87 ± 2,12*
25% atributos	59,87 ± 1,55	54,58 ± 2,48*	**63,38 ± 3,23**	57,46 ± 1,76*	60,33 ± 2,11	58,84 ± 2,65	57,63 ± 2,95*
50% atributos	59,87 ± 1,08	55,46 ± 2,52*	**64,17 ± 2,42**	57,17 ± 1,36*	61,12 ± 2,20	57,75 ± 1,99*	57,75 ± 1,91*

As figuras 40 e 41 mostram a comparação de resultados feita entre a aplicação do método de projeção aleatória e a base de dados original DLBCL-NIH, em que a utilização do método de projeção aleatória melhorou a taxa de acerto do classificador.

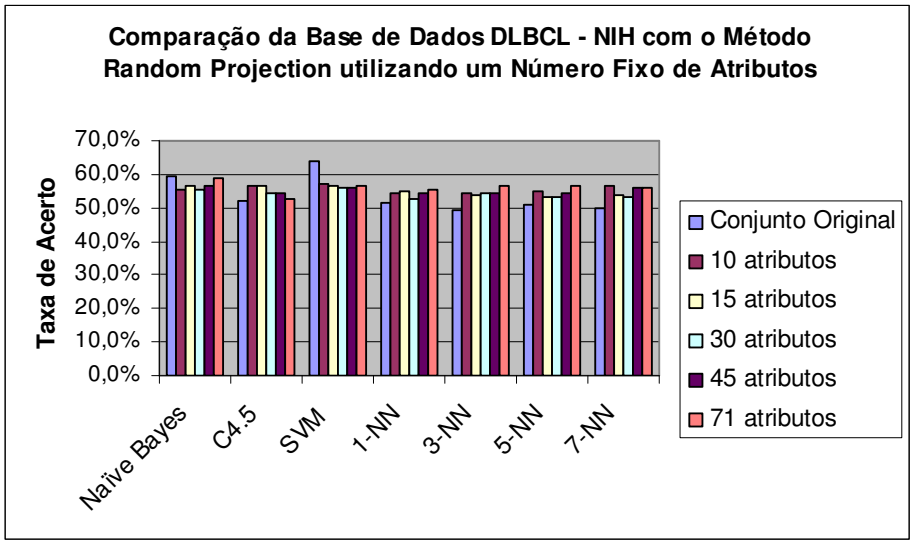

Figura 40: Comparação do Resultado da Base de Dados DLBCL – NIH com o Método de Projeção Aleatória utilizando um número fixo de atributos para a formação do subconjunto de atributos.

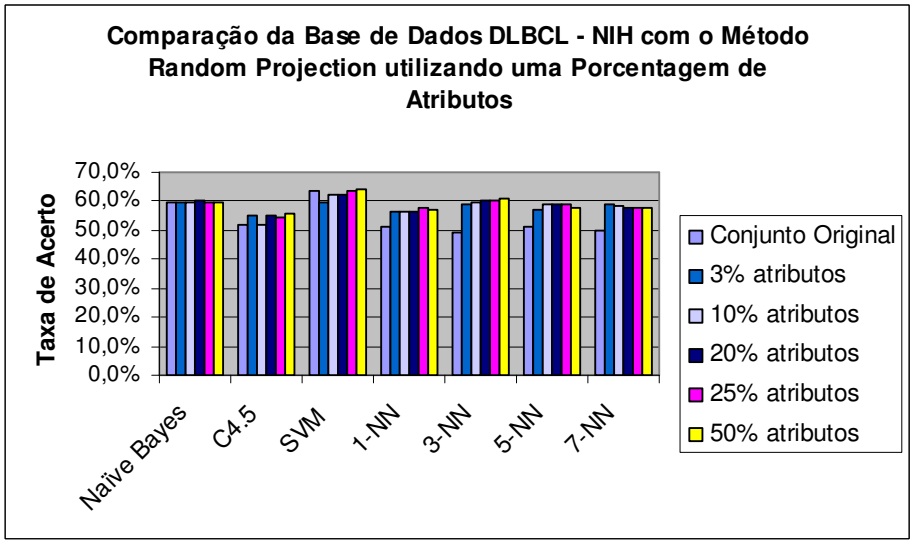

Figura 41: Comparação do Resultado da Base de Dados DLBCL – NIH com o Método de Projeção Aleatória utilizando uma porcentagem de atributos para a formação do subconjunto de atributos.

Analisando os resultados obtidos pela execução do método de projeção aleatória na base de dados ALL/AML quando utilizado um número fixo de atributos (tabela 26), observa-se que estatisticamente o algoritmo C4.5 apresentou os piores resultados em todos os experimentos comparados com o algoritmo *Naïve Bayes*.

Nota-se ainda que o algoritmo 1-NN teve resultados estatisticamente pior nos subconjuntos com 10 e 45 atributos, o algoritmo 3-NN no subconjunto com 71 atributos, o algoritmo 5-NN nos subconjuntos com 45 e 71 atributos e o algoritmo 7-NN nos subconjuntos com 15, 30, 45 e 71 atributos.

Na tabela 27 pode ser observado os resultados obtidos quando se utilizou a porcentagem de atributos. O algoritmo C4.5 e o 7-NN teve resultados estaticamente piores em todos os experimentos. Também é possível observar que no subconjunto formado por 3% de atributos os algoritmo 1-NN, 3-NN e 5-NN tiveram seus resultados considerados estaticamente piores comparados com o algoritmo base *Naïve Bayes*. O mesmo também aconteceu no subconjunto com 20% de atributos com o algoritmo 5-NN.

Observa-se ainda que o algoritmo SVM é considerado estatisticamente melhor em todos os experimentos que se utilizou a porcentagem para a formação do subconjunto na base de dados ALL/AML.

Tabela 26: Resultado do Método de Projeção Aleatória na Base de Dados ALL/AML quando utilizado um número fixo de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 atributos	80,41 ± 7,56	75,02 ± 7,08*	81,29 ± 7,13	76,71 ± 6,98*	77,98 ± 5,68	78,18 ± 4,88	77,55 ±5,22
15 atributos	84,09 ± 6,41	75,18 ± 8,48*	84,11 ± 6,68	80,89 ± 4,70	80,68 ± 5,45	80,54 ± 5,47	80,11 ± 4,78*
30 atributos	86,80 ± 5,09	74,38 ± 5,22*	87,68 ± 7,55	85,21 ± 4,06	86,52 ± 2,74	84,77 ± 3,77	82,30 ± 2,97*
45 atributos	89,04 ± 4,94	73,27 ± 6,65*	**93,34 ± 2,88**	85,27 ± 2,63*	87,29 ± 3,29	85,95 ± 2,84*	84,93 ± 2,87*
71 atributos	90,50 ± 2,56	78,59 ± 3,39*	**95,30 ± 3,27**	89,20 ± 3,52	88,84 ± 2,02*	87,12 ± 1,78*	85,73 ± 3,11*

Tabela 27: Resultado do Método de Projeção Aleatória na Base de Dados ALL/AML quando utilizado a porcentagem de atributos para a formação do subconjunto de atributos.

	Algoritmos						
Nº Atributos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% atributos	91,45 ± 2,42	80,48 ± 5,97*	**96,30 ± 1,80**	89,29 ± 2,31*	89,47 ± 1,84*	88,77 ± 2,65*	88,07 ± 1,91*
10% atributos	93,00 ± 2,65	83,16 ± 5,40*	**97,14 ± 1,17**	91,88 ± 1,77	91,95 ± 2,31	90,82 ± 1,62	90,06 ± 1,65*
20% atributos	93,43 ± 2,15	81,14 ± 6,62*	**97,86 ± 1,01**	91,86 ± 1,54	93,04 ± 1,00	91,11 ± 1,59*	90,63 ± 1,57*
25% atributos	93,00 ± 2,07	80,84 ± 7,32*	**97,86 ± 0,75**	91,84 ± 1,26	92,90 ± 1,12	91,29 ± 1,88	90,40 ± 1,36*
50% atributos	92,43 ± 1,66	82,34 ± 6,92*	**97,86 ± 0,75**	92,54 ± 1,58	93,15 ± 0,60	92,20 ± 1,21	90,13 ± 1,66*

Nas figuras 42 e 43 é feita uma comparação da taxa da de classificação dos algoritmos sobre a base de dados ALL/AML original e os subconjuntos de atributos gerados pelo método de projeção aleatória.

É possível observar que na maioria dos casos a utilização do método de projeção aleatória ajudou a melhorar a taxa de acerto, principalmente quando se utilizou a porcentagem de atributos para a formação do subconjunto de atributos (figura 43).

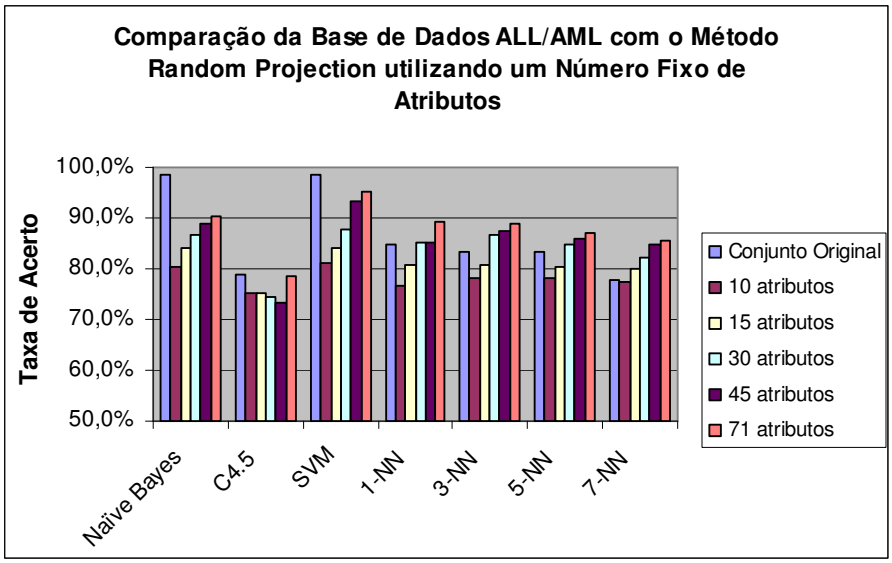

Figura 42: Comparação do Resultado da Base de Dados ALL/AML com o Método de Projeção Aleatória utilizando um número fixo de atributos para a formação do subconjunto de atributos.

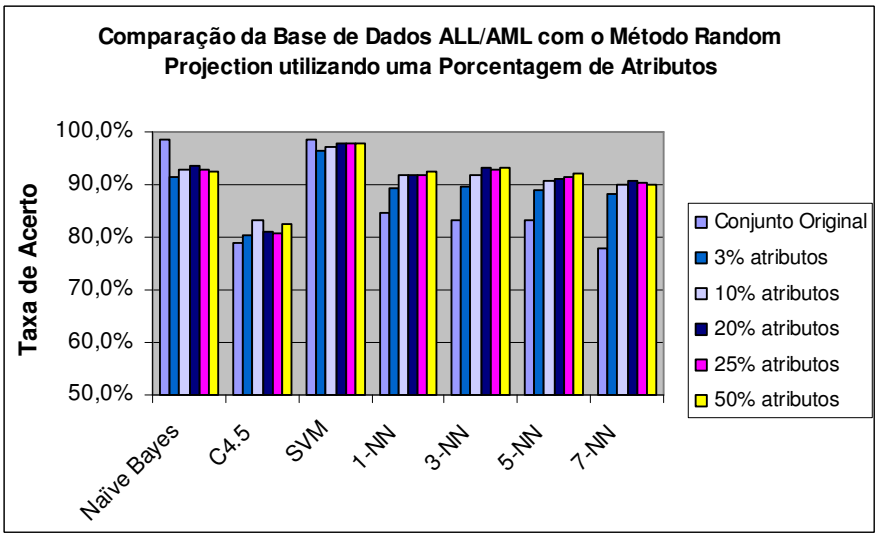

Figura 43: Comparação do Resultado da Base de Dados ALL/AML com o Método de Projeção Aleatória utilizando uma porcentagem de atributos para a formação do subconjunto de atributos.

Em geral, algumas conclusões podem ser tiradas dos experimentos que se utilizou o método de projeção aleatória. Primeiramente, o algoritmo SVM foi o que

apresentou melhores resultados estatisticamente significantes e em nenhum dos experimentos foi considerado estatisticamente pior que o algoritmo base *Naïve Bayes*.

4.4 Resultado da Utilização Conjunta do Método de Projeção Aleatória com a Seleção de Atributos

4.4.1 Seleção de Atributos

Para a identificação do método de seleção de atributos aplicado será utilizada a mesma descrição da tabela mostrada na figura 4.

Devido a execução dos métodos de seleção de atributos para todos os 1800 subconjuntos de atributos gerados para cada base de dados, calculou-se a média aritmética do número de atributos selecionados por cada método em todos os subconjuntos de atributos de cada base de dados.

Dessa forma, obteve-se a média de cada método de seleção de atributos utilizado de cada subconjunto de atributos gerados pelo método de projeção aleatória. Essas médias são mostradas no Apêndice B.

Calculou-se também a média aritmética também da quantidade de atributos selecionados por cada método de seleção de atributos, pois serão esses os resultados para futuras comparações. A tabela 28 mostra os resultados obtidos quando se utilizou um número fixo de atributos para a formação dos subconjuntos de atributos e a tabela 29 mostra os resultados obtidos quando se utilizou a porcentagem para a formação dos subconjuntos de atributos.

É importante deixar claro que, quando é feita referência aos resultados obtidos pela utilização conjunta dos métodos de redução da dimensionalidade serão esses os dados comparados.

Nota-se que em alguns experimentos não houve atributos selecionados. Isso aconteceu devido à maneira como a escolha dos atributos é feita. Por exemplo, na busca seqüencial para frente, o algoritmo primeiramente testa com um subconjunto vazio. Quando ele adiciona um atributo a este subconjunto e testa dando resultado pior que a iteração anterior o algoritmo pára e a busca por um subconjunto de atributos termina. Isso foi o que aconteceu em alguns casos, principalmente quando se estava aplicando a seleção de atributos em um subconjunto com poucos atributos, ou seja, nos casos em que se utilizou um número fixo de atributos.

Tabela 28: Média Geral de Atributos Selecionados nos Subconjuntos de Atributos das Cinco Bases de Dado utilizando o Método de Projeção Aleatória com um Número Fixo de Atributos.

	Média Geral da Quantidade de Atributos Selecionados nos Subconjuntos de Atributos das Cinco Bases de Dados utilizando o Método *Projeção aleatória* com um Número Fixo de Atributos				
Subconjuntos	DLBCL	DLBCL-Tumor	DLBCL-Outcome	DLBCL-NIH	ALL/AML
1.1	4	7	0	0	8
1.2	2	3	0	0	4
1.3	4	4	3	3	5
1.4	2	2	2	2	3
1.5	5	1	3	0	6
1.6	3	4	2	2	5
1.7	3	4	3	2	5
1.8	3	5	2	2	4
1.9	3	4	2	2	4
2.1	6	8	1	3	8
2.2	10	11	3	6	12
2.3	15	14	9	15	16
2.4	10	11	9	10	10
2.5	15	17	11	16	16
2.6	15	15	12	18	16
2.7	16	15	12	18	16
2.8	16	15	12	17	16
2.9	15	16	11	17	16

Tabela 29: Média Geral de Atributos Selecionados nos Subconjuntos de Atributos das Cinco Bases de Dado utilizando o Método de Projeção Aleatória com a Porcentagem de Atributos.

	Média Geral da Quantidade de Atributos Selecionados nos Subconjuntos de Atributos das Cinco Bases de Dados utilizando o Método *Projeção aleatória* com a Porcentagem de Atributos				
Subconjuntos	DLBCL	DLBCL-Tumor	DLBCL-Outcome	DLBCL-NIH	ALL/AML
1.1	25	52	11	13	35
1.2	3	3	5	11	4
1.3	5	5	4	5	4
1.4	3	3	6	11	3
1.5	4	4	6	5	4
1.6	4	4	3	3	4
1.7	4	4	3	3	4
1.8	4	4	4	4	4
1.9	4	4	4	3	4
2.1	163	313	293	273	310
2.2	83	73	654	587	71
2.3	335	399	399	459	529
2.4	287	477	491	394	473
2.5	306	483	422	507	442
2.6	318	525	368	490	535
2.7	325	373	436	542	476
2.8	342	463	467	474	558
2.9	343	516	392	422	447

4.4.2 Classificação dos Subconjuntos de Atributos

4.4.2.1 Abordagem Filtro

Da mesma forma que foi utilizada para chegar aos resultados da seleção de atributos foi utilizada para chegar ao resultado da classificação. Seria inviável apresentar os 1800 resultados obtidos em cada base de dados. Então primeiramente foi calculada a média aritmética de todas as 10 execuções de cada subconjunto de atributos gerados pelo método de projeção aleatória. Assim obteve-se uma média dos resultados de cada método de seleção de atributos de cada subconjunto de atributos. Após, foi determinada a média novamente de todos os resultados obtidos anteriormente, e são esses os resultados que estão apresentados a seguir.

Na tabela 30 são apresentados os resultados da utilização conjunta do método projeção aleatória e da seleção de atributos na base de dados DLBCL. Observa-se que no subconjunto 1.1 os algoritmos C4.5 e o k-NN, para os valores de k=1, k=3 e k=5, foram considerados estatisticamente piores comparados com algoritmo base *Naïve Bayes*. No subconjunto 1.2 apenas o algoritmo 1-NN é estatisticamente pior que o *Naïve Bayes*, os demais são considerados equivalentes. Já nos subconjuntos 2.1 e 2.2 apenas o algoritmo SVM é considerado equivalente ao algoritmo base o restante é estatisticamente pior.

Fazendo uma comparação entre esses resultados obtidos e o resultado dos algoritmos sobre a base de dados original (figura 44), observa-se que apenas em dois casos, no algoritmo *Naïve Bayes* e no SVM, o resultado sem a aplicação do método de redução da dimensionalidade foi melhor.

Tabela 30: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem Filtro nos Subconjuntos de Atributos da Base de Dados DLBCL.

	Média Geral dos Subconjuntos de Atributos da Base de Dados DLBCL utilizando a Abordagem Filtro						
Subconjuntos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	87,70 ± 12,17*	77,30 ± 4,99	87,79 ± 11,45	83,62 ± 12,38*	85,62 ± 11,71*	85,95 ± 12,13*	86,43 ± 12,06
1.2	83,09 ± 8,23	80,96 ± 7,55	82,89 ± 8,24	80,31 ± 10,26*	82,48 ± 9,40	82,08 ± 9,42	82,66 ± 9,51
2.1	83,00 ± 12,20	73,17 ± 7,29*	83,62 ± 11,91	75,03 ± 10,03*	77,83 ± 10,17*	79,14 ± 10,72*	79,95 ± 10,92*
2.2	82,26 ± 7,03	72,57 ± 3,15*	82,76 ± 7,18	72,80 ± 5,27*	76,33 ± 6,29*	77,25 ± 6,09*	77,18 ± 6,00*

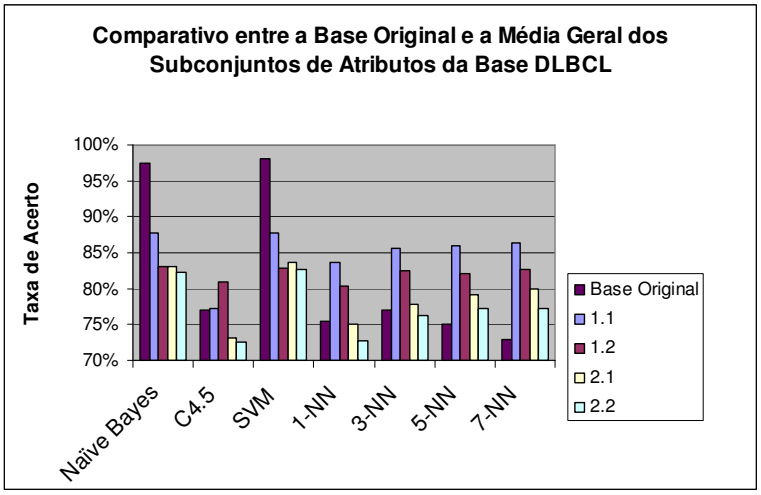

Figura 44: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base DLBCL utilizando a Abordagem Filtro.

Na tabela 31 são apresentados os resultados da utilização conjunta do método projeção aleatória e da seleção de atributos na base de dados DLBCL-Tumor. Nesses experimentos nota-se que o algoritmo C4.5 teve seu resultados considerados estatisticamente piores comparados com algoritmo base *Naïve Bayes* em todos os casos. Além do algoritmo C4.5, o algoritmo 1-NN nos subconjuntos 1.1 e 2.2 e o algoritmo SVM no subconjunto 1.2 tiveram seus resultados estatisticamente piores em seus subconjuntos. Ainda vale destacar o algoritmo 7-NN que teve resultados estatisticamente melhores, comparados com o algoritmo base, nos subconjuntos de atributos 1.2, 2.1 e 2.2.

Se compararmos esses resultados com os resultados obtidos quando se utilizou a base de dados com todos os atributos (figura 45), nota-se que quando se utilizou o algoritmo SVM e o 3-NN os resultados foram inferiores, porém o restante dos resultados foram pelo menos iguais.

Tabela 31: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem Filtro nos Subconjunto de atributos da Base de Dados DLBCL-Tumor.

	Média Geral dos Subconjuntos de Atributos da Base de Dados DLBCL-TUMOR usando a Abordagem Filtro						
Subconjuntos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	91,81 ± 6,04	85,27 ± 4,18*	90,57 ± 8,20	89,69 ± 6,68*	91,39 ± 6,22	91,47 ± 6,59	91,88 ± 5,76
1.2	89,21 ± 4,64	87,21 ± 5,57*	84,12 ± 5,11*	87,71 ± 6,66	89,25 ± 6,10	89,06 ± 7,00	**90,17 ± 5,37**
2.1	86,95 ± 5,01	82,88 ± 4,55*	88,74 ± 9,36	85,74 ± 6,37	86,84 ± 5,60	87,52 ± 6,01	**88,47 ± 5,58**
2.2	86,52 ± 2,29	82,01 ± 2,14*	88,88 ± 7,54	84,79 ± 3,61*	86,13 ± 2,46	86,79 ± 2,87	**87,96 ± 3,10**

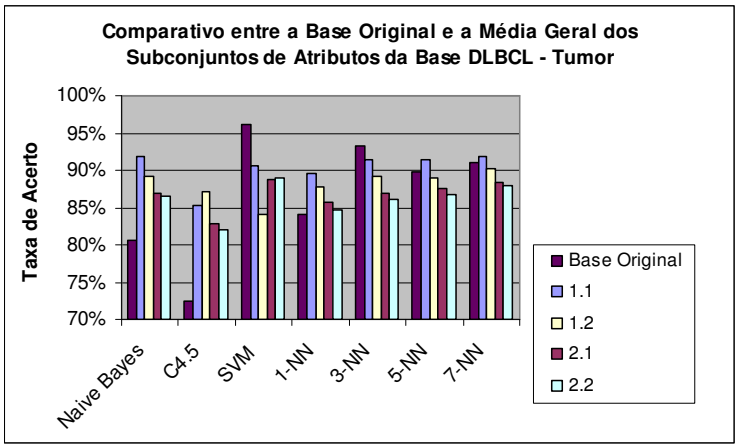

Figura 45: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base DLBCL-Tumor utilizando a Abordagem Filtro.

Analisando os resultados obtidos da utilização conjunta do método projeção aleatória e a seleção de atributos nos subconjuntos de atributos da base de dados DLBCL-Outcome (tabela 32), observa-se que todos os resultados dos algoritmos do subconjunto 1.1 são considerados estatisticamente piores. Isso também acontece para os algoritmos SVM, 1-NN, 3-NN e 5-NN no subconjunto 1.2. Já os resultados dos algoritmos dos subconjuntos 2.1 e 2.2 são estatisticamente melhores que o algoritmo base *Naïve Bayes*.

A figura 46 mostra um comparativo entre a base de dados original e os subconjuntos de atributos gerados pela utilização conjunta no método projeção aleatória e da seleção de atributos. É visível a diferença dos resultados do conjunto original e dos subconjuntos de atributos. Observa-se que quando se utilizou a busca seqüencial os resultados foram superiores aos resultados de quando se utilizou a busca aleatória e até mesmo superior a base de dados original.

Tabela 32: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem Filtro nos Subconjuntos de Atributos da Base de Dados DLBCL-Outcome.

	Média Geral dos Subconjuntos de Atributos da Base de Dados DLBCL-Outcome usando a Abordagem Filtro						
Subconjuntos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	63,29 ± 8,64	60,74 ± 5,79*	60,34 ± 6,18*	59,98 ± 5,69*	61,28 ± 6,18*	61,28 ± 6,33*	61,13 ± 6,77*
1.2	62,03 ± 6,81	61,90 ± 7,52	59,62 ± 4,82*	59,05 ± 6,03*	61,43 ±6,30*	61,00 ± 6,12*	61,27 ± 6,77
2.1	23,29 ± 18,28	**29,43 ± 24,95**	**27,71 ± 23,11**	**29,64 ± 25,52**	**30,00 ± 25,42**	**29,31 ± 25,12**	**28,09 ± 24,14**
2.2	45,07 ± 6,25	**53,29 ± 1,39**	**51,83 ± 3,35**	**52,60 ± 2,89**	**52,71 ± 3,42**	**53,31 ± 1,46**	**51,90 ± 1,77**

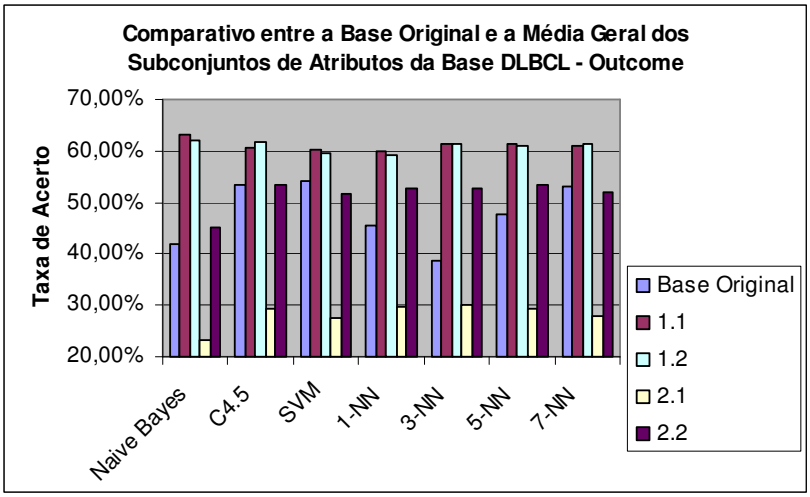

Figura 46: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base DLBCL-Outcome utilizando a Abordagem Filtro.

A tabela 33 apresenta os resultados obtidos da utilização conjunta do método projeção aleatória e a seleção de atributos nos subconjuntos de atributos da base de dados DLBLC-NIH. Definindo o algoritmo *Naïve Bayes* como algoritmo base a ser comparado estatisticamente com os demais algoritmos, para o subconjunto de atributos 1.1 todos os outros algoritmos são considerados piores. Para o subconjunto de atributos 1.2 apenas o resultado do algoritmo C4.5 é considerado equivalente ao resultado do algoritmo *Naïve Bayes*. Já para os subconjuntos de atributos 2.1 e 2.2 apenas o resultado do algoritmo SVM é considerado equivalente ao resultado do algoritmo base *Naïve Bayes*, o restante dos resultados são estatisticamente piores.

A figura 47 mostra um comparativo entre os resultados da utilização conjunta dos métodos de redução da dimensionalidade, em que os resultados dos algoritmos comprovam que houve um aumento na taxa de acerto do classificador.

Tabela 33: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem Filtro nos Subconjuntos de Atributos da Base de Dados DLBCL-NIH.

	Média Geral dos Subconjuntos de Atributos da Base de Dados DLBCL-NIH usando a Abordagem Filtro						
Subconjuntos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	62,95 ± 5,48	59,95 ± 1,77*	61,10 ± 4,11*	55,93 ± 2,52*	57,29 ± 3,65*	58,22 ± 4,45*	58,48 ± 4,83*
1.2	63,28 ± 5,65	60,64 ± 2,40	61,52 ± 4,68*	56,27 ± 3,30*	57,47 ± 4,50*	57,89 ± 5,40*	58,12 ± 5,73*
2.1	41,58 ± 23,07	40,04 ± 21,44*	42,05 ± 23,75	38,45 ± 21,40*	39,21 ± 22,03*	39,98 ± 21,91*	39,92 ± 22,05*
2.2	59,08 ± 2,01	56,40 ± 1,66*	59,21 ± 1,87	54,23 ± 2,98*	55,55 ± 3,87*	56,05 ± 2,92*	56,32 ± 2,32*

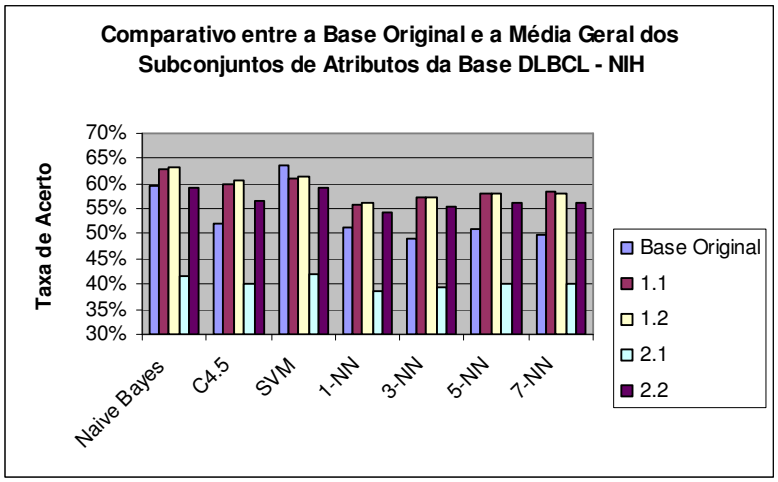

Figura 47: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base DLBCL-NIH utilizando a Abordagem Filtro.

A tabela 34 mostra os resultados da utilização conjunta dos métodos de redução de dimensionalidade nos subconjuntos de atributos da base de dados ALL/AML. Nos subconjuntos de atributos 1.1 e 2.1 os resultados dos algoritmos C4.5 e k-NN são estatisticamente piores comparados com algoritmo *Naïve Bayes*. Já no subconjunto 1.2 além dos algoritmos C4.5 e k-NN o resultado do algoritmo SVM também é estatisticamente pior. No subconjunto de atributos 2.2 o algoritmo SVM possui resultado estatisticamente melhor comparado com algoritmo base e os algoritmos C4.5 e k-NN, para k=1, k=5 e k=7) resultados estatisticamente piores.

Uma comparação entre os resultados obtidos pela utilização conjunta do método projeção aleatória e da seleção de tributos pode ser observada na figura 48. Este gráfico mostra que houve um aumento na taxa de acerto do classificador, na maioria das vezes, quando se utilizou o método de redução de atributos.

Tabela 34: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem Filtro nos Subconjuntos de Atributos da Base de Dados ALL/AML.

	Média Geral dos Subconjuntos de Atributos da Base de Dados ALL/AML usando a Abordagem Filtro						
Subconjuntos	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1.1	93,26 ± 6,69	81,86 ± 4,34*	92,37 ± 7,77	90,35 ± 9,82*	90,56 ± 8,96*	90,72 ± 8,48*	91,01 ± 8,1*
1.2	88,82 ± 4,98	85,24 ± 5,96*	87,45 ±5,74*	85,40 ± 7,80*	86,43 ± 6,93*	86,75 ± 6,27*	86,69 ± 6,10*
2.1	89,23 ± 6,35	77,95 ± 4,56*	89,68 ± 8,64	86,42 ± 8,95*	87,05 ± 8,19	86,84 ± 7,49*	86,70 ± 6,97*
2.2	87,01 ± 3,78	77,74 ± 1,24*	**88,96 ± 6,19**	85,23 ± 5,18*	85,90 ± 5,85	85,03 ± 4,96*	84,06 ± 4,17*

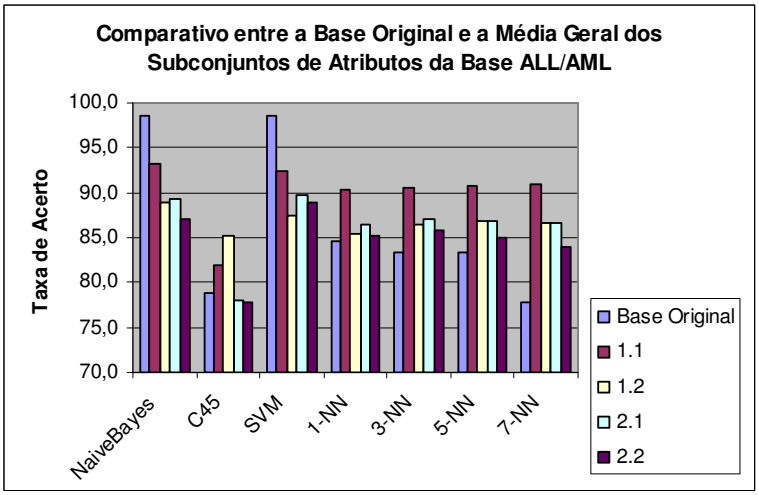

Figura 48: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base ALL/AML utilizando a Abordagem Filtro.

4.4.2.2 Abordagem *Wrapper*

As tabelas 35, 36, 37, 38 e 39 mostram os resultados obtidos por cada uma das bases de dados em estudo quando se utilizou a abordagem *wrapper* como critério de avaliação dos subconjuntos. O método de busca identificado com o número 1 refere-se a busca seqüencial e o número 2 a busca aleatória.

A tabela 35 refere-se aos resultados obtidos da utilização conjunta do método projeção aleatória e da seleção de atributos nos subconjuntos de atributos da base DLBCL. Analisando o método de busca 1, observa-se que o resultado dos algoritmos C4.5, 1-NN e 3-NN são estatisticamente piores comparados com o algoritmo *Naïve Bayes*. Já para o método de busca 2 os resultados dos algoritmos C4.5 e o *k*-NN (para os valores de k=1, k=3, k=5 e k=7).

A figura 49 mostra um comparativo entre os resultados da utilização conjunta do método projeção aleatória e da seleção de atributos da abordagem *wrapper* com a base de dados original em que os resultados da foram superiores na maioria dos experimentos, apenas o resultado dos algoritmos *Naïve Bayes* e SVM foram inferiores comparado com o resultado obtido no conjunto de dados original.

A tabela 36 refere-se aos resultados obtidos da utilização conjunta do método projeção aleatória e da seleção de atributos nos subconjuntos de atributos da base DLBCL-Tumor. Observa-se que os resultados do algoritmo C4.5 nos dois métodos de busca foram estatisticamente piores comparados com o algoritmo base e o resultado do algoritmo 5-NN foi estatisticamente melhor no método de busca seqüencial.

Tabela 35: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem *Wrapper* nos Subconjuntos de Atributos da Base de Dados DLBCL.

	Média Geral dos Subconjuntos de Atributos da Base de Dados DLBCL usando a Abordagem *Wrapper*						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	91,16 ± 7,67	85,72 ± 6,82*	91,32 ± 6,60	89,72 ± 9,03*	89,63 ± 9,02*	89,62 ± 9,64	90,83 ± 7,61
2	92,19 ± 6,93	82,88 ± 4,34*	**94,38 ± 6,19**	88,61 ± 5,16*	89,42 ± 4,16*	89,55 ± 4,06*	89,70 ± 4,39*

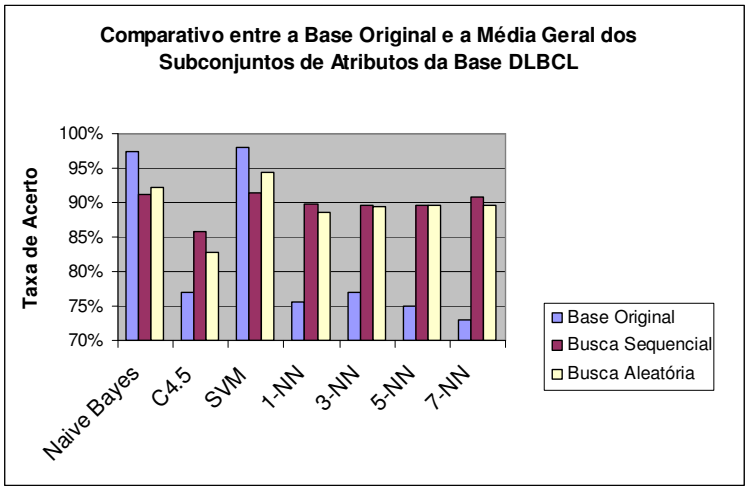

Figura 49: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base DLBCL utilizando a Abordagem *Wrapper*.

Na tabela 36, referente a base de dados DLBCL-Tumor, observa-se que, apenas o resultado do algoritmo C4.5 foi estatisticamente pior que o algoritmo *Naïve Bayes* nos dois experimentos de busca. Além disso, nota-se que o algoritmo 5-NN teve seu resultado estatisticamente melhor no método de busca 1.

Se compararmos esses resultados com os resultados obtidos quando se utilizou todos os atributos veremos que alguns casos a diferença da taxa de acerto do classificador foi bastante grande, melhorando com os métodos de redução de atributos (figura 50).

A tabela 37 refere-se aos resultados obtidos da base de dados DLBCL-Outcome em que o algoritmo k-NN teve os melhores resultados estatisticamente comparados com o algoritmo *Naïve Bayes*.

A figura 51 mostra um comparativo entre os resultados do conjunto original e dos subconjuntos, gerados quando se utilizou os métodos de redução de atributos em

conjunto. Percebe-se que em todos os casos houve uma melhora significativa dos resultados dos classificadores.

Tabela 36: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem *Wrapper* nos Subconjuntos de Atributos da Base de Dados DLBCL-Tumor.

	Média Geral dos Subconjuntos de Atributos da Base DLBCL-Tumor usando a Abordagem *Wrapper*						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	92,78 ± 6,06	88,86 ± 6,91*	88,40 ± 10,47	94,58 ± 6,66	94,68 ± 4,72	**95,09 ± 4,74**	94,60 ± 5,43
2	92,09 ± 2,84	88,60 ± 5,78*	91,77 ± 11,23	92,34 ± 7,74	91,13 ± 7,58	91,91 ± 7,58	92,99 ± 5,01

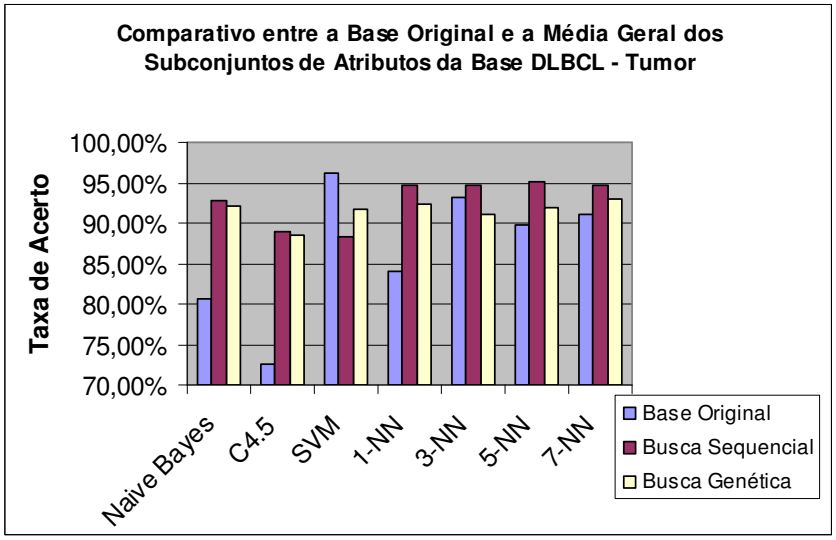

Figura 50: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base DLBCL-Tumor utilizando a Abordagem *Wrapper.*

Tabela 37: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem *Wrapper* nos Subconjuntos de Atributos da Base de Dados DLBCL-Outcome.

	Média Geral dos Subconjuntos de Atributos da Base DLBCL-Outcome usando a Abordagem *Wrapper*						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	71,48 ± 7,13	69,93 ± 12,46	70,28 ± 10,61	**75,17 ± 7,29**	**75,22 ±6,81**	**74,40 ± 7,29**	72,66 ± 8,20
2	55,45 ± 8,54	58,12 ± 1,81	62,10 ± 4,03	**69,10 ± 3,24**	**68,38 ±2,59**	**65,71 ± 3,52**	**63,81 ± 3,45**

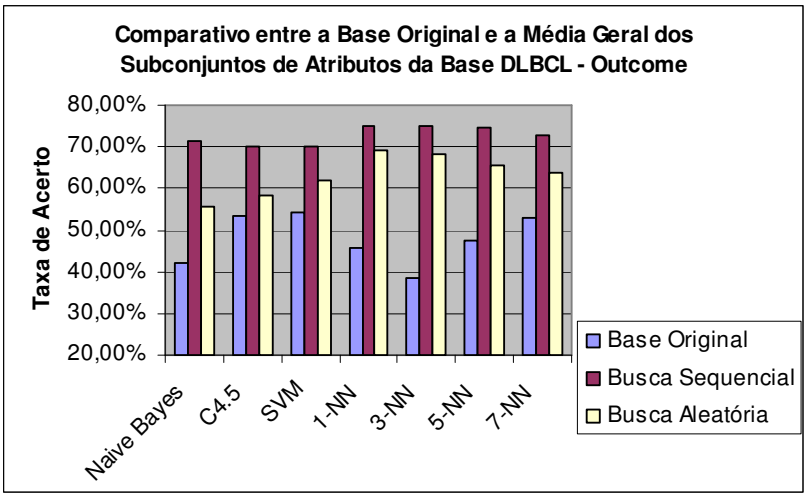

Figura 51: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base DLBCL-Outcome utilizando a Abordagem *Wrapper.*

A tabela 38 mostra os resultados obtidos da base de dados DLBCL-NIH em que o algoritmo k-NN teve os piores resultados estatisticamente avaliados quando se utilizou o método de busca 1. Já no método de busca 2 o algoritmo C4.5 teve o pior resultado. Porém, todos esses resultados foram superiores aos resultados obtidos quando utilizado a base de dados com todos os atributos, isso pode ser notado na figura 52.

A tabela 39 mostra os resultados obtidos da base de dados ALL/AML em que o algoritmo C4.5 teve resultados estatisticamente piores que o algoritmo *Naïve Bayes* nos dois métodos de busca. Além desse, os algoritmos 5-NN e 7-NN também tiveram resultados estatisticamente piores no método de busca aleatória.

A figura 53 mostra que em apenas dois casos os resultados dos algoritmos, quando aplicados sobre a base de dados com todos atributos, tiveram uma taxa de acerto maior.

Tabela 38: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem *Wrapper* nos Subconjuntos de Atributos da Base de Dados DLBCL-NIH.

	Média Geral dos Subconjuntos de Atributos da Base DLBCL-NIH usando a Abordagem *Wrapper*						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	65,74 ± 3,72	66,48 ± 7,10	63,41 ± 6,91	63,90 ± 2,60*	64,44 ± 3,18*	64,69 ± 2,92*	65,05 ± 3,00*
2	63,68 ± 1,42	58,32 ± 1,06*	64,58 ± 5,01	63,96 ± 0,89	64,17 ± 1,29	63,92 ± 0,86	63,48 ± 1,29

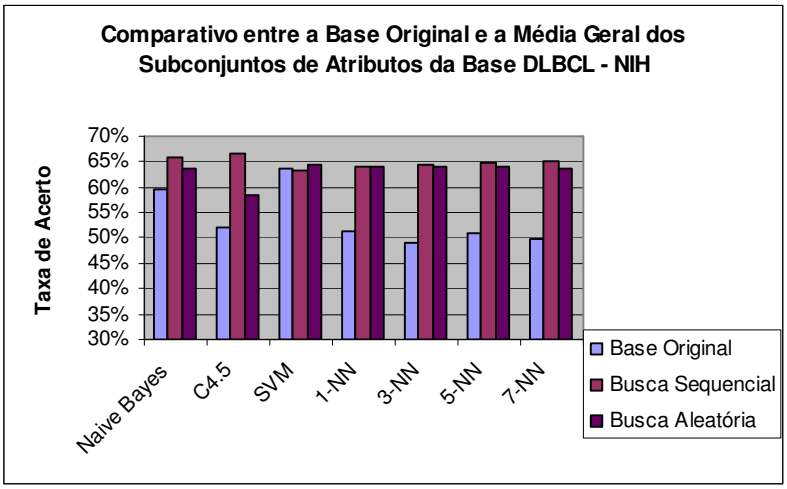

Figura 52: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base DLBCL-NIH utilizando a Abordagem *Wrapper*.

Tabela 39: Média Geral do Resultado da utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos utilizando a Abordagem *Wrapper* nos Subconjuntos de Atributos da Base de Dados ALL/AML.

	Média Geral dos Subconjuntos de Atributos da Base ALL/AML usando a Abordagem *Wrapper*						
Método de Busca	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
1	93,97 ± 5,13*	88,75 ± 6,21	92,96 ± 7,17	93,55 ± 6,20	93,58 ± 5,37	91,37 ± 6,89	91,45 ± 6,85
2	94,33 ± 4,71*	86,48 ± 3,54	94,72 ± 5,89	94,47 ± 4,20	93,95 ± 3,85	91,91 ± 4,56*	91,06 ± 4,47*

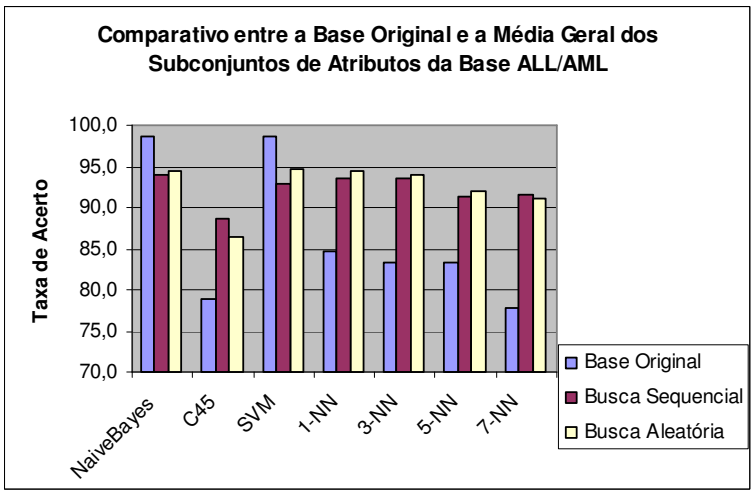

Figura 53: Comparativo entre a Base Original e a Média Geral dos Subconjuntos de Atributos da Base ALL/AML utilizando a Abordagem *Wrapper*.

Se calcularmos a média aritmética, dos resultados dos algoritmos de classificação, de cada método de seleção de atributos aplicado, teremos a média de cada método de seleção de atributos em cada base de dados. Esse resultado pode ser visualizado na tabela 40.

Tabela 40: Comparativo entre os Métodos de Seleção de Atributos quando aplicado juntamente com o Método de Projeção Aleatória nas Cinco Bases de Dados.

	Comparação de Performance entre os Métodos de Seleção de Atributos				
Subconjuntos	DLBCL	DLBCL-Tumor	DLBCL-Outcome	DLBCL-NIH	AML/ALL
1.1	84,92%	90,30%	61,15%	59,13%	90,02%
1.2	82,07%	88,10%	60,90%	59,31%	86,68
2.1	78,82%	86,73%	28,21%	40,18%	86,27%
2.2	77,31%	86,16%	51,53%	56,69%	84,85
1.3	91,16%	92,78%	71,48%	65,74%	93,97
1.4	85,72%	88,86%	69,93%	66,48%	88,75
1.5	91,32%	88,40%	70,28%	63,41%	92,96
1.6	89,72%	94,58%	75,17%	63,90%	93,55
1.7	89,63%	94,68%	75,22%	64,44%	93,58
1.8	89,62%	95,09%	74,40%	64,69%	91,37
1.9	90,83%	94,60%	72,66%	65,05%	91,45
2.3	92,19%	92,09%	55,45%	63,68%	94,33
2.4	82,88%	88,60%	58,12%	58,32%	86,48
2.5	94,38%	91,77%	62,10%	64,58%	94,72
2.6	88,61%	92,34%	69,10%	63,96%	94,47
2.7	89,42%	91,14%	68,38%	64,17%	93,95
2.8	89,55%	91,91%	65,71%	63,92%	91,91
2.9	89,70%	92,99%	63,81%	63,48%	91,06

4.5 Comparação Geral

Após feita uma análise em cada experimento, foi calculado a média dos algoritmos de classificação. Para os experimentos que se aplicou a seleção de atributos os resultados estão divididos nas duas abordagens utilizadas, a abordagem filtro e a abordagem *wrapper*. Em seguida, é calculada a média geral dos resultados das duas abordagens. Nos experimentos em que se usou o método de projeção aleatória os resultados estão divididos em duas partes: a primeira quando se utilizou um número fixo de atributos e a segunda quando se utilizou a porcentagem de atributos para a formação dos novos subconjuntos de atributos.

Esses resultados também são analisados estatisticamente através de testes de hipóteses com amostras pareadas, tomando como base o algoritmo *Naïve Bayes*.

Para podermos comparar todos esses resultados com o resultado obtido pelos algoritmos de classificação quando aplicado sobre as bases de dados com todos os atributos calculou-se a média aritmética também desses resultados.

A tabela 41 mostra a média aritmética obtida pelos algoritmos de classificação nas cinco bases de dados. Analisando estatisticamente esses resultados nota-se que apenas o algoritmo C4.5 é estatisticamente pior.

Tabela 41: Média do Resultado da Classificação das Bases de Dados Originais.

Média do Resultado da Classificação das Bases de Dados Originais						
Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
75,6 ± 24,6	66,8 ± 13,1*	82,1 ± 21,4	68,2 ± 18,5	68,3 ± 23,3	69,3 ± 19,1	69,0 ± 17,3

4.5.1 Comparação da Aplicação da Seleção de Atributos

Na tabela 42 é mostrada a média dos resultados da seleção de atributos, quando usou-se a abordagem filtro, de cada algoritmo de classificação, das cinco bases de dados.

Observa-se que nas bases de dados DLBCL, DLBCL-NIH e ALL/AML o algoritmo SVM teve resultado estatisticamente equivalente ao algoritmo Naïve Bayes. Já os outros algoritmos tiveram resultados estatisticamente piores comparados com o algoritmo base.

Na base de dados DLBCL-Tumor os algoritmos SVM, 3-NN, 5-NN e 7-NN tiveram resultados estatisticamente melhores e o algoritmo C4.5 resultado estatisticamente pior.

Na base DLBCL-Outcome apenas o algoritmo SVM teve resultado estatisticamente melhor.

Na tabela 43 é mostrada a média dos resultados da seleção de atributos, quando usou-se a abordagem *wrapper*, de cada algoritmo de classificação, das cinco bases de dados.

A base de dados DLBCL e ALL/AML obteve-se o mesmo resultado da análise estatística feita na abordagem filtro. Na base de dados DLBCL apenas o algoritmo SVM teve resultado estatisticamente equivalente ao algoritmo base e que os demais algoritmos tiveram resultados estatisticamente piores. Já na base de dados ALL/AML o algoritmo SVM teve resultado estatisticamente equivalente ao *Naïve Bayes* e os outros algoritmos resultados estatisticamente piores. Na base DLBCL-Tumor os algoritmos SVM, 3-NN e 7-NN tiveram resultados estatisticamente melhores. Na base DLBCL-Outcome todos os resultados foram estatisticamente equivalentes. Já na base DLBCL-NIH o algoritmo SVM

teve resultado estatisticamente melhor e os outros algoritmos resultados estatisticamente piores comparados com o algoritmo *Naïve Bayes*.

Calculando-se a média aritmética dos resultados obtidos da abordagem filtro e da abordagem *wrapper* chega-se a média geral dos algoritmos de classificação, a qual pode ser observada na tabela 44. Observa-se que os algoritmos que tiveram os piores resultados analisados estatisticamente nas duas abordagens separadamente continuam tendo resultados piores na média geral de cada base de dados. O mesmo também acontece no melhor resultado.

Calculando a media aritmética, de cada algoritmos de classificação, de todas as bases de dados, como é mostrada na tabela 45, observa-se que o resultado dos algoritmos C4.5 e 1-NN são estatisticamente piores que o algoritmo *Naïve Bayes*.

Comparando esses resultados com a média dos resultados obtidos da base de dados originais, nota-se o aumento da taxa de acerto, isso pode ser observado claramente na figura 54. Além disso, analisando e comparando estatisticamente os resultados originais e os resultados da seleção de atributos, observa-se que apenas o algoritmo SVM foi estatisticamente equivalente, os demais algoritmos apresentaram resultados estatisticamente melhor na aplicação da seleção de atributos.

Tabela 42: Média por Algoritmo de Classificação de cada Base de Dados usando a Seleção de Atributos – Abordagem Filtro.

	Media por Algoritmo de Classificação de cada Base de Dados - Abordagem Filtro						
Base de Dados	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	95,1 ± 4,8	82,1 ± 5,3*	93,1 ± 7,1	85,5 ± 12,2*	86,4 ± 14,6*	85,8 ± 10,5*	83,8 ± 12,7*
DLBCL-Tumor	87,7 ± 8,3	84,6 ± 5,1*	**91,3 ± 9,8**	89,5 ± 4	**92,5 ± 2,4**	**92,2 ± 3,4**	**92,1 ± 3,5**
DLBCL-Outcome	58,5 ± 22,5	61,7 ± 10,5	**66,2 ± 9,3**	58,8 ± 17,3	59,2 ± 17,8	58,7 ± 13,2	62,6 ± 10,5
DLBCL-NIH	65,3 ± 5,7	60,2 ± 7,8*	62,4 ± 4,6	55,6 ± 2,5*	52,9 ± 4,9*	54,8 ± 3,8*	54,2 ± 3,8*
ALL/AML	92,7 ± 8,2	86,8 ± 5,2*	91,6 ± 7,8	87,2 ± 7,9*	88,7 ± 7,6*	89 ± 5,6*	86,2 ± 9,1*

Tabela 43: Média por Algoritmo de Classificação de cada Base de Dados usando a Seleção de Atributos – Abordagem *Wrapper.*

	Media por Algoritmo de Classificação de cada Base de Dados – Abordagem *Wrapper*						
Base de Dados	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	99,0 ± 1,4	91,5 ± 0*	99,0 ± 1,4	91,3 ± 9,5*	88,8 ± 13,1*	88,5 ± 13,4*	90,8 ± 13,1*
DLBCL-Tumor	91,0 ± 12,8	91,4 ± 2,9	**98,8 ± 0**	94,0 ± 8,5	**95,3 ± 3.0**	93,6 ± 1,8	**94,8 ± 3,5**
DLBCL-Outcome	71,2 ± 26,2	69,3 ± 30,2	76,8 ± 24,3	65.0 ± 17,9	77,5 ± 17,2	72,2 ± 29,9	73,0 ± 18,4
DLBCL-NIH	69,6± 7,7	65,6 ± 14,4*	**72,9 ± 0,6**	59,6 ± 1,2*	61,9 ± 3,2*	62,5 ± 6,5*	61,9 ± 9,2*
ALL/AML	100 ± 0	93,9 ±1,0*	98,6 ± 2,1	95,2 ± 6,8*	93,8 ± 8,8*	93,8 ± 8,8*	93,0 ± 9,8*

Tabela 44: Média Geral dos Algoritmos de Classificação de cada Base de Dados usando Seleção de Atributos.

	Media por Algoritmo de Classificação de cada Base de Dados usando Seleção de Atributos						
Base de Dados	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	97,1 ± 2,8	86,8 ± 6,6*	96,1 ± 4,2	88,5 ± 4,2*	87,6 ± 1,7*	87,2 ± 1,9*	87,3 ± 4,9*
DLBCL-Tumor	89,3 ± 2,3	88,0 ± 4,8	**95,0 ± 5,3**	**91,8 ± 3,2**	**93,9 ± 1,9**	**92,9 ± 1,0**	**93,5 ± 1,9**
DLBCL-Outcome	64,8 ± 9,0	65,5 ± 5,4	**71,5 ± 7,5**	61,9 ± 4,4*	68,4 ± 12,9	65,4 ± 9,5	67,8 ± 7,4
DLBCL-NIH	67,4 ± 3,0	62,9 ± 3,8*	67,7 ± 7,4	57,6 ± 2,8*	57,4 ± 6,3*	58,7 ± 5,4*	58,0 ± 5,4*
ALL/AML	96,4 ± 5,2	90,4 ± 5,0*	95,1 ± 4,9	91,2 ± 5,6*	91,2 ± 3,6*	91,4 ± 3,4*	89,6 ± 4,8*

Tabela 45: Média Geral dos Algoritmos de Classificação de todas as Bases de Dados usando Seleção de Atributos.

Média Geral dos Algoritmos de Classificação usando Seleção de Atributos						
Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
83,0 ± 15,8	78,7 ± 13,3*	85,1 ± 14,2	78,2 ± 17,0*	79,47 ± 16,0	79,1 ± 15,9	79,2 ± 15,5

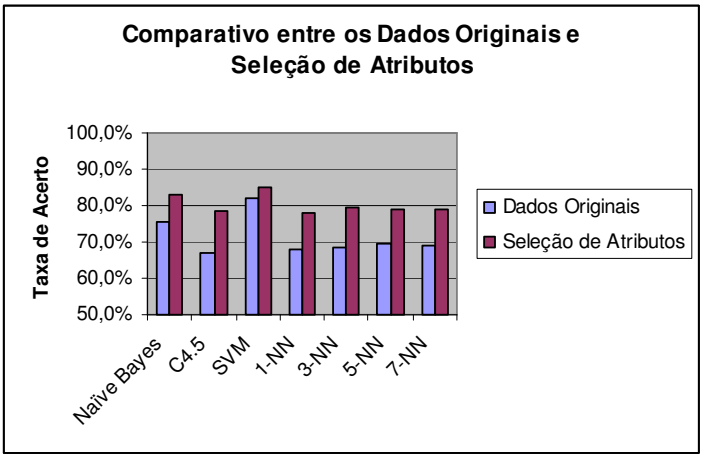

Figura 54: Comparativo entre os Dados Originais e a Seleção de Atributos.

4.5.2 Comparação da Aplicação do Método de Projeção Aleatória

Na tabela 46 é mostrada a média dos resultados do método de projeção aleatória, quando usou-se um número fixo de atributos, de cada algoritmo de classificação, das cinco bases de dados.

Através de análises estatísticas foi possível observar que o algoritmo SVM teve resultado estatisticamente melhor comparado com o algoritmo *Naïve Bayes* nas cinco bases de dados. Observa-se ainda que na base de dados DLBCL-*Outcome* os outros algoritmos também se destacaram por apresentar resultados estatisticamente melhores que o algoritmo base. Nota-se ainda que nas outras bases de dados o algoritmo C4.5 e o *k*-NN tiveram resultados estatisticamente piores.

A tabela 47 mostra a média aritmética obtida pelo método de projeção aleatória, usando uma porcentagem de atributos, através da média aritmética dos algoritmos de classificação.

Na base de dados DLBCL apenas o algoritmo SVM teve resultado estatisticamente equivalente ao algoritmo base, os outros algoritmos tiveram resultados estatisticamente piores.

Na base de dados DLBCL-*Tumor* os algoritmos SVM, 5-NN e 7-NN tiveram resultados estatisticamente melhores e os algoritmos C4.5, 1-NN e 3-NN resultados estatisticamente piores.

Na base de dados DLBCL-*Outcome* todos os resultados dos algoritmos foram estatisticamente melhores que o algoritmo *Naïve Bayes*.

Nas bases de dados DLBCL-NIH e ALL/AML o algoritmo SVM teve resultado estatisticamente melhor e os demais algoritmos resultados estatisticamente pior comparados com o algoritmo base.

Tabela 46: Média por Algoritmo de Classificação de cada Base de Dados usando o Método de Projeção Aleatória com Número Fixo de Atributos.

	Média Geral por Algoritmo de Classificação de cada Base de Dados - Número Fixo de Atributos						
Base de Dados	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	78, 0 ± 6,4	71,2 ± 4,4*	**82,3 ± 4,5**	69,8 ± 2,6*	72,1 ± 3,7*	73,6 ± 3,9*	74,6 ± 4,0*
DLBCL-Tumor	85,3 ± 3,4	81,3 ± 1,9*	**87,1 ± 6,9**	83,7 ± 3,6*	83,6 ± 1,7*	85,0 ± 2,5	85,4 ± 2,8
DLBCL-Outcome	43,3 ± 2,2	**48,8 ± 3,9**	**50,1 ± 2,3**	**53,7 ± 1,7**	**54,4 ± 0,7**	**52,2 ± 0,9**	**51,1 ± 1,7**
DLBCL-NIH	56,6 ± 1,2	55,0 ± 1,7*	**56,4 ± 0,4**	54,2 ± 1,0*	54,7 ± 1,2*	54,5 ± 1,3*	55,1 ± 1,5*
ALL/AML	86,2 ± 4,0	75,3 ± 2,0*	**88,3 ± 6,0**	83,5 ± 4,8*	84,3 ± 4,7*	83,3 ± 3,8*	82,1 ± 3,4*

Tabela 47: Média por Algoritmo de Classificação de cada Base de Dados usando o Método de Projeção Aleatória com Porcentagem de Atributos.

	Média Geral por Algoritmo de Classificação de cada Base de Dados - Porcentagem de Atributos						
Base de Dados	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	93,5 ± 1,8	75,3 ± 3,7*	93,0 ±1,4	80,0 ± 2,3*	84,6 ± 3,1*	85,6 ± 2,7*	85,2 ± 2,9*
DLBCL-Tumor	88,5 ± 0,4	84,8 ± 1,6*	**96,4 ± 0,4**	87,9 ± 0,2*	86,8 ± 0,4*	**90,1 ± 0,9**	**91,0 ± 0,8**
DLBCL-Outcome	38,4 ± 1,3	**47,5 ± 1,3**	**52,0 ± 0,7**	**54,0 ± 2,0**	**56,8 ± 0,9**	**53,9 ± 1,2**	**50,9 ± 1,0**
DLBCL-NIH	60,0 ± 0,3	54,5 ± 1,4*	**62,4 ± 1,6**	56,9 ± 0,4*	60,0 ± 0,9	58,4 ± 0,8*	58,1 ± 0,5*
ALL/AML	92,7 ± 0,8	81,6 ± 1,1*	**97,4 ± 0,7**	91,5 ± 1,3*	92,1 ± 1,6*	90,8 ± 1,3*	89,9 ± 1,0*

Na tabela 48 é mostrada a média geral da aplicação do método de projeção aleatória de cada base de dados. Observa-se que os resultados estatísticos obtidos foram muito parecidos com os resultados das duas tabelas anteriores, em que na maioria dos casos o algoritmo SVM teve resultado estatisticamente melhor.

Já a tabela 49 apresenta a média geral da aplicação do método de projeção aleatória em todas as bases de dados. Analisando esses resultados observa-se que todos os algoritmos possuem resultados estatisticamente equivalentes comparados com algoritmo base *Naïve Bayes*.

Comparando ainda esses resultados com os resultados obtidos na base de dados originais observa-se que os algoritmos *Naïve Bayes* e SVM apresentaram resultados inferiores, isso pode ser visualizado na figura 55. Porém, analisando estatisticamente os resultados obtidos desse método com os resultados obtidos da base de dados original observa-se que os resultados são estatisticamente equivalentes.

Tabela 48: Média por Algoritmo de Classificação de cada Base de Dados usando o Método de Projeção Aleatória.

	Media Geral por Algoritmo de Classificação de cada Base de Dados						
Base de Dados	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	85,8 ± 9,4	73,3 ± 4,4*	87,6 ± 6,4	74,9 ± 5,9*	78,6 ± 7,3*	79,6 ± 7,0*	79,9 ± 6,5*
DLBCL-Tumor	86,9 ± 2,9	83,1 ± 2,5*	**91,8 ± 6,7**	85,8 ± 3,3*	85,2 ± 2,1*	87,5 ± 3,2	**88,2 ± 3,5**
DLBCL-Outcome	40,9 ± 3,1	**48,2 ± 2,8**	**51,1 ± 1,9**	**53,8 ± 1,8**	**55,6 ± 1,5**	**53,0 ± 1,3**	**51,0 ± 1,3**
DLBCL-NIH	58,3 ± 2,0	54,7 ± 1,5*	**59,4 ± 3,4**	55,6 ± 1,6*	57,4 ± 3,0*	56,4 ± 2,3*	56,6 ± 1,9*
ALL/AML	89,5 ± 4,4	78,4 ± 3,7*	**92,9 ± 6,2**	87,5 ± 5,4*	88,2 ± 5,3	87,1 ± 4,8*	86,0 ± 4,7*

Tabela 49: Média Geral dos Algoritmos de Classificação usando o Método de Projeção Aleatória.

Média Geral dos Algoritmos de Classificação usando o Método de Projeção Aleatória						
Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
72,2 ± 21,6	67,5 ± 15,3	76,5 ± 19,7	71,5 ±16,1	72,9 ± 15,5	72,7 ± 16,8	72,3 ± 17,3

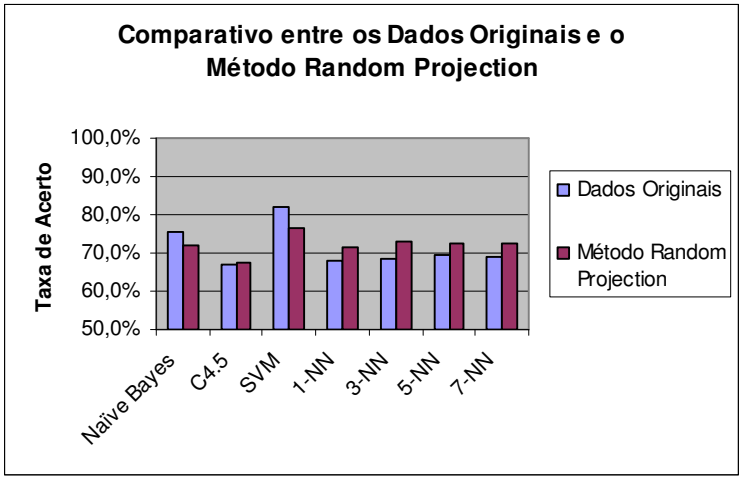

Figura 55: Comparativo entre os Dados Originais e o Método de Projeção Aleatória.

4.5.3 Comparação da Aplicação Conjunta do Método de Projeção Aleatória e a Seleção de Atributos

Na tabela 50 é mostrada a média das execuções da utilização conjunta do método de projeção aleatória com a seleção de atributos quando se usou a abordagem filtro, de cada base de dados.

Observa-se que em todas as bases de dados o algoritmo SVM teve resultado estatisticamente equivalente ao algoritmo base. Nota-se ainda que os algoritmos C4.5 e *k*-NN tiveram resultados estatisticamente piores na média de quase todas as bases de dados.

Na tabela 51 é mostrada a média das execuções da utilização conjunta do método de projeção aleatória com a seleção de atributos quando se usou a abordagem *wrapper*, de cada base de dados.

Na base de dados DLBCL observa-se que o algoritmo SVM teve resultado estatisticamente melhor e os outros algoritmos resultados estatisticamente pior comparados como o algoritmo base.

Na base de dados DLBCL-Tumor o algoritmo 1-NN teve resultado estatisticamente melhor, e os algoritmos C4.5, SVM, 5-NN e 7-NN resultados estatisticamente piores comparados com o algoritmo *Naïve Bayes*.

Na base de dados DLBCL-Outcome o algoritmo *k*-NN apresentou estatisticamente melhores resultados e o algoritmo C4.5 o pior resultado.

Na base de dados DLBCL-NIH todos os resultados dos algoritmos foram estatisticamente piores que o algoritmo *Naïve Bayes*.

Já na base de dados ALL/AML os algoritmos C4.5 e o *k*-NN apresentaram os piores resultados estatisticamente comparados com o algoritmo base.

Na tabela 52 é mostrada a média geral da aplicação do método de projeção aleatória de cada base de dados.

Já a tabela 53 mostra a média geral dos algoritmos de classificação usando a utilização conjunta do Método de Projeção Aleatória e da Seleção de Atributos. Analisando os resultados obtidos e feita a analise estatística conclui-se que todos os resultados são estatisticamente equivalentes comparados com o algoritmo *Naïve Bayes*.

Se compararmos, ainda, esse resultado com o resultado obtido quando se usou-se todos os atributos observa-se o resultado do algoritmo SVM na utilização conjunta foi inferior, isso pode ser visualizado na figura 56. Porém, analisando estatisticamente esses resultados observa-se que os resultados são estatisticamente equivalentes.

Tabela 50: Média por Algoritmo de Classificação de cada Base de Dados usando o Método de Projeção Aleatória e a Seleção de Atributos – Abordagem Filtro.

	Média Geral dos Algoritmos de Classificação de cada Base de Dados - Abordagem Filtro						
Base de Dados	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	84,0 ± 2,5	76,0 ± 3,9*	84,3± 2,4	77,9 ± 4,9*	80,57 ± 4,3*	81,1 ± 3,8*	81,6 ± 4,0*
DLBCL-Tumor	88,6 ± 2,43	84,3 ± 2,4*	88,1 ± 2,8	87,0 ± 2,2*	88,4 ± 2,4	88,7 ± 2,1	**89,6 ± 1,8**
DLBCL-Outcome	48,4 ± 18,7	51,4 ± 15,1	49,9 ± 15,3	50,3 ± 14,2	51,4 ± 14,8	51,22± 15,1	50,6 ± 15,6
DLBCL-NIH	56,7 ± 10,3	54,7 ± 9,7*	56,0 ± 9,3	51,2 ± 8,6*	52,4 ± 8,8*	53,0 ± 8,8*	53,2 ± 8,9*
ALL/AML	89,6 ± 2,6	80,7 ± 3,6*	89,6 ± 2,1	86,9 ± 2,4*	87,5 ± 2,1*	87,4 ± 2,4*	87,1 ± 2,9*

Tabela 51: Média por Algoritmo de Classificação de cada Base de Dados usando o Método de Projeção Aleatória e a Seleção de Atributos – Abordagem *Wrapper*.

	Média Geral dos Algoritmos de Classificação de cada Base de Dados - Abordagem *Wrapper*						
Base de Dados	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	91,7 ± 0,8	84,3 ± 20,1*	**92,9 ± 2,2**	89,2 ± 0,8*	89,5 ± 0,2*	89,9 ± 0,1*	90,3 ± 0,8*
DLBCL-Tumor	92,4 ± 0,5	88,7 ± 0,2*	90,1 ± 2,4*	**93,5 ± 1,6**	92,9 ± 2,5	93,5 ± 2,3*	93,8 ± 1,1*
DLBCL-Outcome	63,5 ± 11,3	64,0 ± 8,4*	66,2 ± 5,8	**72,1 ± 4,3**	**71,8 ± 4,8**	**70,1 ± 6,1**	**68,2 ± 6,3**
DLBCL-NIH	64,7 ± 1,5	62,4 ± 5,8*	64,0 ± 0,8*	63,9 ± 0,1*	64,3 ± 0,2*	64,3 ± 0,6*	64,3 ± 1,1*
ALL/AML	94,2 ± 0,3	87,6 ± 1,6*	93,9 ± 1,3*	94,0 ± 0,6	93,8 ± 0,3*	91,6 ± 0,4*	91,3 ± 0,3*

Tabela 52: Média Geral dos Algoritmos de Classificação de todas as Bases de Dados usando o Método de Projeção Aleatória e a Seleção de Atributos.

	Média Geral dos Algoritmos de Classificação de cada Base de Dados						
Base de Dados	*Naïve Bayes*	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	86,6 ± 4,4	78,8 ± 5,3*	87,1 ± 4,9	81,7 ± 7,0*	83,6 ± 5,7*	83,9 ± 5,7*	84,5 ± 5,5*
DLBCL-Tumor	89,9 ± 2,7	85,8 ± 2,9*	88,8 ± 2,6*	89,1 ± 3,8	89,9 ± 3,2	90,3 ± 3,1	**91,0 ± 2,6**
DLBCL-Outcome	53,4 ± 17,2	55,6 ± 13,9	55,3 ± 14,8	57,6 ± 15,9	58,2 ± 15,7	57,5 ± 15,4	56,5 ± 15,4
DLBCL-NIH	59,4 ± 9,0	**57,0 ± 9,0**	58,6 ± 8,3	55,5 ± 9,3*	56,4 ± 9,2*	56,8 ± 9,0*	56,9 ± 9,0*
ALL/AML	91,1 ± 3,1	83,0 ± 4,6*	91,1 ± 2,8	89,2 ± 4,2*	89,6 ± 3,6*	88,8 ± 2,9*	88,5 ± 3,1*

Tabela 53: Média Geral dos Algoritmos de Classificação usando o Método de Projeção Aleatória e a Seleção de Atributos.

Média Geral dos Algoritmos de Classificação						
Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
76,1 ± 18,1	72 ± 14,6	76,2 ± 17,6	74,6 ± 16,8	75,5 ± 16,9	75,5 ± 16,9	75,5 ± 17,3

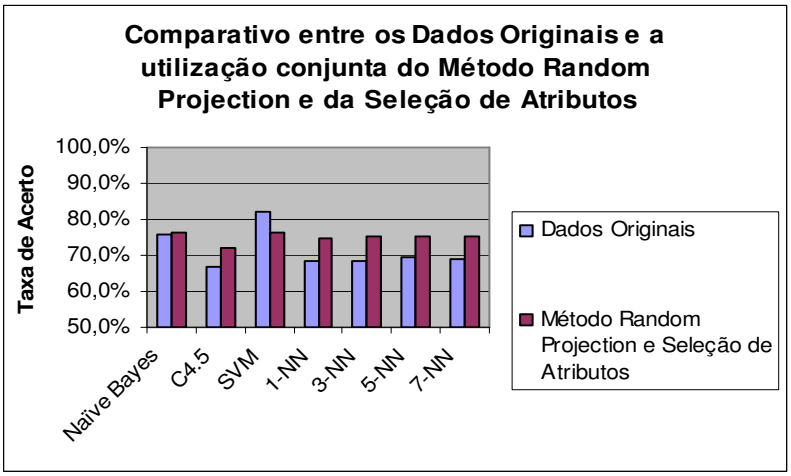

Figura 56: Comparativo entre os Dados Originais e a utilização conjunta do Método de Projeção Aleatória e a Seleção de Atributos.

4.5.4 Comparação entre a Seleção de Atributos e o Método de Projeção Aleatória

Analisando os resultados obtidos na seleção de atributos e no método de projeção aleatória é possível fazer um comparativo entre ambos.

Com a análise estatística, através do teste de hipótese, conclui-se que a seleção de atributos foi melhor que o método de projeção aleatória.

Na figura 57 é possível visualizar o resultado de ambos os métodos de redução de dimensionalidade.

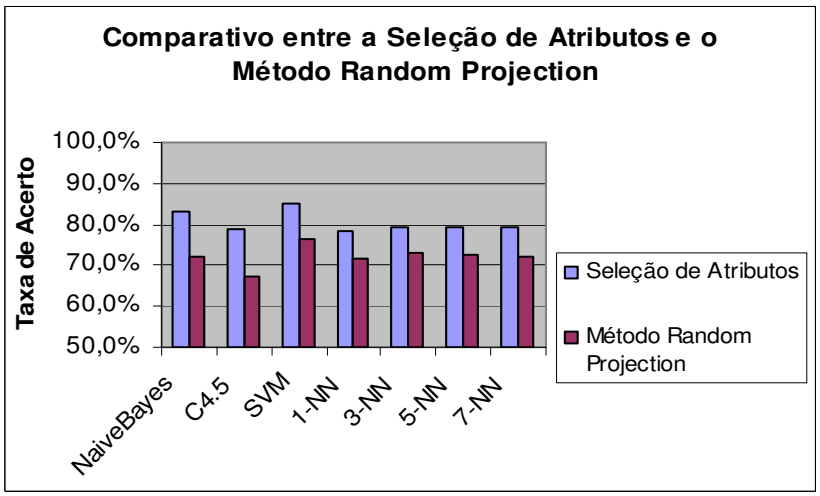

Figura 57: Comparativo entre a Seleção de Atributos e o Método de Projeção Aleatória.

4.5.5 Comparação entre a Seleção de Atributos e a Utilização Conjunta do Método de Projeção Aleatória com a Seleção de Atributos

Analisando os resultados obtidos na seleção de atributos e na utilização conjunta do método de projeção aleatória com a seleção de atributos é possível fazer um comparativo entre os métodos.

Na figura 58 é possível visualizar os resultados de ambos os métodos. Observa-se que os resultados da seleção de atributos é superior ao resultado da utilização conjunta, por todos os algoritmos de classificação. Análises estatística comprovam essa afirmação nos resultados dos algoritmos *Naïve bayes*, C4.5 e SVM, porém o resultado do algoritmo *k*-NN (para os valores de *k* usado) é equivalente nos dois métodos.

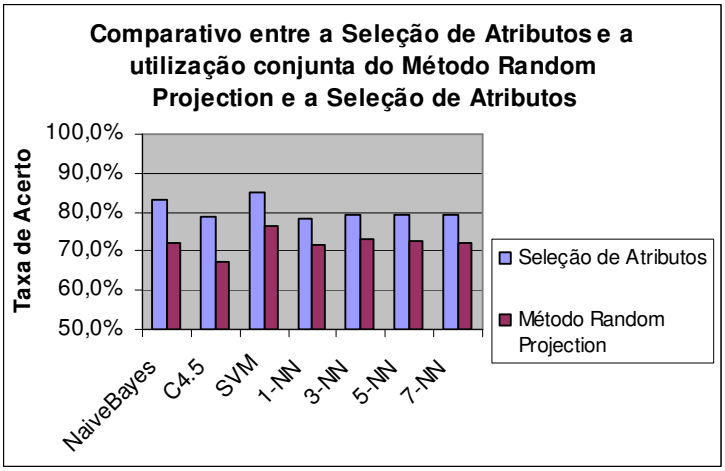

Figura 58: Comparativo entre a Seleção de Atributos e a utilização conjunta Método de Projeção Aleatória com a Seleção de Atributos.

4.5.6 Comparação entre a Base de Dados Original e os Métodos de Redução de Dimensionalidade

Após obter-se a média do resultado de cada método é possível comparar todos os métodos entre si juntamente com a média do resultado da base de dados com todos os atributos (figura 59).

A seleção de atributos foi o método de redução de dimensionalidade que apresentou melhores resultados, seguido da utilização conjunta do método de projeção aleatória com a seleção de atributos.

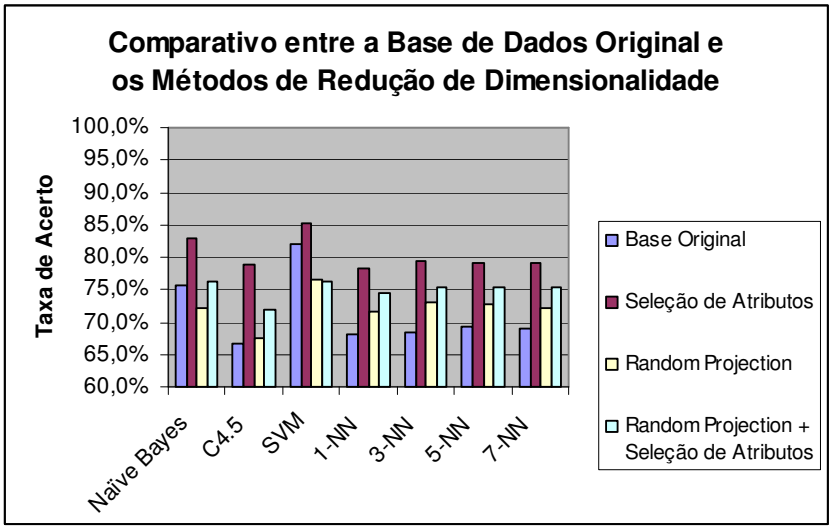

Figura 59: Comparativo entre a Base de Dados Original e os Métodos de Redução de Dimensionalidade.

Através dessa comparação geral feita, foi possível identificar os melhores resultados como mais facilidade. Observa-se que os resultados dos dois métodos de redução de dimensionalidade, a seleção de atributos e o método de projeção aleatória, nas três formas em que foi aplicada teve resultados melhores comparados com os resultados obtidos quando aplicados nas bases de dados com todos os atributos.

Analisando os resultados nota-se que a seleção de atributos foi o método de redução que produziu os melhores resultados. A utilização do método de projeção aleatória contribuiu no tempo de processamento computacional dos algoritmos, porém, o resultado obtido por esse método foi inferior ao resultado da seleção atributos. A utilização conjunta do método de projeção aleatória e da seleção de atributos teve resultados melhores do que a aplicação somente do método de projeção aleatória, porém ainda em uma escala menor que o resultado da seleção de atributos.

5 Conclusão

Existe um pouco de dificuldade em definir algumas técnicas de mineração a serem utilizadas em bases de expressão gênica, obtidas pela técnica de microarranjos, pois este tipo de base requer um tratamento diferenciado devido as suas características no que diz respeito a grande quantidade de atributos e relativamente poucas amostras. Com isso, vários algoritmos de aprendizagem de máquinas tornam-se sua aplicação inviável como é caso das redes neurais artificiais. Assim, métodos de redução de dimensionalidade têm sido utilizados nesses tipos de dados.

Nesse trabalho foram aplicados dois métodos de redução de dimensionalidade: a seleção de atributos e o método de projeção aleatória. Primeiramente cada método de redução foi executado separadamente e depois utilizou-se os dois métodos em conjunto, ou seja, foi aplicado os algoritmos de seleção de atributos no novo conjunto de atributos obtidos pelo método de projeção aleatória. Para a seleção de atributos duas abordagens principais foram usadas: a filtro e a *wrapper*.

Analisando os resultados obtidos por esses métodos observou-se uma melhora significativa nos resultados. Quando usado os algoritmos de seleção de atributos mesmo nos piores casos, a taxa de acerto do classificador foi superior a de quando aplicado sobre as bases de dados com todos os atributos. Além disso, houve uma grande redução na quantidade de atributos selecionados, principalmente, quando se aplicou as medidas de avaliação com a busca seqüencial.

Comparando os resultados obtidos das duas abordagens de seleção de atributos, observou-se a abordagem *wrapper* quando aplicado juntamente com a busca seqüencial produziu melhores resultados, seguida da medida de avaliação dependência pertencente a abordagem filtro. Embora a diferença do resultado dos classificadores tenha sido pequena no que diz a taxa de acerto, o custo computacional foi muito superior quando usou-se a abordagem *wrapper*.

Em geral, a execução dos algoritmos pertencente a abordagem filtro teve um tempo de processamento na ordem de segundos a minutos. Já os algoritmos pertencentes a abordagem *wrapper* teve um tempo de processamento da ordem de horas a dias de processamento, o que pode, em alguns casos, torna-se inviável a sua aplicação.

O algoritmo SVM foi, em geral, o que produziu melhores resultados, porém o seu tempo de processamento foi bastante elevado comparado com os outros algoritmos de classificação devido ao tamanho das bases de dados. Outro algoritmo que teve bons resultados na classificação foi o *Naïve Bayes*.

A aplicação do método de projeção aleatória foi uma tentativa de melhorar o tempo de execução dos algoritmos, principalmente os algoritmos de seleção de atributos, e aumentar ainda mais a taxa de acerto do classificador.

Quando se utilizou o método de projeção aleatória observou um pequeno aumento na taxa de acerto na maioria dos classificadores comparado com os resultados obtidos quando se usou a base de dados com todos os atributos.

Quando se aplicou os dois métodos em conjunto, houve uma melhora nos resultados um pouco maior de quando usado somente o método de projeção aleatória.

Embora o resultado da aplicação do método de projeção aleatória, nos dois casos que foi utilizado, tenha sido superior aos resultados obtidos quando não usado nenhum método de redução, não foi superior ao desempenho dos algoritmos de classificação quando usado somente a seleção de atributos.

Portanto, tem-se como recomendação geral que, a seleção de atributos é um método de redução de dimensionalidade que trás bons resultados quando aplicado em bases de expressão gênica. O método de projeção aleatória é um método alternativo, pois, além de diminuir o custo computacional quando aplicado, principalmente em conjunto com a seleção de atributos, produz bons resultados.

Os resultados dos experimentos comprovam que as aplicações desses métodos de redução de dimensionalidade produzem uma taxa de acerto do classificador maior do que quando aplicado somente o algoritmo de mineração sobre as bases de dados com todos os atributos.

5.1 Trabalhos Futuros

Diversas extensões deste trabalho podem ser exploradas. Uma das propostas é aplicar os métodos de seleção de atributos utilizados nesse trabalho em outras bases de dados para comparar o desempenho dos classificadores entre as bases.

Outra extensão está relacionada em se fazer uma otimização nos algoritmos de classificação utilizados, alterando os parâmetros de entrada dos algoritmos; por exemplo, utilizar outros valores de *k*, para o algoritmo *k*-NN, etc. Também é possível utilizar outros algoritmos de classificação, como os algoritmos baseados em indução de regras, já que esses tipos de algoritmos podem trazer algum conhecimento surpreendente ou inesperado.

Outra possibilidade está na aplicação de outros algoritmos de seleção de atributos como o algoritmo *relief* [KIR92], que é um algoritmo que mede a relevância dos atributos no treinamento, o algoritmo seqüencial para trás, visto que, o que foi aplicado nesse trabalho o método seqüencial para frente.

E finalmente, outro trabalho futuro está na aplicação de outras técnicas de redução de dimensionalidade de dados, bem como a combinação de outros métodos de redução de dimensionalidade. A utilização do método de análise de componentes principais (PCA) [KAR92], que é um método que tem por finalidade básica, a redução de dados a partir de combinações lineares das variáveis originais, em conjunto com a seleção de atributos pode ser uma alternativa a ser aplicada.

Referências

[ACH01] ACHLIOPTAS, D. Database-friendly random projections. In Proc. ACM Symp. on the Principles of Database Systems, 2001, p. 274-281.

[ALB97] ALBERTS, B. et al. Biologia Molecular da Célula. Artes Médicas, 3ª Edição, 1997.

[ALI00] ALIZADEH, A. et al. Distinct types of diffuse large B-cell Lymphoma Identified by gene expression profiling. Nature 4051, p. 503–511, fev., 2000.

[ALM91] ALMUALLIM, H. E DIETTERICH T. G. Learning with Many Irrelevant Features. In Proc. of the 9th National Conf. on Artificial Intelligence, Anaheim, CA, 1991, v. 2, p. 547–552.

[BAL96] BALA, J.; JONG K. DE; HUANG, J.; VAFAIE, H.; WECHSLER, H. Using Learning to Facilite the Evolution of Features for Recognizing Visual Concepts, In: Special Issue of Evolutionary Computatuion – Evolution, learning and Instinct: 100 years of Baldwin Effect, 1996, v. 4, p. 297-311.

[BAL01] BALDI, P.; BRUNAK, S. Bioinformatics: the Machine Learning approach. MIT Press, 2ª Edição, 2001.

[BIN01] BINGHAM, E.; MANNILA, H. Random projection in dimensionality reduction: applications to image and text data. In Knowledge Discovery and Data Mining, 2001, p. 245-250.

[BOR05a] BORGES, H. B.; NIEVOLA, J. C. Attribute Selection Methods Comparison for Classification of Diffuse Large B-Cell Lymphoma. In: The 4th International Conference on Machine Learning and Applications - ICMLA´05, 2005, Los Angeles.v.1,p. 201-206.

[BOR05b] BORGES, H. B.; NIEVOLA, J. C. Attribute Selection Methods Comparison for Classification of Diffuse Large B-Cell Lymphoma. In: VI International Enformatika Conference - IEC2005, 2005, Budapeste, Hungria. v. 8, p.193-197.

[CAS92] CASLEY, D. Primer on molecular biology. Technical report, U. S. Department of Energy, Office of Health and Environmental Research, 1992.

[CRI00] CRISTIANINI, N. E SHAWE-TAYLOR, J. *An Introduction to Support Vector Machinesand other kernel-based learning methods*. 2000, Cambridge University Press.

[DAS97] DASH M. E LIU H. Feature Selection for Classification. Intelligent Data Analysis: An Int'l J. 1(3): 131-156, 1997.

[DEV82] DEVIJVER, P. A. E KITTLER, J. Pattern Recognition – A Statistical Approach. Prentice Hall, 1982, London, GB.

[DUG99] DUGGAN, D. J. et al. Expression profiling using cDNA microarrays. Nature Genetics, 21:10–14, 1999.

[EIS98] EISEN, M. B. et al. Cluster analysis and display of genome-wide expression pattern. In Proc. of National Academy of Sciences USA, 1998, v. 95, p. 14863–14868.

[FAY96] FAYYAD, USAMA M. et al. KDD for science data analysis: issues and examples. Second International Conference on Knowledge Discovery and Data Mining, 1996 Portland, Oregon, Ago.1996, AAAI Press.

[FRE02] FREITAS, A. A., Data Mining and Knowledge Discovery with Evolutionary Algorithms, 2002, Berlin: Springer-Verlag.

[GEO95] GEORGE H. J.; PAT L. Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995. p. 338-345. Morgan Kaufmann, San Mateo.

[GOL89] GOLDBERG, D. E., Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company, 1989.

[GOL99] GOLUB T. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537, 1999.

[GOR02] GORDON et al. Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gege Expression Ratios in Lung Cancer And Mesothelioma. Cancer Research, 62:4963-4967, 2002.

[HAL99] HALL, M. A. Correlation-based Feature Selection for Machine Learning. 198f. 1999. Tese (Doutorado em Ciência da Computação) - Universidade de Waikato.

[HAL00] HALL, M., Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the Seventeenth International Conference on Machine Learning, 2000, p. 359-366.

[HAY99] HAYKIN, S. Neural Networks*: A Comprehensive Foundation*. Prentice Hall, 1999.

[HEA98] HEARST, M. A. et al. Trends and controversies - Support Vector Machines. *IEEE Intelligent Systems*, 1998, 13(4), p.18–28.

[HOL91] HOLSHEIMER, M.; SIEBES, A., Data Mining – The Search for Knowledge in Databases, 1991, Report CS-R9406, Amsterdam.

[KAR92] KARLSSON, A. The use of Principal Component Analysis (PCA) for evaluating results from pig meat quality measurements, *Meat Science,* 1992, v. 31, p.423-33.

[KAS98] KASKI, S. Dimensionality Reduction by Random Mapping. In Proc. Int. Joint Conf. On Neural Networks, 1998, v. 1, p. 413-418.

[KIR83] KIRKPATRICK, S. Optimization by Simulated Annealing – Quantitative Studies", J. Stat. Phys., 1984, p.34.

[KIR92] KIRA, K.; RENDELL, L. A., The Feature Selection Problem: Traditional Methods and a New Algorithm, In: Proc. 10th Conference on Artificial Intelligence, 1992, p. 129-136, Menlo Park, CA.

[KOH98] KOHAVI, R.; JOHN, G. H., The Wrapper Approach, In: H. Liu and H. Motoda (Eds.) Feature Extraction, Construction and Selection: a data mining perspective, 1998, p. 33-49.

[KUR99] KURIMO, M. Indexing audio documents by using latent semantic analysis and SOM. In E. Oja and S. Kaski, editors, Kohonen Maps, p. 363–374. Elsevier, 1999.

[LAP05] LAPPONI, J. C. Estatística usando Excel. Rio de Janeiro: Elsevier, 2005.

[LAU03] LAU, M., SCHULTZ, M. A Feature Selection Method for Gene Method for Gene Expression Data with Thousands of Features. 2003, Yale University, New Heven, CT 06511.

[LEW01] LEWIS, R. Human Genetics - Concepts and Applications. Mc Graw Hill, 4ª Edição, 2001.

[LIM02] LIMA, A. R. G. Máquinas de Vetores Suporte na Classificação de Impressões Digitais. 81f. 2002. Dissertação (Departamento de Computação) Universidade Federal do Ceará, Fortaleza-CE.

[LIN03] LIN, J.; GUNOPULOS, D. Dimensionality Reduction by Random Projection and Latent Semantic Indexing. In proceedings of the Data Mining Workshop, at the 3th SIAM International Conference on Data Mining. 2003, San Francisco, CA. Mai 3.

[LIU02] LIU, H., LI, J., WONG, L. A comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome Informatics 13:51-60, 2002.

[LIU98] LIU, H., MOTODA, H., Feature Selection for Knowledge Discovery and Data Mining, 1998, Kluwer academic Publishers.

[LIU03] LIU, H., MOTODA, H., YU, L., The Handbook of Data Mining, Lawrence Erlbaum Associates, 2003, Inc. Publishers. Editor: N. Y., p.409-423.

[LIU96] LIU, H.; SETIONO, R. A Probabilistic Approach to Feature Selection: a Filter Solution. In Proc. of the 13th Int. Conf.on Machine Learning, 1996, p. 319–327. Morgan Kaufmann.

[LIU98] LIU, H.; SETIONO, R. Scalable Feature Selection for Large Sized Databases. In Proc. of the 4th World Congress on Expert System, 1998, p. 68–75. Morgan Kaufmann.

[LIU05] LIU H. E YU L. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering. 2005, 17(4): p. 491-502.

[MET53] METROPOLIS et. al. Equations of State Calculations by Fast Computing Mavhines. J. Chem. Phys., 1953, p.1087-1092.

[MIL02] MILLER et al. Optimal gene expression analysis by microarrays. Cancer Cell. 2(5):353-61, Nov., 2002.

[MÜL01] MÜLLER, K. R. et al. An introduction to kernel-based learning algorithms. *IEEE Trans. on Neural Networks*, 2001, 12(2):181–201.

[MIT97] MITCHELL, T., Machine Learning. New York, United States of America: McGraw-Hill, 1997.

[MOL02] MOLINA L. C., BELANCHE L., NEBOT A. Feature Selection Algorithms: A Survey and experimental Evaluation. 2002, Technical Report LSI-02-62-R, Universidade Politécnica de Catalunya, Barcelona, Espanha.

[PAE98] PAES, A.T. Itens essenciais em bioestatística. Arquivos Brasileiros de Cardiologia, São Paulo, SP, 1998, v. 71, n. 4, p. 575-580.

[PAG04] PAGANO, M., GAUVREAU, K. Princípios de Bioestatística. São Paulo: Pioneira Thonson Learning, 2004.

[PAP98] PAPADIMITRIOU, C. H.; RAGHAVAN, P.; TAMAKI, H.; VEMPALA, S. Latent semantic indexing: A probabilistic analysis. In 17th Annual Symposium on Principles of Database Systems, Seattle, WA, 1998, p. 159–168.

[POM02] POMEROY, S. L. et al. Prediction of central nervous system embryonal tumor outcome based on gene expression. Nature Lond., *415:* 436–442, 2002.

[QUI86] QUINLAN, J. R. Induction of decision trees. Machine Learning, 1986, v.1, n.1, p. 81-106.

[QUI93] QUINLAN, J. R. C4.5: Programs for Machine Learning. 1993, San Mateo, CA: Morgan Kaufmann Publishers.

[ROS02] ROSENWALD, A. et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med, 346:25:1937-1947, 2002.

[RUB03] RUBINSTEIN et al. Machine Learning in Low-level Microarray Analysis. Position paper in *ACM SIGKDD Explorations (Special Issue on Microarray Data Mining)*, 5(2), December, 2003.

[RUS04] RUSSEL S.; NORVIG P., Inteligência Artificial. Tradução da 2ª Edição, Editora Campus, 2004.

[SHA03] SHAPIRO, G. P; KHABAZA. T.; RAMASWAMY S. Capturing best practice for microarray gene expression data analysis. 2003, p.407-415 Electronic Edition (DOI: 10.1145/956797) BibTeX.

[SCU04] SCUSE, D E REUTEMANN, P. Weka Experimenter Tutorial for Version 3.4, Mar 8, 2001. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/ - Acessado em: 10/09/2004.

[SCU06] SCUSE, D E REUTEMANN, P. Weka Experimenter Tutorial for Version 3.4, Feb. 16, 2006. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/ - Acessado em: 20/03/2006.

[SET97] SETÚBAL, J. C.; MEIDANIS, J. Introduction to Computational Molecular Biology. PWS Publishing Company, 1997.

[SHI02] SHIPP, M. et al. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nature Medicine 8:1:68-74, 2002.

[SMO02] SMOLA, A. J.; SCHÖOLKOPF, B. *Learning with Kernels*. The MIT Press, 2002, Cambridge, MA.

[SOU03] SOUTO M. C. P., LORENA A. C., DELBEM A. C. B., CARVALHO A. C. P. L. F. Técnicas de Aprendizado de Máquina para Problemas de Biologia Molecular. p.103-152.

Editora SBC. Disponível em: http://www.dimap.ufrn.br/~marcilio/ENIA2003/jaia2003-14-08.pdf - Acessado em: 08/05/2006.

[TAM99] TAMAYO, P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. In Proc. Natl. Acad. Sci. USA, 1999, 96:2907–2912.

[TAM03] TAMAYO, G. P., TAMAYO, P. Microarray Data Mining: Facing the Challenges. SIGKDD Explorations, 2003, v. 5.

[VAP95] VAPNIK, V. N. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[YUL04] YU, L., LIU, H. Redundancy Based Feature Selection for Microarray Data. In Proceedings of the Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2004, p. 737-742.

[WAN05] WANG et al. "Gene selection from microarray data for cancer classification - a machine learning approach", Computational Biology and Chemistry, 29(1):37-46, 2005.

[WEN98] WEN, X. et al. Large-scale temporal gene expression mapping of central nervous system development. In Proc. Natl. Acad. Sci. USA, 1998, v. 95, p. 334-339.

[WIT05] WITTEN I. H.; IAN H.; FRANK, E. Data Mining: Practical machine learning tools and techniques, 2nd Edition, 2005, Morgan Kaufmann, San Francisco.

[WON02] WONG, L. Datamining: Discovering Information from Bio-data. *Current Topics in Computational Biology*, edited by Tao Jiang, Ying Xu, and Michael Zhang, 2002, c. 13, p. 317-342, MIT Press, Cambridge, MA.

Apêndice A

Genes Selecionados

Os genes são organizados conforme o número de vezes que foram selecionados nos experimentos de seleção de atributos. Dessa forma, são mostrados os genes que apareceram ao menos oito vezes nos experimentos. A tabela A.1 refere-se aos genes mais selecionados da base de dados DLBCL, a tabela A.2 aos genes selecionados da base de dados DLBCL-Tumor, a tabela A.3 os genes selecionados da base de dados DLBCL-Outcome, a tabela A.4 os genes selecionados da base de dados DLBCL-NIH e a tabela A.5 os genes selecionados da base de dados ALL/AML.

É importante deixar claro que os genes identificados são baseados na identificação conforme o autor de cada base de dados. Maiores informações sobre os genes podem ser encontradas em sites como o NCBI (*National Center for Biotechnology Informations*)[4].

Esses genes selecionados podem ser o inicio de um grande estudo na área biológica, auxiliando os especialistas na identificação de possíveis genes que estão diretamente relacionados com a doença (nesse caso os dois tipos de bases de dados de câncer estudado: o linfoma e a leucemia).

Tabela A1: Genes Selecionados com mais freqüência pelos Métodos de Seleção de Atributos da Base de Dados DLBCL.

Base de Dados DLBCL		
Atributo	Gene	Anotação Gênica
GENE3330X	19288	*Unknown; Clone=825199
GENE2065X	21008	*Unknown UG Hs.27774 ESTs; Clone=1269030
GENE2283X	13348	(Unknown UG Hs.44628 ESTs; Clone=1333781)
GENE3262X	20569	*Unknown UG Hs.186709 ESTs, Weakly similar to !!!! ALU SUBFAMILY SB WARNING ENTRY !!!! [H.sapiens]; Clone=1341225
GENE3388X	16299	*Immunoglobulin J chain; Clone=117806
GENE2807X	17314	*N-CoR=transcriptional corepressor; Clone=783355
GENE2746X	16245	*RLF=Zn-15 related zinc finger protein; Clone=53251
GENE2905X	15542	(KIAA0594; Clone=1370282)
GENE349X	15149	(Unknown; Clone=1371980)
GENE2952X	14009	(Unknown; Clone=1340660)
GENE910X	16942	(CA150=putative transcription factor; Clone=471761)
GENE448X	17399	*TFAR15=apoptosis-related protein; Clone=950376
GENE802X	16122	*MSH2=DNA mismatch repair mutS homologue; Clone=630013

[4] http://www.ncbi.nlm.nih.gov/

GENE632X	20420	*ATP5A=mitochondrial ATPase coupling factor 6 subunit; Clone=825312
GENE505X	21187	(Unknown UG Hs.56421 ESTs, Weakly similar to Similarity to H.influenza ribonuclease PH [C.elegans]; Clone=1300230)
GENE72X	20289	(Unknown UG Hs.193922 ESTs, Weakly similar to !!!! ALU SUBFAMILY SP WARNING ENTRY !!!! [H.sapiens]; Clone=686395)
GENE3466X	21415	(Unknown UG Hs.84153 dynamitin (dynactin complex 50 kD subunit); Clone=1338059)
GENE3607X	16637	*Similar to SER/THR-protein kinase NEK2=NIMA-related protein kinase 2; Clone=280376

Tabela A2: Genes Selecionados com mais freqüência pelos Métodos de Seleção de Atributos da Base de Dados DLBCL-Tumor.

Base de Dados DLBCL - Tumor	
Atributo	Anotação Gênica
D55716_at	DNA REPLICATION LICENSING FACTOR CDC47 HOMOLOG
AFFX-HUMTFRR/M11507_3_at	AFFX-HUMTFRR/M11507_3_at (endogenous control)
D87119_at	Cancellous bone osteoblast mRNA for GS3955
HG1733-HT1748_at	Moloney Murine Sarcoma Viral Oncogene Homolog
HG4144-HT4414_at	Zinc Finger Protein Hzf6
J02888_at	NMOR2 Quinone oxidoreductase (NQO2)
L07590_at	PPP2R3 Protein phosphatase 2 (formerly 2A), regulatory subunit B" (PR 72), alpha isoform and (PR 130), beta isoform
L14269_at	SLC18A2 Solute carrier family 18 (vesicular monoamine), member 2
L20815_at	S protein mRNA
L37127_at	(clone mf.18) RNA polymerase II mRNA
M14218_at	ASL Argininosuccinate lyase
M24194_at	Alpha-tubulin mRNA
M68516_rna1_at	PCI gene (plasminogen activator inhibitor 3) extracted from Human protein C inhibitor gene
M99564_at	P PROTEIN
U21931_at	FBP1 Fructose-bisphosphatase 1
U22963_at	Class I histocompatibility antigen-like protein mRNA
U50535_at	BRCA2 region, mRNA sequence CG005
X66899_at	EWSR1 Ewing sarcoma breakpoint region 1
X76538_at	MPV17 MpV17 transgene, murine homolog, glomerulosclerosis
X82693_at	E48 antigen
Z33642_at	V7 mRNA for leukocyte surface protein
D16105_at	LTK Leukocyte tyrosine kinase
D86988_at	KIAA0221 gene
L04569_at	Calcium channel L-type alpha 1 subunit (CACNL1A1) mRNA
HG2809-HT2920_s_at	Lung Surfactant Protein D
HG3417-HT3600_s_at	Gtp Cyclohydrolase I, Alt. Splice 1
U58837_s_at	CNCG2 Cyclic nucleotide gated channel (photoreceptor), cGMP gated 2 (beta)
M11313_s_at	A2M Alpha-2-macroglobulin
HG3636-HT3846_at	Myosin, Heavy Polypeptide 9, Non-Muscle

Tabela A3: Genes Selecionados com mais freqüência pelos Métodos de Seleção de Atributos da Base de Dados DLBCL-Outcome.

Base de Dados DLBCL - Outcome	
Atributo	Anotação Gência
L42354_at	(clone 48ES4) mRNA fragment
M19720_rna1_at	L-myc gene (L-myc protein) extracted from Human L-myc protein gene
U49278_at	Putative DNA-binding protein mRNA, partial cds
L09229_s_at	FACL1 Long chain fatty acid acyl-coA ligase
U49020_cds2_s_at	MEF2A gene (myocyte-specific enhancer factor 2A, C9 form) extracted from Human myocyte-specific enhancer factor 2A (MEF2A) gene, first coding

Tabela A4: Genes Selecionados com mais freqüência pelos Métodos de Seleção de Atributos da Base de Dados DLBCL-NIH.

Base de Dados DLBCL - NIH	
Atributo	Anotação Gência
31618	|NM_016403|*AA458914|Hs.42743|hypothetical protein HSPC148
16804	|||LC_16804
28394	|X98296|~R93207|Hs.77578|ubiquitin specific protease 9, X chromosome (fat facets-like Drosophila)
32869	|U50196|*AA748750|Hs.94382|adenosine kinase
22131	||*AI083695||

Tabela A5: Genes Selecionados com mais freqüência pelos Métodos de Seleção de Atributos da Base de Dados ALL/AML.

Base de Dados ALL/AML		
Atributo	Gene	Anotação Gênica
attribute134	AF005043_at	Poly(ADP-ribose) glycohydrolase (hPARG) mRNA
attribute1497	L35240_at	Enigma gene
attribute83	AB002365_at	KIAA0367 gene, partial cds
attribute189	D12686_at	EIF4G Eukaryotic translation initiation factor 4 (eIF-4) gamma
attribute618	D83735_at	Adult heart mRNA for neutral calponin
attribute3650	U71092_at	GB DEF = Somatostatin receptor-like protein (GPR24) gene
attribute4973	Y08612_at	RABAPTIN-5 protein
attribute5103	Z27113_at	DNA-DIRECTED RNA POLYMERASE II 14.4 KD POLYPEPTIDE
attribute6091	L25286_s_at	COL15A1 Collagen, type XV, alpha 1
attribute6822	HG3636-HT3846_at	Myosin, Heavy Polypeptide 9, Non-Muscle

Apêndice B

Quantidade de Atributos Selecionados

Tabela B1: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL transformados pelo Método de projeção aleatória utilizando um número fixo de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados DLBCL – Números Fixos de Atributos				
Subconjuntos	10 Atributos	15 Atributos	30 Atributos	45 Atributos	71 Atributos
1.1	2	2	2	5	7
1.2	1	2	2	3	4
1.3	3	3	4	4	5
1.4	2	2	2	3	3
1.5	4	4	5	5	6
1.6	3	3	3	3	4
1.7	2	2	3	3	4
1.8	3	2	2	4	5
1.9	3	3	3	4	4
2.1	2	2	3	6	17
2.2	2	4	8	15	23
2.3	4	7	12	19	32
2.4	3	4	8	12	24
2.5	5	7	14	18	30
2.6	6	7	12	20	30
2.7	5	7	14	21	31
2.8	6	7	13	20	33
2.9	5	7	13	20	32

Tabela B2: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL transformados pelo Método de projeção aleatória utilizando a porcentagem de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados DLBCL – Porcentagem de Atributos				
Subconjuntos	3% Atributos	10% Atributos	20% Atributos	25% Atributos	50% Atributos
1.1	11	24	27	31	33
1.2	4	4	3	3	3
1.3	4	5	5	5	4
1.4	3	3	3	3	2
1.5	4	4	4	3	4
1.6	5	4	4	4	4
1.7	4	4	5	4	4
1.8	5	3	4	3	3
1.9	4	4	4	4	4
2.1	18	123	192	247	235
2.2	37	56	68	79	176
2.3	48	162	331	384	752
2.4	33	136	251	336	681
2.5	48	158	284	371	670
2.6	47	153	284	357	749
2.7	57	161	310	413	686
2.8	50	165	341	441	711
2.9	49	146	347	446	726

Tabela B3: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL-Tumor transformados pelo Método de projeção aleatória utilizando um número fixo de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados DLBCL-Tumor – Números Fixos de Atributos				
Subconjuntos	10 Atributos	15 Atributos	30 Atributos	50 Atributos	71 Atributos
1.1	3	4	7	10	13
1.2	1	2	4	5	5
1.3	2	3	5	5	4
1.4	1	1	2	3	3
1.5	0	0	1	2	3
1.6	3	3	4	6	5
1.7	4	3	4	4	4
1.8	4	4	5	5	6
1.9	3	3	5	5	5
2.1	3	4	5	8	21
2.2	3	4	10	17	23
2.3	5	7	13	19	27
2.4	4	5	8	14	25
2.5	5	7	17	23	32
2.6	4	7	13	21	31
2.7	5	8	14	18	30
2.8	5	7	12	20	31
2.9	5	8	13	20	32

Tabela B4: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL-Tumor transformados pelo Método de projeção aleatória utilizando a porcentagem de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados DLBCL-Tumor – Porcentagem de Atributos				
Subconjuntos	3% Atributos	10% Atributos	20% Atributos	25% Atributos	50% Atributos
1.1	24	41	56	59	82
1.2	4	3	3	3	3
1.3	5	4	5	5	4
1.4	3	3	3	3	2
1.5	5	4	4	5	4
1.6	6	4	4	4	4
1.7	5	4	4	4	4
1.8	4	4	4	4	3
1.9	5	4	4	4	3
2.1	74	214	373	456	449
2.2	47	61	88	99	69
2.3	87	218	463	566	660
2.4	85	251	517	511	1023
2.5	90	257	527	606	934
2.6	89	286	492	642	1117
2.7	80	204	397	555	631
2.8	89	279	527	603	816
2.9	94	264	508	681	1031

Tabela B5: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL-Outcome transformados pelo Método de projeção aleatória utilizando um número fixo de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados DLBCL-Outcome – Números Fixos de Atributos				
Subconjuntos	10 Atributos	15 Atributos	30 Atributos	45 Atributos	71 Atributos
1.1	0	0	0	0	1
1.2	0	0	0	0	1
1.3	2	2	2	3	4
1.4	1	1	2	3	4
1.5	1	1	3	3	5
1.6	2	2	2	2	2
1.7	3	2	2	3	3
1.8	2	2	2	3	3
1.9	2	2	2	2	2
2.1	0	0	0	0	3
2.2	1	1	1	1	12
2.3	2	4	7	8	23
2.4	1	4	7	10	23
2.5	4	5	10	12	25
2.6	3	6	12	14	25
2.7	4	5	12	15	25
2.8	4	5	9	14	29
2.9	4	5	8	12	25

Tabela B6: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL-Outcome transformados pelo Método de projeção aleatória utilizando a porcentagem de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados DLBCL-Outcome – Porcentagem de Atributos				
Subconjuntos	3% Atributos	10% Atributos	20% Atributos	25% Atributos	50% Atributos
1.1	1	6	12	13	24
1.2	1	4	6	6	6
1.3	4	4	5	4	4
1.4	5	5	6	6	6
1.5	6	6	6	6	6
1.6	3	3	3	3	5
1.7	4	3	3	3	3
1.8	3	4	4	4	3
1.9	3	4	4	5	4
2.1	39	153	319	302	654
2.2	53	275	611	776	1554
2.3	79	183	344	570	821
2.4	86	246	515	549	1060
2.5	93	294	357	642	725
2.6	90	225	455	554	517
2.7	87	242	525	547	779
2.8	89	264	457	548	976
2.9	98	248	424	498	690

Tabela B7: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL-NIH transformados pelo Método de projeção aleatória utilizando um número fixo de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados DLBCL-NIH – Números Fixos de Atributos				
Subconjuntos	10 Atributos	15 Atributos	30 Atributos	50 Atributos	71 Atributos
1.1	0	0	0	1	1
1.2	0	0	0	0	0
1.3	2	3	3	4	5
1.4	1	1	1	3	6
1.5	0	0	0	0	0
1.6	3	2	2	2	2
1.7	2	2	1	2	2
1.8	2	2	1	2	2
1.9	1	2	2	2	3
2.1	0	0	0	3	14
2.2	1	1	1	3	26
2.3	4	6	8	16	41
2.4	1	2	6	13	30
2.5	3	4	9	19	43
2.6	5	6	12	21	44
2.7	5	7	13	18	47
2.8	5	6	11	19	45
2.9	4	6	11	19	44

Tabela B8: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados DLBCL-NIH transformados pelo Método de projeção aleatória utilizando a porcentagem de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados DLBCL - NIH – Porcentagem de Atributos				
Subconjuntos	3% Atributos	10% Atributos	20% Atributos	25% Atributos	50% Atributos
1.1	5	10	13	16	19
1.2	4	10	11	15	17
1.3	5	4	5	5	5
1.4	11	10	11	11	10
1.5	2	6	6	6	7
1.6	4	2	2	2	3
1.7	3	3	3	3	3
1.8	2	4	5	4	5
1.9	3	3	4	2	3
2.1	112	214	131	318	590
2.2	165	426	570	800	972
2.3	172	361	397	716	648
2.4	169	321	370	512	597
2.5	171	375	458	683	846
2.6	170	432	427	615	805
2.7	191	435	463	691	932
2.8	163	406	443	600	758
2.9	169	353	407	553	630

Tabela B9: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados ALL/AML transformados pelo Método de projeção aleatória utilizando um número fixo de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados ALL/AML – Números Fixos de Atributos				
Subconjuntos	10 Atributos	15 Atributos	30 Atributos	45 Atributos	71 Atributos
1.1	3	4	7	10	15
1.2	2	3	4	6	6
1.3	4	5	6	6	6
1.4	2	2	3	4	5
1.5	3	4	7	7	7
1.6	4	4	5	5	5
1.7	3	4	4	6	6
1.8	4	4	4	4	5
1.9	3	4	4	6	5
2.1	3	4	6	8	17
2.2	3	5	10	15	25
2.3	5	7	16	22	31
2.4	3	4	7	13	24
2.5	6	8	16	20	32
2.6	6	8	14	22	32
2.7	5	9	14	22	31
2.8	6	7	14	21	32
2.9	5	8	13	20	33

Tabela B10: Média de Atributos Selecionados pelos Métodos de Seleção de Atributos nos Subconjuntos de Atributos da Base de Dados ALL/AML transformados pelo Método de projeção aleatória utilizando a porcentagem de atributos.

	Média de Nº. de Atributos Selecionados nos Subconjuntos de Atributos da Base de Dados ALL/AML– Porcentagem de Atributos				
Subconjuntos	3% Atributos	10% Atributos	20% Atributos	25% Atributos	50% Atributos
1.1	25	32	39	42	35
1.2	5	4	4	4	3
1.3	4	4	4	4	4
1.4	4	3	3	3	3
1.5	5	4	4	4	4
1.6	5	4	4	4	4
1.7	5	5	4	4	4
1.8	5	5	4	4	4
1.9	5	4	5	4	4
2.1	83	214	315	404	532
2.2	46	54	87	99	69
2.3	90	267	479	664	1146
2.4	88	235	471	613	957
2.5	88	276	441	550	855
2.6	101	275	535	583	1179
2.7	99	278	560	516	929
2.8	99	274	588	673	1156
2.9	95	228	482	596	835

Apêndice C

Teste t Pareado

Para podermos comparar a performance dos algoritmos de aprendizagem e analisar o se o seu resultado é estatisticamente significante, ou que tem significância estatística é necessária a aplicação de técnicas estatísticas. Alguns dos métodos mais aplicados são conhecidos como teste de hipóteses. Ele nos auxilia a decidir se o valor postulado da média parece estar correto ou não além de fornecer o valor de p específico.

Pode-se definir como hipóteses questões levantadas relacionadas ao problema em estudo e que, se respondidas, podem ajudar a solucioná-lo. O papel fundamental da hipótese na pesquisa cientifica é sugerir explicações para os fatos. Uma vez formuladas as hipóteses essas devem ser comprovadas ou não através do estudo com a ajuda de testes estatísticos. Num teste estatístico são formuladas duas hipóteses chamada hipótese nula (H_0) e hipótese alternativa (H_1). Hipótese nula é aquela que é colocada a prova, enquanto a hipótese alternativa é aquela que será considerada como aceitável, caso a hipótese nula seja rejeitada.

Todo teste de hipótese possui erros associados a ele. Um dos mais importantes é chamado “erro tipo 1” que corresponde à rejeição da hipótese nula quando esta for verdadeira. A probabilidade do erro tipo I chama-se nível de significância e é expressa através da letra grega α. O nível de significância adotado foi de 0,05 ou 5%.

O “valor do p” ou *p-value* é conhecido na estatística como nível descritivo e está associado ao que chamamos de testes de hipóteses.

Formalmente, o nível descritivo (p) é definido como o “menor nível de significância (α) que pode ser assumido para rejeitar H_0”. É importante ressaltar que o nível de significância α é um valor arbitrado previamente, enquanto que o nível descritivo (p) é calculado de acordo com os dados obtidos. A grande vantagem de se utilizar o nível descritivo é a possibilidade de “quantificar” a significância, ou seja, no lugar de uma resposta do tipo “sim ou não” temos a informação de “quanto” [PAE98].

Devido ao tipo de resultados que se obteve nos experimentos foi feita uma comparação de esquemas utilizando a estatística do teste-t pareado, que segue a distribuição *t de Student*. O objetivo fundamental do teste t para amostras pareadas é avaliar o comportamento das diferenças observadas em cada elemento. Dessa forma, em

vez de se considerar os dois conjuntos de observações como amostras distintas, focalizar-se a diferença de medições dentro de cada par. Suponha-se que os dois grupos de observações sejam:

Amostra 1	Amostra 2
X_{11}	X_{12}
X_{21}	X_{22}
X_{31}	X_{32}
:	:
:	:
X_{n1}	X_{n2}

Nessas amostras, X_{11} e X_{12} são um par, X_{21} e X_{22} são outro par e assim por diante. Usa-se esses dados para criar um novo conjunto de observações que representam as diferenças dentro de cada par:

$d_1 = X_{11} - X_{12}$

$d_2 = X_{21} - X_{22}$

$d_3 = X_{31} - X_{32}$

:

:

$d_n = X_{n1} - X_{n2}$

Em vez de analisar as observações individuais usa-se a diferença entre os membros de cada um dos pares como a variável de interesse. Como a diferença é uma única medida, a análise reduz-se ao problema de uma amostra e aplica-se o procedimento do teste de hipóteses para uma amostra.

Para fazê-lo, primeiro verifica-se que a média do conjunto de diferenças é:

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n} \qquad \text{(C.1)}$$

Essa média da amostra fornece uma estimativa por ponto para a verdadeira diferença das médias das populações $\mu_1 - \mu_2$. O desvio padrão das diferenças é:

$$S_d = \sqrt{\frac{\sum_{i=1}^{n} (d_1 - \overline{d})^2}{n-1}} \qquad \text{(C.2)}$$

Se a verdadeira diferença das médias das populações for representada por $\delta = \mu_1 - \mu_2$ e deseja-se testar se essas duas médias são iguais, pode-se escrever a hipótese nula como $H_0 : \delta = 0$ e a alternativa como $H_A : \delta \neq 0$

Ao se assumir que a população das diferenças é normalmente distribuída, H_0 pode ser testada calculando a estatística

$$t = \frac{\bar{d} - \delta}{S_d / \sqrt{n}} \tag{C.3}$$

Nota-se que $S_d / \sqrt{n}$ é o erro padrão de $\bar{d}$. Se a hipótese nula é verdadeira, essa quantidade tem uma distribuição t com $n-1$ grau de liberdade. Compara-se o resultado de t com os valores da tabela da Distribuição t para encontrar p, a probabilidade de se observar uma diferença media tão grande ou maior que $\bar{d}$ sendo $\delta = 0$. Se $p \leq \alpha$ rejeita-se H_0 . Se $p > \alpha$ não se rejeita a hipótese nula [PAG04], [LAP05].