Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Seleção de Características através de Nearest Shrunken Centroids

Diego Ricardo de Araujo

JUIZ DE FORA
DEZEMBRO, 2011

Seleção de Características através de Nearest Shrunken Centroids

Diego Ricardo de Araujo

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Custódio Gouvêa Lopes da Motta

JUIZ DE FORA

DEZEMBRO, 2011

Seleção de Características através de Nearest Shrunken Centroids

Diego Ricardo de Araujo

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Custódio Gouvêa Lopes da Motta
Doutor em Engenharia Civil/Sistemas Computacionais - COPPE/UFRJ

Saulo Moraes Villela
Mestre em Engenharia de Sistemas e Computação - COPPE/UFRJ

Carlos Cristiano Hasenclever Borges
Doutor em Engenharia Civil - COPPE/UFRJ

JUIZ DE FORA

08 DE DEZEMBRO, 2011

Ao meu pai (in memorian),
pelos ideais e valores transmitidos.

Resumo

A tarefa de classificação de dados faz parte de uma gama de problemas de grande importância atualmente: a descoberta de conhecimento. Em muitos casos, o processo de classificação pode se tornar complexo devido ao alto nível dimensional do modelo de dados envolvido que, em geral, interfere negativamente nos aspectos de desempenho e acurácia dos classificadores utilizados. Uma alternativa para tratar esse tipo de problema é conhecida como seleção de características, que reduz a dimensionalidade através da identificação dos atributos mais significativos à classificação, excluindo os demais do processo. O principal objetivo do presente trabalho é a realização de um estudo a respeito do impacto causado pelo processo de seleção de características em problemas de classificação de dados. Para tal, foi implementado um sistema inteligente que utiliza o método de seleção de características chamado *Nearest Shrunken Centroids*. Além de uma descrição detalhada sobre o funcionamento do método, são apresentados, também, os resultados de testes comparativos realizados em diferentes bases de dados e diversos classificadores disponíveis atualmente. Finalizando, o sistema é disponibilizado por meio de uma licença de software livre.

Palavras-chave: Classificação de Dados, Seleção de Características, *Nearest Shrunken Centroids*.

Abstract

The data classification task takes part of several important problems at the present days: the Knowledge Discovery in Databases, or KDD. In many cases, the classification process can become complex due to the high-level dimensional data model involved that, usually, impairs aspects of performance and accuracy of the classifiers. An alternative to treat this problem is known as feature selection, which reduces the dimensionality by identifying the most important atributes to the classification process. The main objective of this work is to conduct a study on the impact caused by the feature selection process in data classification problems. For this, an intelligent system was implemented using the so-called Nearest Shrunken Centroids method, which performs the classification and feature selection. Besides a detailed description of the method's operation, are also presented the comparative tests results realized using different databases and some classifiers currently available. Finally, the system is released under a free software license.

Keywords: Classification, Feature Selection, Nearest Shrunken Centroids.

Agradecimentos

A minha família e amigos, pelo encorajamento e apoio.

Ao professor Custódio pela orientação, amizade e paciência, sem as quais este trabalho não se realizaria.

Aos professores do Departamento de Ciência da Computação por seus ensinamentos e aos funcionários do curso que, durante esses anos, contribuíram de alguma forma para meu enriquecimento pessoal e profissional.

"Ser-se livre não é fazermos aquilo que queremos, mas querer-se aquilo que se pode".

Jean-Paul Sartre

Sumário

Lista de Figuras

Lista de Abreviações

DCC	Departamento de Ciência da Computação
UFJF	Universidade Federal de Juiz de Fora
DM	*Data Mining*
MD	Mineração de Dados
KDD	*Knowledge Discovery in Databases*
AM	*Aprendizado de Máquina*
SVM	*Support Vector Machine*
NSC	*Nearest Shrunken Centroid*

1 Introdução

A tarefa de classificação é um problema globalmente reconhecido e que está presente em uma grande gama de aplicações atuais, tais como detecção de mensagens de spam em e-mails e o diagnóstico de doenças, notoriamente o câncer. No entanto, muitas das vezes, devido ao alto nível de complexidade dos problemas envolvidos, a grande quantidade de informação contida no modelo de dados oferecido e a pouca quantidade de amostras para treinamento representam grandes desafios ao processo de classificação.

Em problemas complexos com muitos atributos irrelevantes e/ou redundantes, o modelo de dados se torna desnecessariamente volumoso, repercutindo em um baixo desempenho no processo de análise dos dados. Em casos em que o espaço amostral da base de dados é reduzido, muitas vezes devido à dificuldade ou impossibilidade de coleta de dados, o processo de descoberta de conhecimento pode indicar um número indesejado de falsos positivos causado pelo processo de aprendizado (Klassen and Kim, 2009).

Este fenômeno, conhecido como *overfitting*, deve-se a ocorrência de super-aprendizado de informações contidas nas amostras provenientes da base de dados, que se revelam como inconsistentes, incoerentes ou que, de fato, não generalizam o conhecimento que se deseja extrair da realidade (Goldschmidt and Passos, 2005).

Surge então a necessidade de se identificar as características mais significativas, expressas pelo modelo de dados, de forma a eliminar informações irrelevantes, ou até mesmo prejudiciais, ao processo de descoberta de conhecimento. Tal processo, conhecido como Seleção de Características (*Feature Selection*), se torna um importante e decisivo passo aplicado previamente ao processo de classificação e tem recebido recentemente grande atenção no contexto científico.

Atualmente uma imensa quantidade dados é gerada diariamente, sendo assim, o bom tratamento e escolha das melhores informações torna-se um fator determinante para a otimização do processo de transformação de informação em conhecimento útil. Dessa forma, cada vez mais problemas de classificação, e fortemente correlacionados a estes, os problemas de seleção de características, têm-se tornado bastante relevantes na área de

Mineração de Dados (*Data Mining*) e descoberta de conhecimento de qualidade.

1.1 Objetivos

Este trabalho tem como finalidade a realização de um estudo sobre os possíveis efeitos impactantes que o processo de seleção de características pode causar anteposto à tarefa de classificação, principalmente em problemas em que o modelo de dados inicial é demasiado complexo e de difícil compreensão.

Nesse sentido, propõe-se a implementação de um método de seleção de características, conhecido como *Nearest Shrunken Centroids*, que possibilite a distinção dos melhores atributos de um modelo de dados necessários ao processo de classificação. O desenvolvimento de um classificador que complemente o processo de seleção de características também se torna necessário, a fim de evidenciar a importância e as consequências que a etapa de seleção de características implica aos problemas de classificação e predição de dados.

Sendo assim, intenciona-se, principalmente, impactar positivamente os aspectos de desempenho e acurácia do processo de classificação de dados. Para tal, serão realizados testes comparativos utilizando-se alguns dos principais classificadores e diferentes bases de dados disponíveis atualmente.

1.2 Descoberta de Conhecimento em Bases de Dados

Em praticamente todas as áreas do conhecimento, dados são coletados e acumulados em um ritmo gradativamente mais acelerado. Torna-se mais evidente a necessidade de ferramentas e tecnologias que consigam realizar a transformação desse crescente fluxo de dados em informação útil (conhecimento). Tais tecnologias e ferramentas fazem parte da área denominada Descoberta de Conhecimento em Bases de Dados, *Knowledge Discovery in Databases - KDD* (Fayyad et al, 1996).

A tarefa de descoberta de conhecimento consiste dos seguintes passos (Han and Kamber, 2006):

- Pré-processamento: os dados são preparados para a tarefa de mineração. De acordo

com a necessidade, pode-se realizar:

  - Limpeza: remoção de inconsistências presentes no modelo de dados;
  - Integração: combinação de diferentes fontes de dados, normalmente representadas de formas distintas. São unidas pra formar uma única e maior base de dados;
  - Seleção: seleção das informações relevantes à análise pretendida;
  - Transformação: dados são transformados e consolidadados em formas mais apropriadas através de operações de agregação, normalização, sumarização, entre outras; a fim de prepará-los para a etapa de mineração.

- Mineração de Dados: utilização de métodos inteligentes de aprendizagem para extração de padrões de interesse, não evidentes, contidos no modelo de dados;
- Avaliação: identificação dos padrões que de fato refletem conhecimento a respeito da base de dados;
- Apresentação: visualização e representação do conhecimento extraído de forma compreensível ao usuário.

Conforme a Figura 1.1, a tarefa de *Data Mining* é definida apenas como uma das etapas do processo de descoberta de conhecimento, mas ainda assim, essencial, pois caracteriza-se por extrair o conhecimento implícito presente na base de dados.

1.3 Mineração de Dados

Mineração de Dados (*Data Mining, DM*) tem atraído grande atenção na indústria de informação e na sociedade em geral nos últimos anos, principalmente devido à disponibilidade de enormes quantidades de dados e a iminente necessidade de transformação desses dados em conhecimento potencialmente útil (Han and Kamber, 2006).

O processo de mineração de dados refere-se à análise de grandes bases de dados previamente existentes, a fim de que, através do comportamento de seus componentes, seja possível a extração de novos padrões implícitos no modelo de dados que, somados,

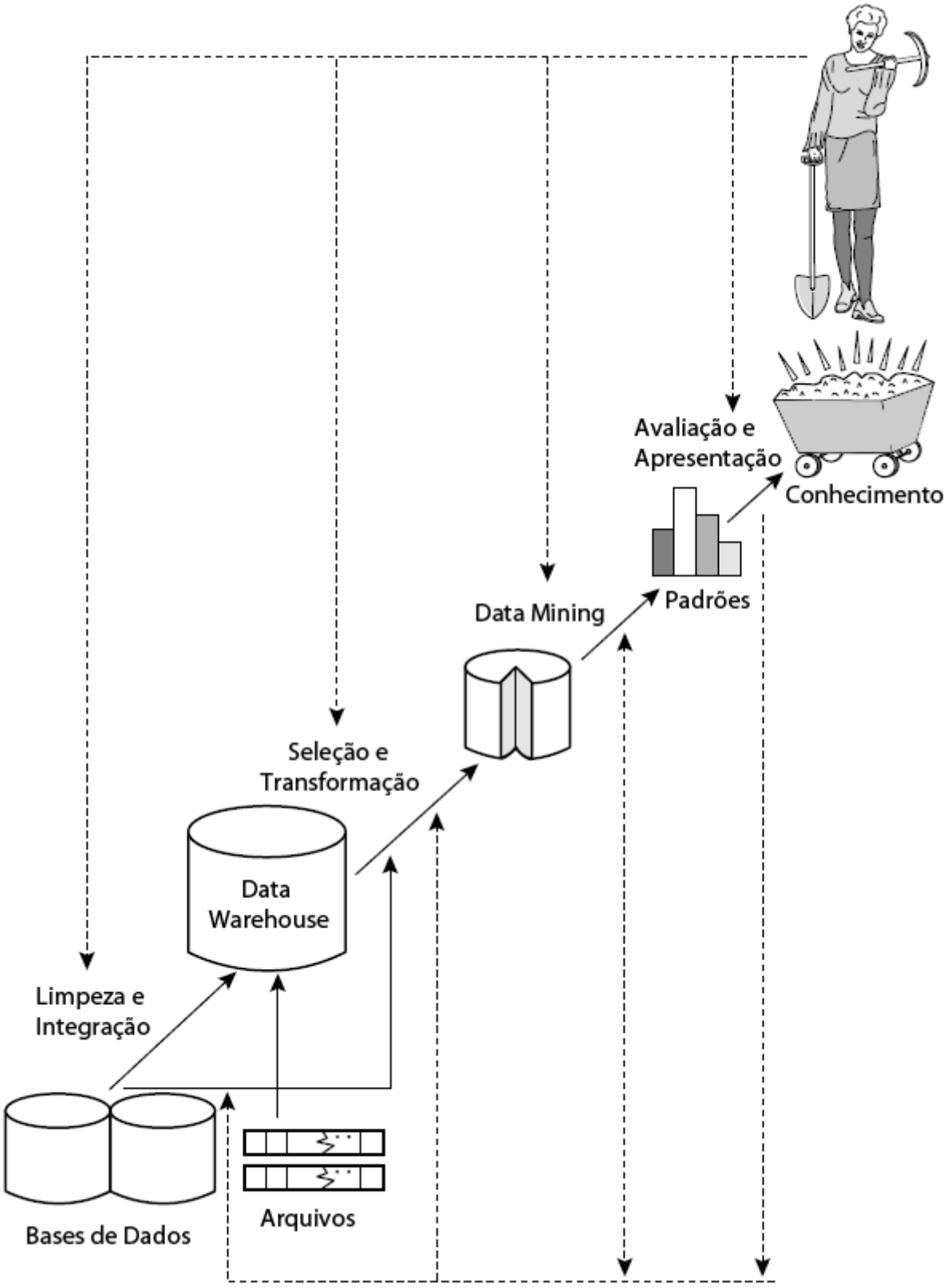

Figura 1.1: Etapas do KDD (Han and Kamber, 2006)

gerem uma gama de conhecimentos úteis capazes de aumentar a compreensão a respeito da própria base de dados.

A arquitetura típica de sistemas de mineração de dados consiste, principalmente, dos seguintes componentes (Han and Kamber, 2006):

- Conjunto de dados: base de dados ou conjunto de bases de dados disponíveis ao processo de mineração. Comumente apresentam grande volume de informações;
- Base de conhecimentos: refere-se ao domínio de conhecimento utilizado na avaliação do conhecimento extraído. Tais conhecimentos podem incluir conceitos de hierarquia de importância de atribtutos e crenças/opniões do usuário que possam guiar a busca por padrões de interesse;
- Métodos de mineração: consistem em um conjunto de diferentes métodos de mineração, como classificação, predição e regressão de dados, análise de correlação e associação entre atributos, sumarização e agregação de dados e análises de agrupamento de amostras;
- Avaliação de padrões: emprega medidas de interesse de forma a avaliar e guiar a busca por padrões descobertos pelos métodos de mineração.

Devido à grande gama de diferentes formas de conhecimento desejadas a respeito de bases de dados, a tarefa de mineração de dados pode empregar a integração de ténicas interdisciplinares, tais quais tecnologias de bases de dados, métodos estatísticos, computação de alto desempenho, métodos de visualização, redes neurais e sistemas inteligentes baseados em aprendizado de máquina, como descrito na Figura 1.2.

Uma característica dos sistemas inteligentes, dentre os quais os relacionados à mineração de dados, é a habilidade de se retro-alimentar através dos conhecimentos adquiridos a partir do modelo de dados. Tal processo, conhecido como ciclo virtuoso, é apresentado na Figura 1.3.

Considera-se como inteligente um sistema computacional que, através de habilidades especificas, é capaz de adquirir conhecimento a respeito da realidade do problema em que está inserido, a fim de solucioná-lo (Rezende et al, 2003).

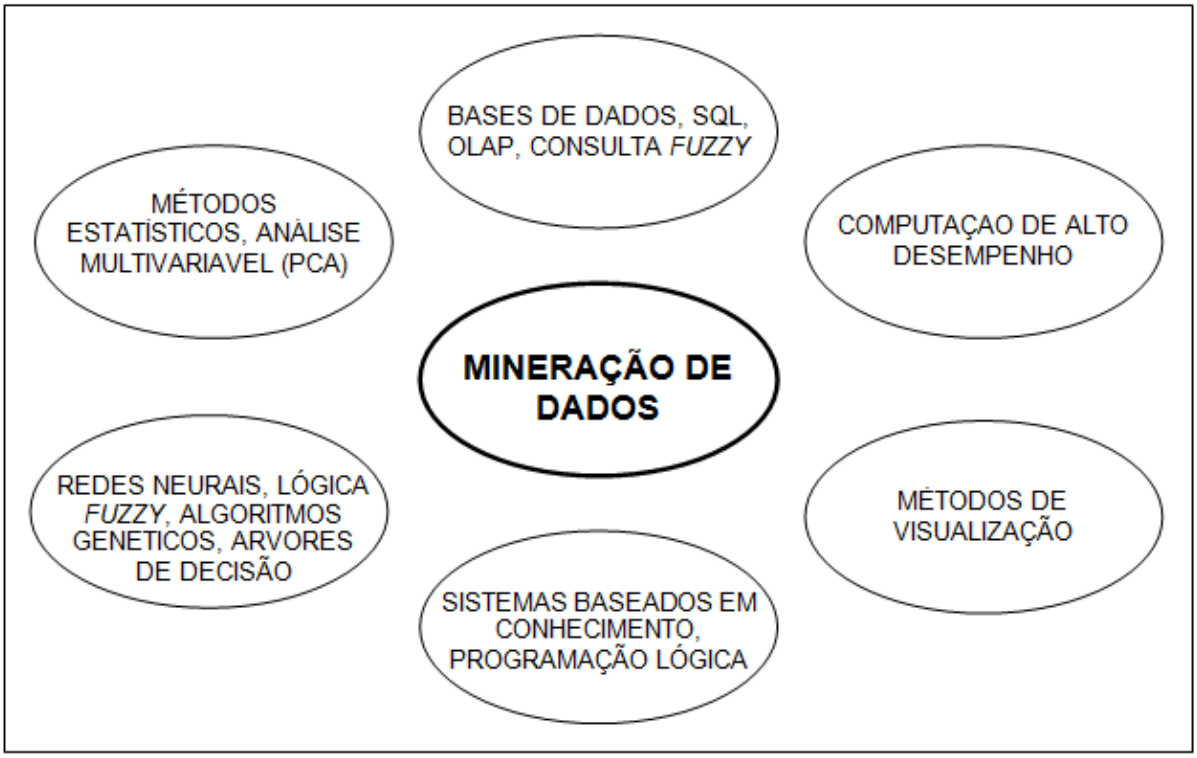

Figura 1.2: Mineração de Dados e suas Tecnologias (Han and Kamber, 2006)

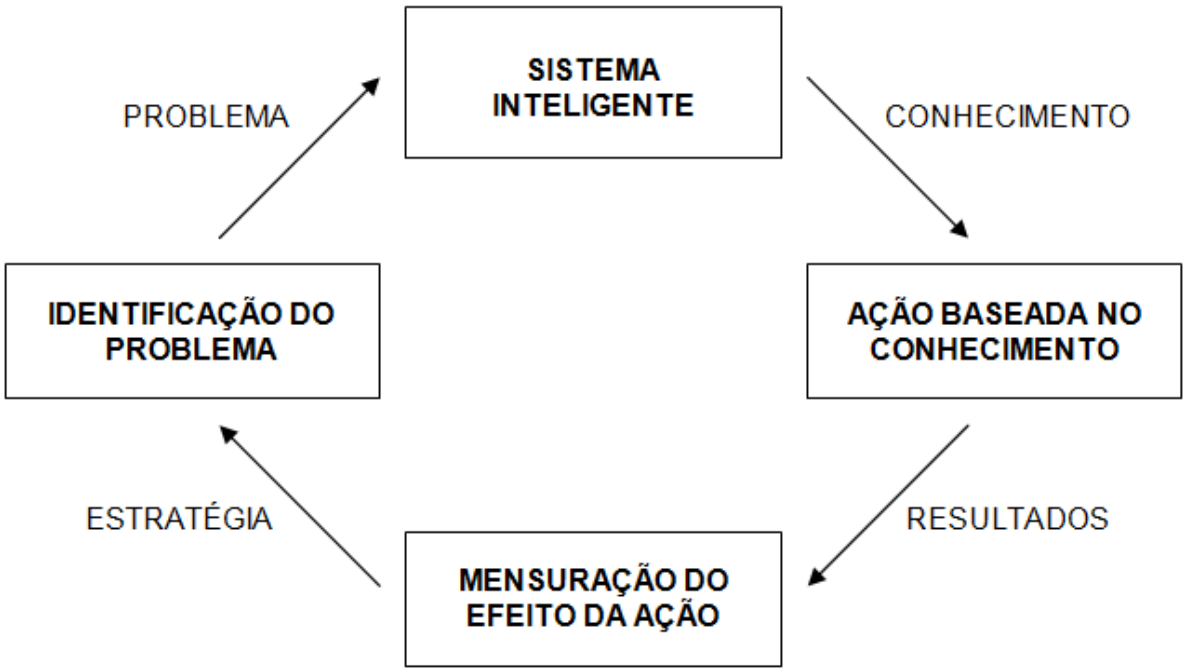

Figura 1.3: Ciclo Virtuoso e Sistema Inteligente (Silver, 1998)

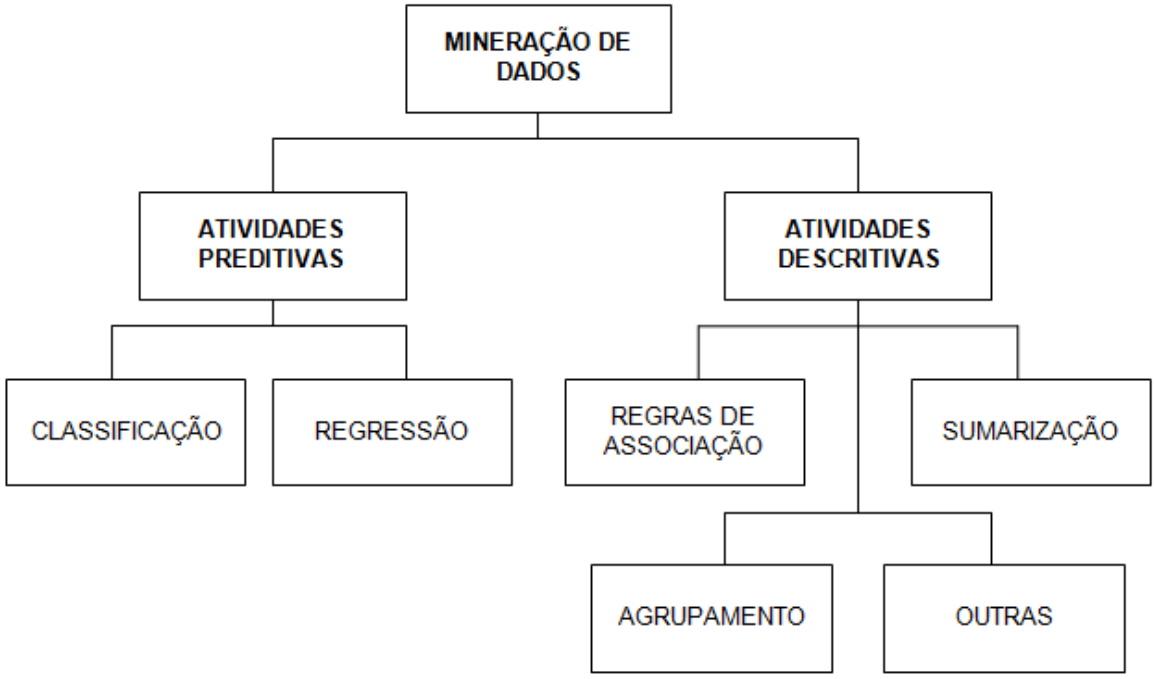

Figura 1.4: Classificação das Tarefas de Mineração de Dados (Rezende et al, 2003)

Do ponto de vista da área de Aprendizado de Máquina (*Machine Learning*), um sistema computacional inteligente, a partir da integração de novos conhecimentos, busca o auto-aprimoramento de sua capicidade de operação (Dietterich, 1990).

Dentro dessa perspectiva, as tarefas de mineração de dados podem ser classificadas em duas categorias: atividades preditivas e atividades descritivas (Figura 1.4). As atividades preditivas procuram descobrir a classe de uma amostra desconhecida a partir de um conhecimento previamente obtido de um conjunto de amostras já classificadas. As atividades descritivas, procuram por padrões comportamentais mais recorrentes, a partir de um conjunto de dados sem classes especificadas (Motta, 2004).

Dentro do escopo da área de *Data Mining*, destaca-se a tarefa de Classificação de Dados. Tal processo consiste em uma forma de análise de dados destinada a extração de padrões comportamentais que descrevem diferentes classes de dados.

2 Classificação de Dados

2.1 Processo de Classificação

A tarefa de classificação de dados consiste em um processo de duas etapas:

- Treinamento ou aprendizagem: extração ou aprendizado do conhecimento a partir de uma base de dados composta por um conjunto de amostras com classes conhecidas;
- Teste: avaliação do conhecimento descoberto no processo de treinamento através da predição da classe de amostras que ainda não foram submetidas ao classificador.

2.1.1 Treinamento

Etapa em que o conhecimento desejado é extraído a partir de um modelo de dados descrito por conjunto pré-determinado de classes. Realiza-se a construção de um classificador a partir das informações contidas em um conjunto inicial de dados, conhecido como conjunto de treinamento ou amostras de treinamento.

Cada amostra de treinamento é definida por um conjunto de atributos com seus valores associados e por um atributo de classe, que identifica à qual categoria a amostra pertence (Santoro, 2005).

Matematicamente, uma amostra X é representada por um conjunto p-dimensional de atributos, $X = (x_1, x_1, \ldots, x_p)$, e por um atributo-rótulo x_k, de ordem discreta, que define uma das K classes pré-definidas pelo modelo de dados, $C_k = (k_1, k_2, \ldots, k_K)$.

A partir do conjunto de treinamento, o conhecimento é adquirido através da definição de uma função de mapeamento, $y = f(X)$, que se responsabiliza por predizer a classe y associada à amostra X.

A partir do conjunto de treinamento, o algoritmo de classificação se encarrega da construção de um modelo de conhecimento que seja capaz de realizar a previsão de classes de futuras amostras desconhecidas (Figura 2.1).

Pelo fato de cada amostra de treinamento estar rotulada com sua classe associada,

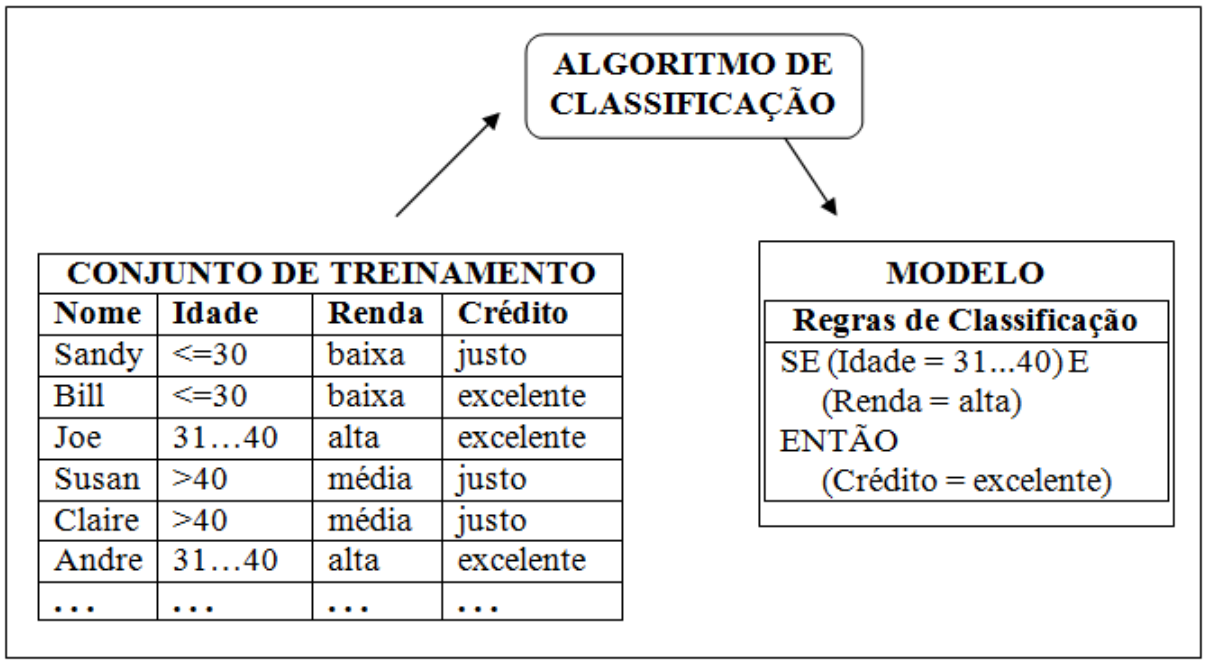

Figura 2.1: Processo de Treinamento (Motta, 2004)

o processo de treinamento é definido como Apredizado Supervisionado. Dessa forma, a tarefa de classificação se distingue de outras áreas da mineração de dados, conhecidas como Aprendizado Não-Supervisido, em que a classificação das amostras de treinamento é desconhecida (Han and Kamber, 2006).

2.1.2 Teste

Neste passo é realizada a avaliação do conhecimento adquirido pela etapa anterior. Realiza-se o processo de classificação de um novo conjunto de amostras independentes do processo de treinamento. Tais amostras, também já previamente classificadas, são denominadas amostras de teste ou conjunto de teste. Após a classificação, é feita a comparação entre as classes preditas pelo classificador com os valores esperados conhecidos.

A Figura 2.2 exemplifica o processo de teste de classificação. A partir de um conjunto de teste, o classificador realiza a predição das classes das novas amostras através do modelo de conhecimento extraído pelo processo de treinamento.

Através do percentual de acertos da classificação realizada nesta etapa, estima-se a acurácia do modelo, definida como a capacidade do classificador em predizer corretamente as classes de amostras do conjunto de teste. Se a acurácia encontrada for considerada aceitável, o classificador pode ser utilizado futuramente na classificação de amostras

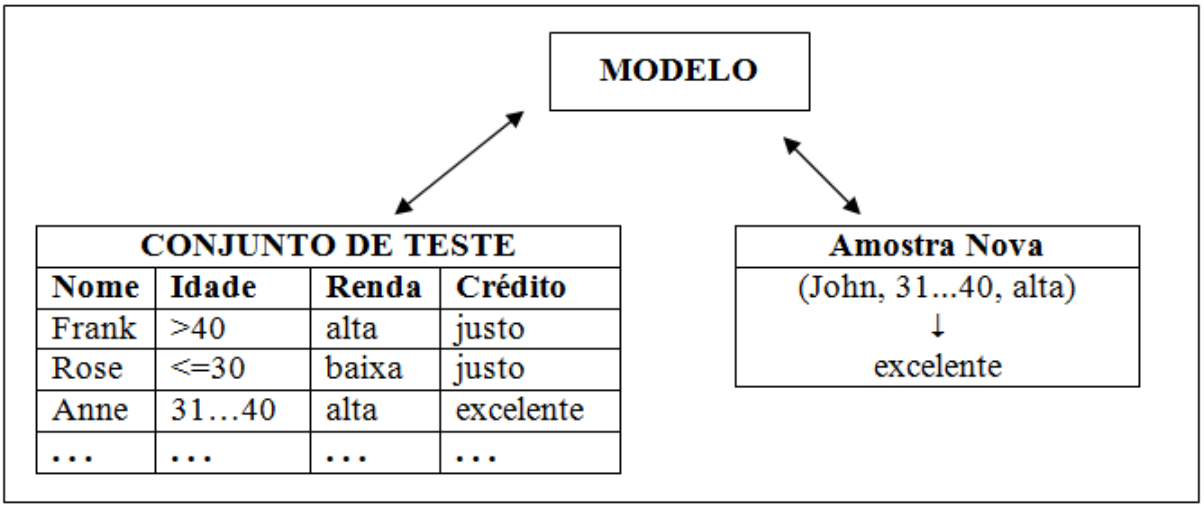

Figura 2.2: Teste de Classificação (Motta, 2004)

desconhecidas.

2.2 Avaliação dos Métodos de Classificação

Os métodos de classificação podem ser avaliados pelos seguintes critérios (Han and Kamber, 2006):

- Acurácia: é a capacidade de predizer com exatidão a classe de amostras desconhecidas ao processo treinamento;
- Desempenho: refere-se a velocidade e custo computacional relacionados à construção e utlização do classificador;
- Robustez: capacidade de realizar predições corretas a partir de conjuntos de dados com amostras incompletas ou com ruído;
- Escalabilidade: é a habilidade de construção de um modelo de conhecimento eficiente a partir de uma grande quantidade de dados;
- Interpretabilidade: refere-se a facilidade de compreensão do modelo de conhecimento extraído do modelo de dados.

Na perpectiva deste trabalho, como melhor exemplificado a se seguir, será dada atenção especial aos aspectos de desempenho e acurácia do processo de classificação.

Sendo assim, serão abordadas algumas técnicas mais complexas de avaliação da acurácia de classificadores.

2.2.1 Métodos de Avaliação de Acurácia

A acurácia de um determinado classificador é estimada a partir da exatidão com que o mesmo classifica um conjunto de amostras distintas das utilizadas no processo de treinamento.

Dessa forma, tornou-se comum, em sistemas inteligentes, a divisão da base de dados inicial em conjuntos de treinamento e teste, a fim de que, no final do processo de aprendizagem, seja realizada a avaliação do conhecimento adquirido pelo sistema de aprendizagem (Santoro, 2005).

Nesse sentido, algumas técnicas de avaliação de acurácia podem diferir em níveis de complexidade ou metodologia que empregam na divisão do conjunto de dados original. Algumas das principais técnicas utlizadas são: método *holdout*, validação cruzada e método *leave-one-out*.

Método *Holdout*

A técnica mais simples de validação é conhecida como método (ou validação) *holdout*, ou teste de estimativa simples. Consiste na divisão do conjunto original de dados em dois subconjuntos disjuntos: treinamento e teste. Usualmente utiliza-se a proporção de 2/3 a 4/5 do volume inicial de dados para o conjunto de treinamento e o restante das amostras forma o conjunto de teste (Reich and Barai, 1999).

Este processo é exemplificado pela Figura 2.3: o conjunto de treinamento é utlizado na indução de conhecimento, enquanto o conjunto de teste para a avaliação do conhecimento e determinação da acurácia do classificador construído.

Este método isoladamente é pouco utilizado, pois pode definir conjuntos de treinamento e teste altamente variantes e, em alguns casos, não significativos ao contexto do conhecimento intencionado. Esta abordagem é aconselhável quando uma grande quantidade de dados está disponível para a classificação, pois caso o conjunto total de dados seja pequeno, o conhecimento extraído pode não ser confiável (Santoro, 2005).

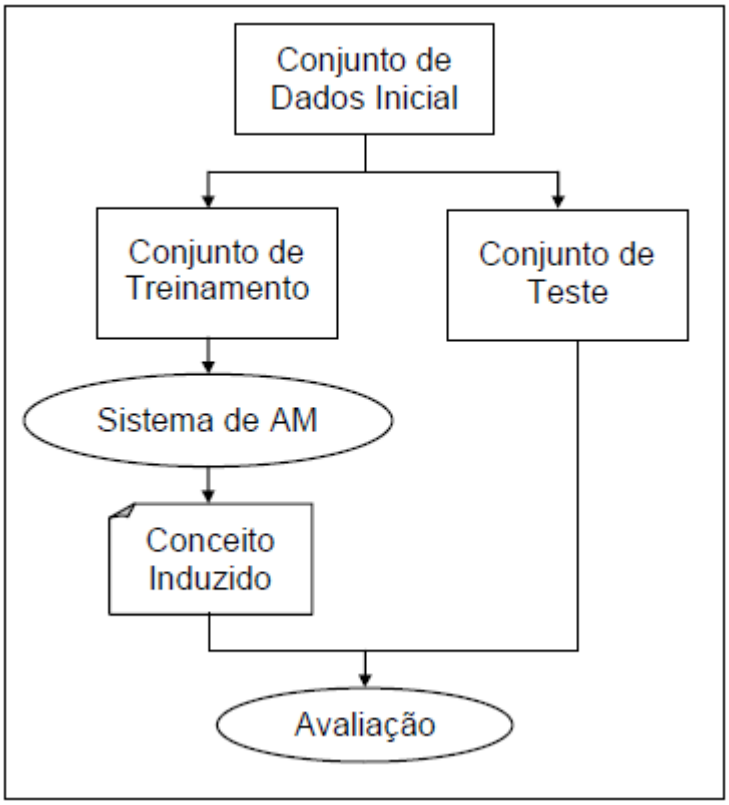

Figura 2.3: Avaliação de Conhecimento adquirido por um Sistema de Aprendizado de Máquina (AM) (Santoro, 2005)

Validação Cruzada

Devido às limitações da validação *holdout*, uma alternativa muito utilizada é conhecida como validação cruzada, ou *cross-validation*. Consiste na divisão da base de dados original em n conjuntos mutuamente exclusivos e de mesmo tamanho aproximado. Os processos de treinamento e teste são realizados n vezes de forma que, a cada iteração, um dos conjuntos se torna o conjunto de teste e os restantes formam o conjunto de treinamento. A acurácia é calculada a partir da média aritmética dos resultados obtidos a cada iteração da validação cruzada.

A Figura 2.4 exemplifica uma aplicação de validação cruzada de 10 folhas num conjunto incial de dados de 1000 amostras. São definidos 10 subconjuntos de 100 amostras, de forma a serem realizados 10 experimentos independentes. Cada experimento consiste na utilização de 9 subconjuntos no processo de treinamento e o subconjunto restante como conjunto de teste.

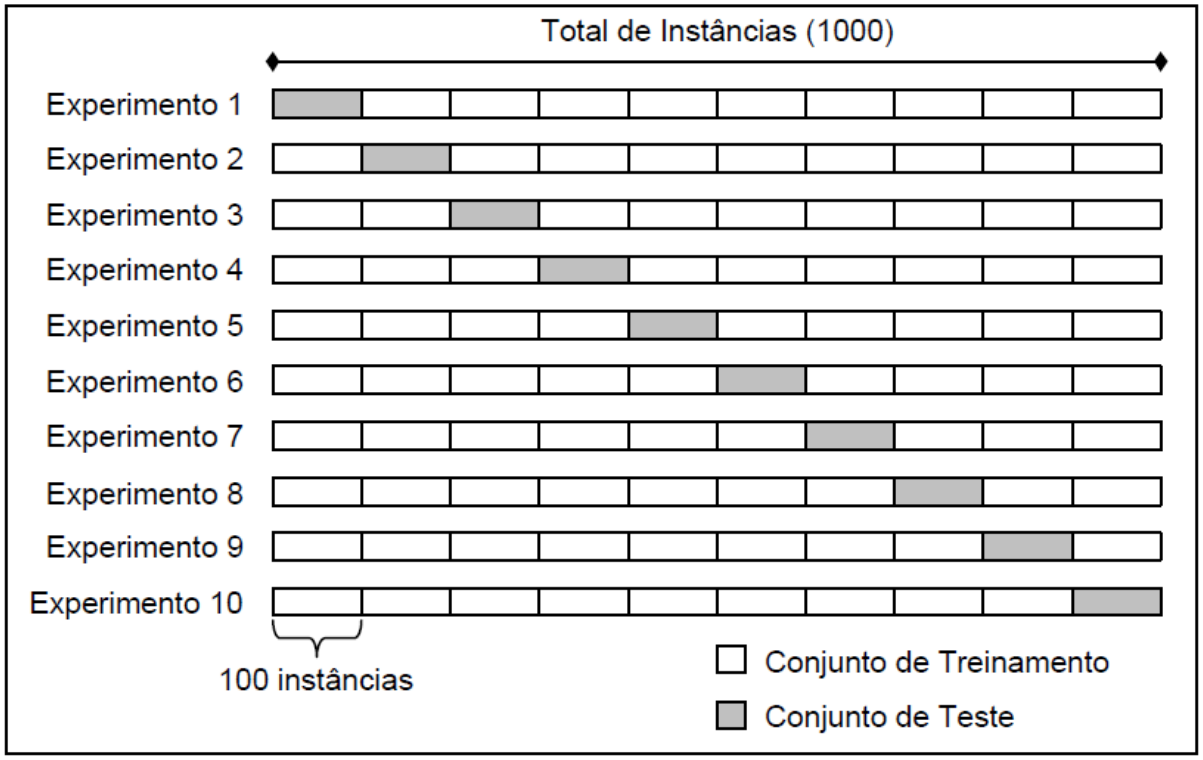

Figura 2.4: Validação cruzada de 10 folhas ou *10-fold cross-validation* (Santoro, 2005)

Método *leave-one-out*

Um caso especial de validação cruzada é conhecido como método *leave-one-out*. Em uma situação em que a base de dados inicial é composta por n amostras, é realizada a validação cruzada de n folhas, de forma que os conjuntos de teste, definidos a cada iteração, sejam compostos por apenas uma instância.

Esta abordagem garante uma estimativa de erro mais aproximada à realidade, mas possui a desvantagem de ser proibitiva devido ao seu alto custo computacional. Em geral, a validação cruzada de 10 folhas é suficiente para a estimativa da acurácia em sistemas de classificação (Han and Kamber, 2006).

2.3 Método de *Nearest Centroid*

Considerando que os principais objetivos do presente trabalho são as avaliações dos processos de classificação e de seleção de características, foram adotados os seguintes pontos de referência:

- Avaliação do processo de classificação: optou-se pela utilização da validação cruzada

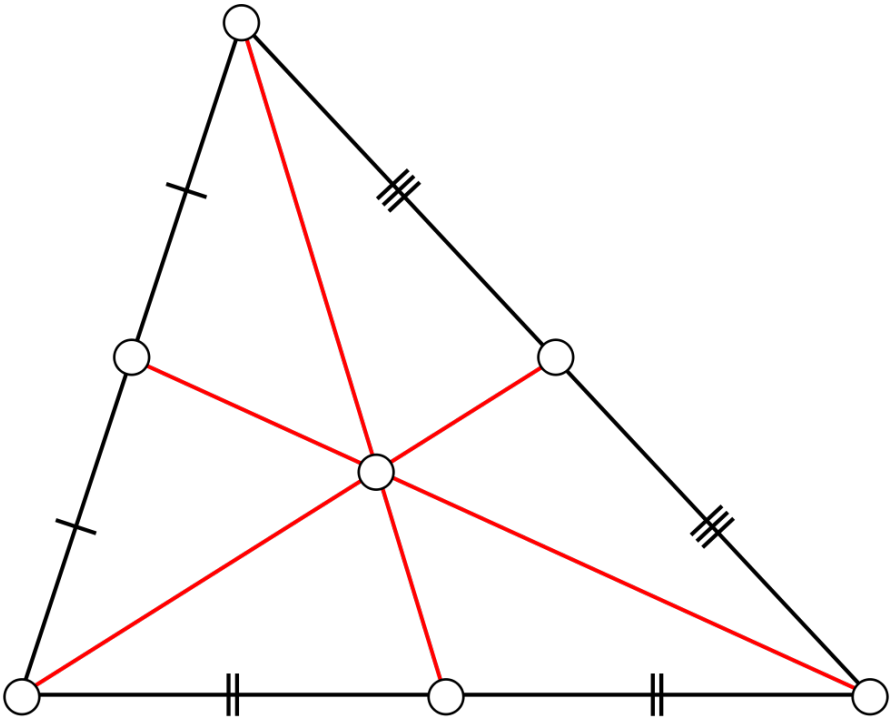

Figura 2.5: Centróide geométrico de um triângulo

de 10 folhas para a estimativa de acurácia;

- Avaliação do processo de seleção de características: construção e utilização de um classificador conhecido como *Nearest Centroid* a fim de direcionar o processo de redução do número de atributos.

2.3.1 Centróides

Geometricamente, centróide é definido como o ponto interno de uma forma geométrica que define seu centro geométrico. Caso a forma geométrica represente uma seção homogênea, então o centróide também define seu centro de massa. Na Figura 2.5 é ilustrado o centróide geométrico de um triângulo.

Em classificação de dados, centróide é definido como o protótipo ou centro de distribuição de um determinado conjunto de amostras. Tal estrutura sintetiza, simplificadamente, a forma como o conjunto de dados é representado pela distribuição amostral.

Geralmente, os centóides são definidos a partir dos seguintes conjuntos de dados:

- Conjunto de treinamento: a partir de todas as amostras de treinamento, com suas diferentes classes associadas, define-se o **centróide global** que representa a dis-

tribuição amostral média do conjunto de treinamento;

- Conjunto de amostras de classe: por meio de um conjunto de amostras de uma classe específica, determina-se o **cetróide de classe** que pode ser utilizado na distinção entre os diferentes conjuntos de classes presentes no modelo de dados.

Dessa forma, pode-se definir centróides como estruturas de dados, de mesma dimensão das amostras de treinamento, em que cada componente é obtido através da média do i-ésimo atributo correspondente, presente nas amostras de dados (Tibshirani et al, 2002a).

Matematicamente: seja um espaço p-dimensional, sendo p o número de atributos $i = 1, 2, \ldots, p$, presentes num conjunto de dados composto de n amostras $j = 1, 2, \ldots, n$. Define-se x_{ij} como a expressão do i-ésimo atributo da amostra j. Cada amostra está associada a uma classe k, pertence a um conjunto discreto de K classes $C_k = (1, 2, \ldots, K)$. A cada classe k, estão associadas n_k amostras que compõem o modelo de dados.

Através do conjunto total de treinamento define-se o cetróide global, representado por $\overline{x}$. Dessa forma, o i-ésimo componente do centróide global é expressado por:

$$\overline{x}_i = \sum_{j=1}^{n} x_{ij}/n \tag{2.1}$$

De forma análoga, a partir do conjunto de amostras de determinada classe k, determina-se o i-ésimo componente do centróide de classe, $\overline{x}_{ik}$:

$$\overline{x}_{ik} = \sum_{i \in C_k} x_{ij}/n_k \tag{2.2}$$

A partir do momento em que a base de dados de treinamento é analisada, são criados e calculados $K+1$ centróides que correspondem a cada uma das K classes presentes no modelo de dados e ao centróide global.

2.3.2 *Nearest Centroid*

O método de classificação *Nearest Centroid* se assemelha ao método do vizinho mais próximo (ou *k-nearest neighbor, kNN*). A classificação é feita de acordo com o centróide

mais próximo à amostra a ser classificada, x^*, a partir de uma função discriminante de distância, $\delta(x^*)$, definida como (Tibshirani et al, 2002a):

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \overline{x}_{ik})^2}{s_i^2} - 2\log \pi_k \tag{2.3}$$

Sendo s_i definido como o desvio padrão do atributo i, isto é:

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{i \in C_k} (x_{ij} - x_{ik})^2 \tag{2.4}$$

O primeiro fator da função discriminante 2.3 é a distância quadrada, de x^* ao k-ésimo centróide, normalizada pela variância, s_i^2. O segundo fator é uma correção baseada na distribuição de probabilidades das classes contidas no modelo de dados, sendo π_k a probabilidade de uma dada amostra ser da classe k, como a seguir:

$$\pi_k = n_k/n \tag{2.5}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.6}$$

Tal correção objetiva a eliminação de ambiguidades decorrentes de situações em que duas ou mais distâncias muito próximas são encontradas, de forma que a classe com o maior número de ocorrências seja beneficiada.

Dessa forma, pode-se definir a classe da amostra x^* como a seguir:

$$C(x^*) = l \text{ se } \delta_l(x^*) = \min_k \delta_k(x^*) \tag{2.7}$$

A classe escolhida é a associada ao centróide que, através da função $\delta_k(x^*)$, minimiza a distância à amostra a ser classificada.

3 Seleção de Características

Um fator decisivo ao processo de descoberta de conhecimento é relativo à escolha das informações disponíveis que participarão do processo de mineração de dados.

Teoricamente, um número maior de atributos deveria resultar em um maior poder de discernimento. Mas, na prática, a adição de informações irrelevantes a um conjunto de dados geralmente confunde os sistemas de aprendizado. Experimentos realizados mostram que a adição randômica de atributos a bases de dados conhecidas causam deterioração do desempenho e da corretude da análise dos dados (Witten et al, 2011).

Sendo assim, a questão da quantidade de dados a ser utilizada no processo de descoberta de conhecimento torna-se, primordialmente, um aspecto de dimensionalidade do modelo de dados. Muitos dos algoritmos de aprendizado de máquina são projetados para realizar a descoberta dos atributos mais apropriados para embasarem suas decisões, de forma a maximizar a habilidade de generalização do conhecimento extraído (Santoro, 2005).

Atualmente muitas bases de dados contém centenas ou milhares de atributos, muitos dos quais irrelevantes ou redundantes à tarefa de mineração. Em uma tarefa de classificar quão eficiente é uma oferta de um novo CD enviada a um determinado cliente, informações como o nome são menos significativas que a idade ou gosto musical. Mesmo que seja possível para um especialista a escolha de atributos úteis, essa tarefa pode ser difícil e demorada. Esse fato é agravado quando a base de dados é desconhecida, motivo este do processo de análise de dados (Han and Kamber, 2006).

Surge então a necessidade de um processo de eliminação de atributos ou características irrelevantes, redundantes ou até mesmo prejudiciais; em busca da melhoria do processo de descoberta de conhecimento e de uma maior compreensão a respeito do próprio modelo de dados como fonte de conhecimento. Este processo é conhecido como Seleção de Características (*Feature Selection*).

Sendo assim, seleção de características é definida como o processo de redução de dimensionalidade de uma base de dados devido à remoção de atributos indesejados.

Tal processo objetiva encontrar o menor subconjunto de atributos cuja distribuição de probabilidades de classes é tão próxima quanto possível à da base de dados original. Dessa forma, posteriores processos de análise de dados são beneficiados em fatores mensuráveis, como o desempenho, e fatores subjetivos, como a facilidade de compreensão e extração de informação útil da base de dados (Han and Kamber, 2006).

O principal estímulo à realização da seleção de características é a existência de bases de dados com alto nível dimensional que acarretam alto custo computacional (baixo desempenho) no processo de mineração de dados. Tendo em vista este fato e que, em uma base de dados com n atributos, existem 2^n possíveis subconjuntos de atributos, uma busca exaustiva pela solução ótima se torna demasiadamente custosa.

Alguns métodos se utilizam de estratégias heurísticas para alcançar uma boa seleção de características. Dentre estes, podemos citar:

- *Stepwise forward selection*: parte-se de um subconjunto vazio de atributos e a cada passo o melhor atributo, determinado por uma função heurística, é adicionado ao conjunto de atributos;
- *Stepwise backward selection*: parte-se do conjunto original de atributos e a cada passo o pior atributo é eliminado do conjunto de atributos;
- Árvores de decisão: a partir da árvore de decisão construída de acordo com o modelo de dados, os atributos ausentes à estrutura são eliminados, pois são irrelevantes.

3.1 Método *Nearest Shrunken Centroids*

Um método simples proposto faz uso de *Shrunken Centroids*, ou centróides reduzidos, como protótipos de classes que identificam os subconjuntos de atributos presentes no modelo de dados que melhor caracterizam cada classe (Tibshirani et al, 2002b). O método é geral, de fácil entendimento e interpretação, e pode ser utilizado em problemas de classificação com altos níveis dimensionais (Tibshirani et al, 2002a).

Esse método foi inicialmente desenvolvido para ser empregado em problemas envolvidos com micro-matrizes genéticas. Micro-matrizes genéticas são representações de

conjuntos de genes que compõem uma parte do DNA de células, humanas ou não. Devido à alta complexidade inerente ao modelo genético, muitas das vezes, tais representações genéticas possuem um elevadíssimo nível dimensional expressado por um número cada vez maior de atributos.

Dessa forma, tornou-se evidente a necessidade da redução dimensional desses modelos a um patamar em que pudessem ser realizados estudos analíticos a um custo computacional acessível.

Considerando as definições de centróide global e centróides de classes, definidos nas Equações 2.1 e 2.2, respectivamente, e a definição de desvio padrão em 2.4, Tibshirani et al (2002b) definem uma função estatística de distância d_{ik} como:

$$d_{ik} = \frac{\overline{x}_{ik} - \overline{x}_i}{m_k.s_i} \tag{3.1}$$

Sendo m_k:

$$m_k = \sqrt{1/n_k - 1/n} \tag{3.2}$$

Dessa forma, o denominador, $m_k.s_i$, torna-se igual ao erro padrão do numerador, $\overline{x}_{ik} - \overline{x}_i$. Funções como a descrita são parte de um grupo denominado *t statistic* e normalmente são empregadas em diversos métodos, computacionais e estatísticos, de aprendizado e de tomada de decisões.

A função d_{ik} denota a comparação entre a classe k e a distribuição amostral da base de dados. Em outras palavras, é uma medida de distância entre o centróide k e o centróide global. Dessa forma podemos transcrever a Equação 3.1 da seguinte forma, para um melhor entendimento:

$$\overline{x}_{ik} = \overline{x}_i + m_k.s_i.d_{ik} \tag{3.3}$$

Sendo assim, os centróides de cada classe podem ser expressos a partir do centróide global e de uma medida de distância definida por $m_k s_i d_{ik}$.

A partir desse princípio, e levando em conta a seleção de características como objetivo, Tibshirani et al (2002b) propõem o encolhimento (*shrinkage*) dos centróides

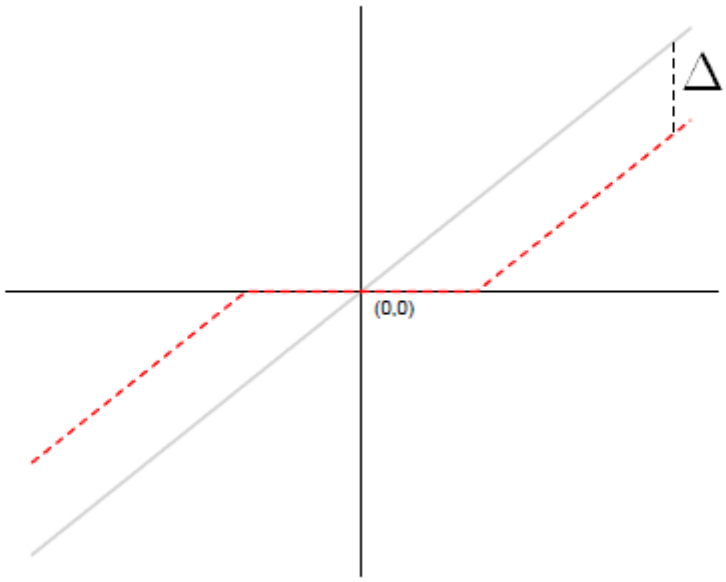

Figura 3.1: Função de limiarização suave ou *soft-thresholding* (Tibshirani et al, 2002b)

em direção ao centróide global. Cada d_{ik} é reduzido em valor absoluto por um valor Δ. Caso esse resultado seja negativo, então d_{ik} é transformado em zero. Tal transformação é conhecida como limiarização suave ou *soft-thresholding* e pode ser expressa matematicamente por:

$$d'_{ik} = \text{sinal}(d_{ik})(|d_{ik}| - \Delta)_+ \tag{3.4}$$

Onde o símbolo + significa a parte positiva, como em $t_+ = t$ se $t > 0$, senão $t_+ = 0$.

Sendo assim, a partir do encolhimento de d_{ik}, agora transformado em d'_{ik} pela função de limiarização, podemos reescrever 3.3 como a seguir:

$$\overline{x}'_{ik} = \overline{x}_i + m_k.s_i.d'_{ik} \tag{3.5}$$

Se for possível escolher um valor de Δ suficientemente grande para que o valor de d'_{ik} seja igual a zero, considerando um determinado atributo i e todas as classes k existentes, chega-se a uma situação em que todos os i-ésimos componentes dos novos centróides, $\overline{x}'_{ik}$, sejam iguais ao do centróide global $\overline{x}_i$. Dessa forma esse atributo i deixa

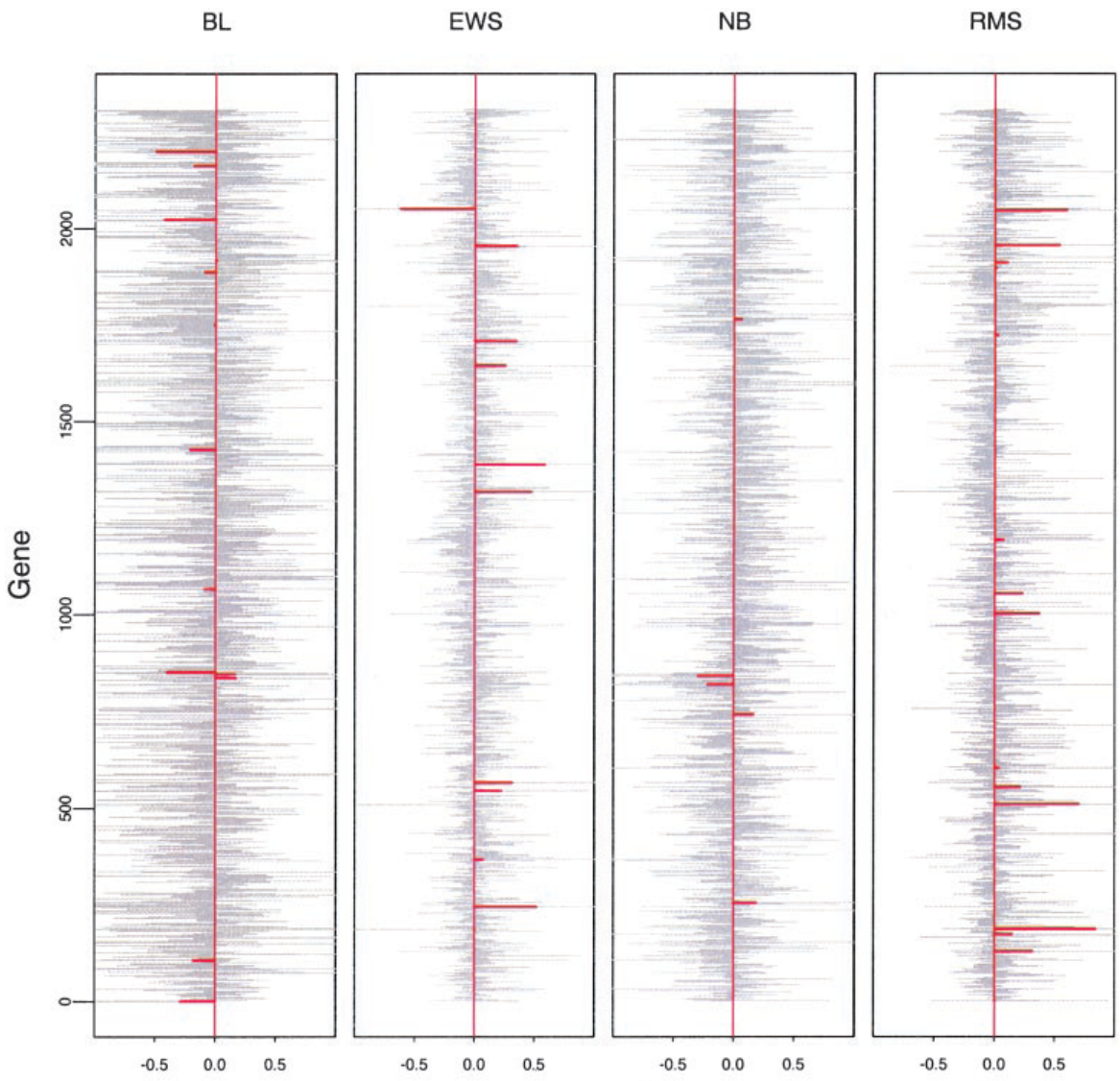

Figura 3.2: Centróides e *Shrunken Centroids* (Tibshirani et al, 2002a)

de contribuir para a classificação e pode, então, ser descartado. Ocorre então os chamados *shrunken centroids*, ou centróides encolhidos, que possuem um menor número de atributos que os centróides originais.

Matematicamente: se dado i, $\exists\Delta$ tal que $\forall k$ tem-se $d'_{ik} = 0$, então o i-ésimo componente dos centróides pode ser suprimido, gerando assim novos e menores *shrunken centroids*.

Na Figura 3.2 estão representados centróides (em cinza) e seus respectivos *shrunken centroids* (em vermelho) da base de dados SRBCT, que representa um conjunto de células cancerígenas de diferentes tipos. Cada tipo de câncer (BL, EWS, NB, RMS) é discriminado por um centróide específico e a partir de seu *shrunken centroid*. Nota-se que diferentes conjuntos de genes discriminam cada uma das classes.

Pode-se escolher o valor de Δ a partir da análise de erros de classificação através de validação cruzada, de forma que Δ não seja suficientemente grande para que o processo de classificação não sofra deterioração apresentando um maior percentual de erros que o encontrado a partir da base de dados original. Esse fato se deve à eliminação de uma quantidade maior de características que a suportada pelo modelo de dados.

4 Estudo de Caso

Um dos objetivos deste trabalho é a análise dos efeitos impactantes que o processo de seleção de características pode causar à tarefa de classificação. Dessa forma, através do método *Nearest Shrunken Centroids*, foi implentado um sistema inteligente capaz de realizar, a partir de uma base de dados, ou conjunto de bases de dados, seleção de características e classificação.

A seguir, são descritos a arquitetura e o funcionamento do sistema inteligente implementado, a metodologia utlizada para a realização dos testes comparativos e os resultados obtidos pelos testes, utilizando diferentes bases de dados e classificadores disponíveis.

4.1 Sistema Inteligente

A estrutura do sistema inteligente é descrita na Figura 4.1. Cada um dos componentes apresentados pelo diagrama de classes é descrito detalhadamente a seguir:

- ***File***: representa o arquivo físico da base de dados utlizada no processo de classificação e seleção de características;
- ***Sample***: representação de cada uma das amostras contidas na base de dados. Cada amostra é composta de uma lista de atribtuos (***values***) e rotulada por sua classe específica (***classe***);
- ***DataBase***: através do arquivo físico (***File***) é construída a representação de toda a base de dados. Cada base de dados é identificada por um nome (***name***), uma lista de atributos (***attributes***), sendo um deles o identificador de classe (***classAttribute***); e composta de um conjunto de amostras (***samples***);
- ***Centroid***: a partir de um determinado conjunto de amostras, é calculado um centróide determinador do centro de distribuição dos atributos. Analogamente às amostras (***Sample***), os centróides são representados por uma lista de atributos (***values***) e pela classe associada (***classe***);

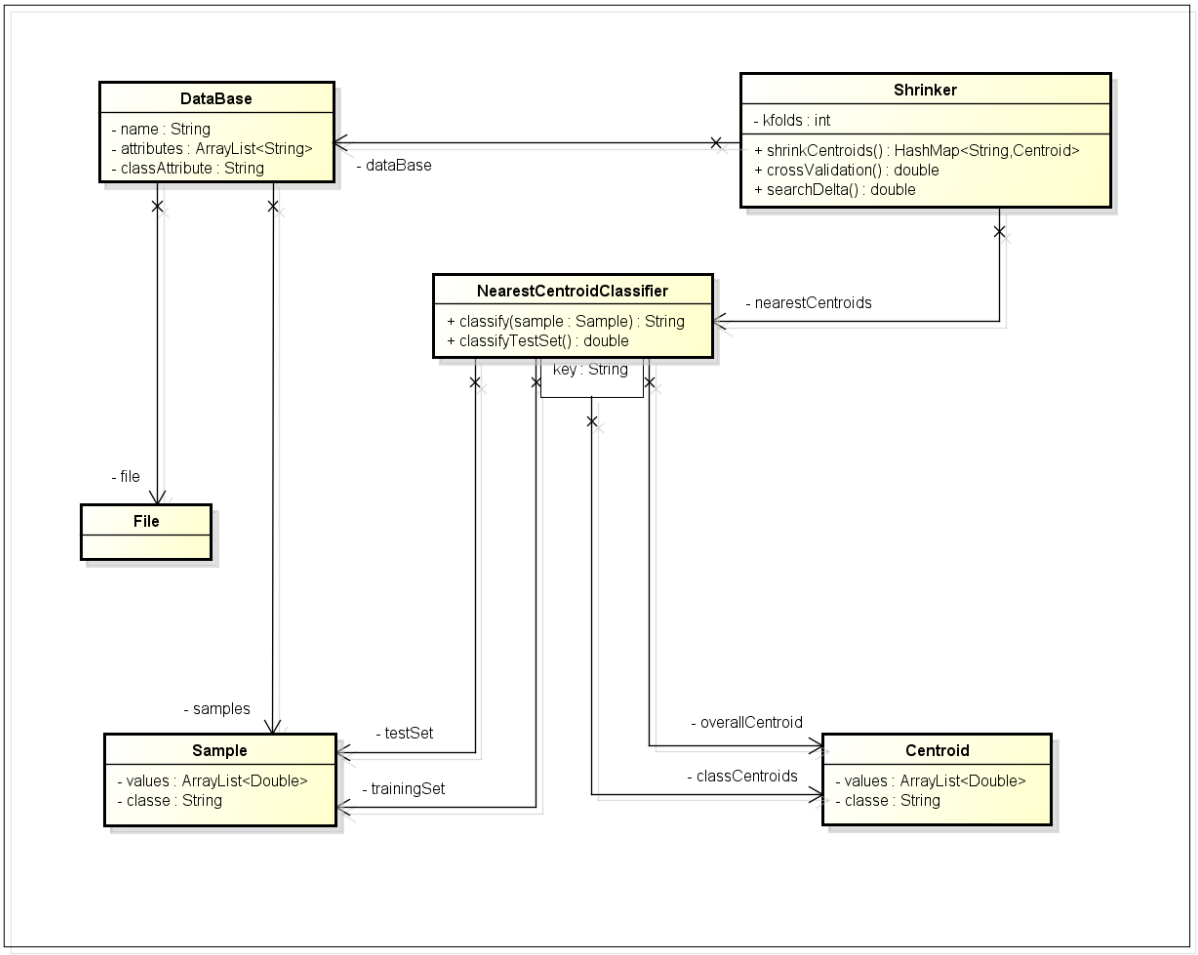

Figura 4.1: Diagrama de Classes

- ***NearestCentroidClassifier***: classificador que implementa o método *Nearest Centroid*. É composto de um conjunto de treinamento (***trainingSet***) e um conjunto de teste (***testSet***). Através do conjuto de classes contido no conjunto de treinamento, o classificador constrói um conjunto de centróides de classe (***classCentroids***) e um centróide geral (***overallCentroid***). A partir do conjunto de teste, novas amostras desonhecidas são classificadas através da minimização da função discriminante ***classify(sample)*** (Equação 2.7);

- ***Shrinker***: executa o processo de seleção de características da base de dados. Realiza o encolhimento dos centróides (***shrinkCentroids()***) de um classificador (***NearestCentroidClassifier***) a partir de um parâmetro Δ, como definido pelas Equações 3.4 e 3.5. Este parâmetro é encontrado pelo método ***searchDelta()*** através da execução de validação cruzada (***crossValidation()***) de ***kfolds*** folhas sobre o conjunto de dados do classificador

NearestCentroidClassifier.

4.2 Metodologia

A metodologia utilizada no proposto estudo de caso é apresentada abaixo a fim de descrver a forma de execução dos testes comparativos realizados.

4.2.1 Sub-divisão das bases de dados

Através da sub-divisão das bases de dados originais em conjunto de treinamento e teste e, posteriormente, da seleção de características, são disponibilizados quatro diferentes conjuntos de dados:

- Dados originais: base de dados original, composta de todas as características iniciais do modelo de dados. Os dados são divididos em:
  - Conjunto de Treinamento: conjunto utilizado no processo de apredizado do modelo de dados original;
  - Conjunto de Teste: conjunto de dados utilizado nos teste de classificação do modelo inicial.
- Dados reduzidos: composto dos mesmos conjuntos de amostras de treinamento e teste originais, mas suprimidos dos atributos eliminados através do processo de seleção de características.

Para a realização deste estudo, o parâmetro utilizado para a divisão da base de dados foi de 75% dos dados originais como conjunto de treinamento e 25% para teste.

4.2.2 Bases de Dados Utilizadas

Ao todo, foram selecionadas sete diferentes bases de dados para a realização do presente estudo de caso. Os conjuntos de dados diferem entre si no número de amostras, número de atributos, quantidade de classes por modelo de dados e na forma como os atributos são representados, discreta ou continuamente.

A seguir é apresentada uma análise estrutural e descritiva de cada base de dados utlizada.

Iris

- **Descrição**: Conjunto de dados de amostras de 3 diferentes tipos da flor íris;
- **Número de amostras**: 150;
- **Atributos**: 4 atributos contínuos que representam a largura e comprimento da sépala e pétala;
- **Classes e Distribuição**: as amostras estão distribuídas em 3 variedades da flor íris:
  - Iris Setosa (50 amostras);
  - Iris Versicolour (50 amostras);
  - Iris Virginica (50 amostras).

Glasses

- **Descrição**: Base de dados de identificação de diferentes tipos de vidro. O estudo da classificação de tipos de vidro é motivado por investigações criminais. Os restos de vidro deixados em cenas de crime podem ser usados como evidência caso identificados corretamente;
- **Número de amostras**: 214;
- **Atributos**: 9 atributos contínuos que representam o índice de refração e concentrações de componentes químicos;
- **Classes e Distribuição**: os 7 diferentes tipos de vidro são:
  - vidro plano de janelas de edifícios (70 amostras);
  - vidro não-plano de janelas de edifícios (76 amostras);
  - vidro plano de veículos (17 amostras);

  - vidro não-plano de veículos (0 amostras);
  - recipientes de vidro (13 amostras);
  - louça (9 amostras);
  - faróis (29 amostras).

Breast

- **Descrição**: Conjunto de dados de diagnóstico de câncer de mama em células;
- **Número de amostras**: 24;
- **Atributos**: 12625 atributos contínuos representando genes;
- **Classes e Distribuição**: a classe de cada amostra representa o diagnóstico da célula:
  - Diagnóstico positivo: célula cancerígena (10 amostras);
  - Diagnóstico negativo: célula não cancerígena (14 amostras).

Colon

- **Descrição**: Conjunto de dados de diagnóstico de câncer de cólon em células;
- **Número de amostras**: 62;
- **Atributos**: 2000 atributos contínuos representando genes;
- **Classes e Distribuição**
  - Diagnóstico positivo: célula cancerígena (22 amostras);
  - Diagnóstico negativo: célula não cancerígena (40 amostras).

Leukemia

- **Descrição**: Base de dados de diagnóstico de leucemia em células;
- **Número de amostras**: 72;

- **Atributos**: 7129 atributos inteiros representando genes;
- **Classes e Distribuição**
  - Diagnóstico positivo: célula cancerígena (47 amostras);
  - Diagnóstico negativo: célula não cancerígena (25 amostras).

Prostate

- **Descrição**: Base de dados de diagnóstico de câncer de próstata em células;
- **Número de amostras**: 102;
- **Atributos**: 12600 atributos inteiros representando genes;
- **Classes e Distribuição**
  - Diagnóstico positivo: célula cancerígena (50 amostras);
  - Diagnóstico negativo: célula não cancerígena (52 amostras).

Lymphoma

- **Descrição**: Conjunto de dados de diagnóstico de 3 tipos de câncer linfático em células;
- **Número de amostras**: 66;
- **Atributos**: 4026 atributos contínuos representando genes;
- **Classes e Distribuição**: as classes representam 3 tipos de câncer linfático:
  - DLBCL (46 amostras);
  - FL (9 amostras);
  - CLL (11 amostras).

4.2.3 Classificadores Utilizados

A fim da realização de testes de classificação das diferentes bases de dados selecionadas neste trabalho, a suíte de mineração de dados Weka (*Waikato Environment for Knowledge Analysis*) foi escolhida. Este pacote de software visa a agregação de algoritmos e métodos de diferentes abordagens e sub-áreas de aprendizado de máquinas.

Através do Weka, são disponibilizados diferentes métodos de clusterização, classificação e análise de associação de atributos. Dentre estes, foram selecionados alguns dos principais classificadores existentes atualmente.

A seguir, uma breve descrição do funcionamento dos classificadores utilizados no presente estudo de caso.

Naive-Bayes

Um dos classificadores mais utilizados em aprendizado de máquina. Consiste em um método probabilístico que aplica o teorema de Bayes e pressupõe a indepedência condicional entre atributos. Dessa forma, assume-se que as características do modelo de dados não possuem correlação entre si.

Apesar da abordagem simplista, tal classificador apresenta desempenho bastante competitivo em tarefas de classificação.

SMO

Sequential Minimal Optimization (SMO) é uma implementação do conceito estatístico de Máquina de Vetor Suporte, *Support Vector Machine (SVM)*. Consiste em uma abordagem geométrica da tarefa de classificação, definindo o espaço amostral como um hiperplano contendo as amostras de treinamento.

Dessa forma, o processo de aprendizado e classificação é definido pela separação das amostras dentro do hiperplano definido. Este método tem sido bastante utilizado por apresentar boa acurácia.

Multilayer Perceptron

Uma das áreas da Inteligência Computacional (IA), conhecida como Redes Neurais Artificiais, baseia-se no funcionamento das redes neurais biológicas. Dessa forma, uma rede neural é estruturada a partir de um conjunto de unidades básicas de processamento, os neurônios, interligadas entre si.

O modelo de neurônio mais utilizado é conhecido como *Perceptron*. É representado por uma estrutura composta de um conjunto de sinais de entrada combinados em uma função de ativação que gera o retorno, como pode ser visto na Figura 4.2.

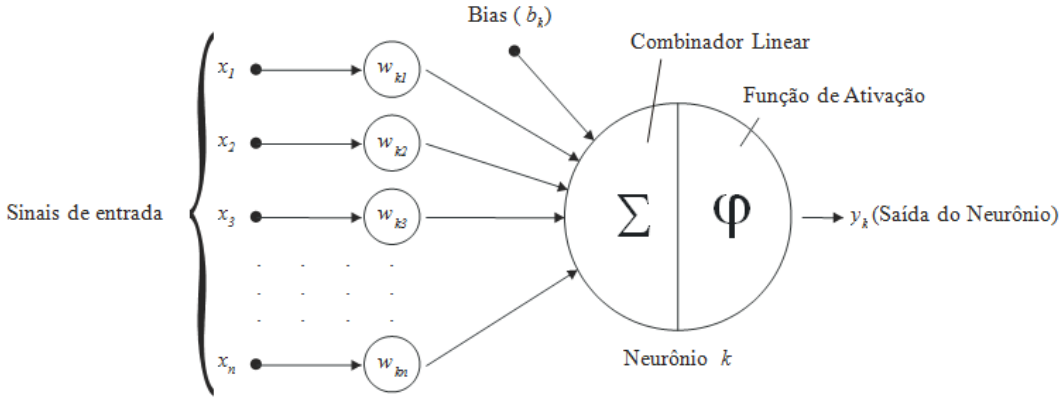

Figura 4.2: *Perceptron* (Motta, 2004)

A partir de camadas de neurônios interligados entre si, forma-se um *Multilayer Perceptron*, uma rede neural bastante utilizada em problemas de classificação de dados.

J48

J48 é uma implementação *open source* do algoritmo C4.5 de classificação por árvore de decisão.

Árvores de decisão são representadas por fluxogramas como o descrito na análise de crédito da Figura 4.3. Cada nó interno simboliza um teste de atributo, o qual embasará a decisão a ser tomada. Os nós terminais representam as classes preditas pelo classificador.

Random Forest

O método *Random Forest* consiste na utilização de várias árvores de decisão como ponto de partida para a tarefa de classificação.

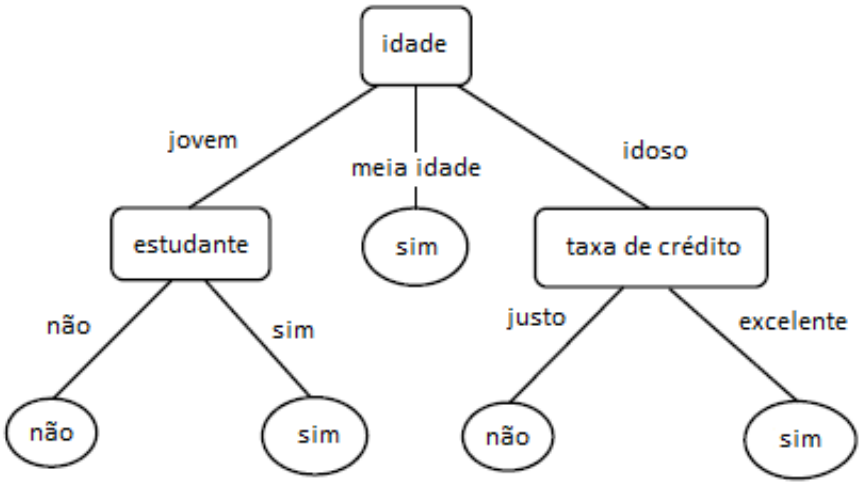

Figura 4.3: Árvore de decisão (Han and Kamber, 2006)

Para determinada amostra desconhecida, a classe predita é a que ocorre com mais frequência no conjunto resultado das sub-árvores de decisão.

4.3 Testes comparativos

Os testes comparativos foram realizados através da contrução dos classificadores a partir do conjunto de treinamento e, com base no conhecimento adquirido, o conjunto de teste é utilizado na classificação. Este processo é realizado a partir dos conjuntos de dados inicial e final.

Os resultados são apresentados, primeiramente, agrupados por base de dados com seu precentual de redução de características e os respectivos índices de desempenho (tempo de execução em milissegundos) e erro dos classificadores utlizados. Em seguida são apresentados os resultados de acordo com o classificador utilizado a partir da classificação de cada uma das base de dados.

Devido ao elevado custo computacional inerente à utilização do classificador *Multilayer Perceptron*, algumas bases de dados de alto nível dimensional não puderam ser classificadas e seus resultados são apresentados como não definido (ND).

Em determinados casos o tempo de execução foi demasiado baixo, atingindo limites abaixo da escala de milissegundos. Seus respectivos resultados são apresentados com índice de desempenho igual a 0 (zero).

4.3.1 Resultados por base de dados

Breast

Nome da Base de Dados			Breast	
Nº de Amostras			24	
Nº de Amostras de Treinamento			19	
Nº de Amostras de Teste			5	
Aspecto	Inicial	Final	Ganho	Média
Nº de Características	12625	17	99,87%	99,87%
Desempenho NSC	20	0	100,00%	95,51%
Desempenho NaiveBayes	330	0	100,00%	
Desempenho SMO	290	20	93,10%	
Desempenho Multilayer Perceptron	ND	17	ND	
Desempenho J48	180	10	94,44%	
Desempenho RandomForest	300	30	90,00%	
Erro NSC	20,00%	40,00%	-20,00%	-8,00%
Erro NaiveBayes	20,00%	40,00%	-20,00%	
Erro SMO	60,00%	40,00%	20,00%	
Erro Multilayer Perceptron	ND	60,00%	ND	
Erro J48	40,00%	60,00%	-20,00%	
Erro RandomForest	40,00%	40,00%	0,00%	

Figura 4.4: Resultados: base de dados *Breast*

Colon

Nome da Base de Dados			Colon	
Nº de Amostras			62	
Nº de Amostras de Treinamento			47	
Nº de Amostras de Teste			15	
Aspecto	Inicial	Final	Ganho	Média
Nº de Características	2000	7	99,65%	99,65%
Desempenho NSC	0	0	0,00%	76,89%
Desempenho NaiveBayes	80	0	100,00%	
Desempenho SMO	190	20	89,47%	
Desempenho Multilayer Perceptron	ND	60	ND	
Desempenho J48	210	0	100,00%	
Desempenho RandomForest	200	10	95,00%	
Erro NSC	40,00%	26,67%	13,33%	1,33%
Erro NaiveBayes	33,33%	26,67%	6,67%	
Erro SMO	26,67%	26,67%	0,00%	
Erro Multilayer Perceptron	ND	33,33%	ND	
Erro J48	46,67%	46,67%	0,00%	
Erro RandomForest	26,67%	40,00%	-13,33%	

Figura 4.5: Resultados: base de dados *Colon*

Glasses

Nome da Base de Dados			Glasses	
Nº de Amostras			214	
Nº de Amostras de Treinamento			162	
Nº de Amostras de Teste			52	
Aspecto	Inicial	Final	Ganho	Média
Nº de Características	9	5	44,44%	44,44%
Desempenho NSC	0	0	0,00%	52,54%
Desempenho NaiveBayes	10	0	100,00%	
Desempenho SMO	270	90	66,67%	
Desempenho Multilayer Perceptron	1170	290	75,21%	
Desempenho J48	30	20	33,33%	
Desempenho RandomForest	50	30	40,00%	
Erro NSC	44,23%	38,46%	5,77%	-1,60%
Erro NaiveBayes	50,00%	50,00%	0,00%	
Erro SMO	32,69%	36,54%	-3,85%	
Erro Multilayer Perceptron	42,31%	40,38%	1,92%	
Erro J48	44,23%	51,92%	-7,69%	
Erro RandomForest	25,00%	30,77%	-5,77%	

Figura 4.6: Resultados: base de dados *Glasses*

Iris

Nome da Base de Dados			Iris	
Nº de Amostras			150	
Nº de Amostras de Treinamento			114	
Nº de Amostras de Teste			36	
Aspecto	Inicial	Final	Ganho	Média
Nº de Características	4	4	0,00%	0,00%
Desempenho NSC	0	0	0,00%	0,00%
Desempenho NaiveBayes	0	0	0,00%	
Desempenho SMO	30	30	0,00%	
Desempenho Multilayer Perceptron	110	110	0,00%	
Desempenho J48	0	0	0,00%	
Desempenho RandomForest	10	10	0,00%	
Erro NSC	2,78%	2,78%	0,00%	0,00%
Erro NaiveBayes	0,00%	0,00%	0,00%	
Erro SMO	2,78%	2,78%	0,00%	
Erro Multilayer Perceptron	2,78%	2,78%	0,00%	
Erro J48	2,78%	2,78%	0,00%	
Erro RandomForest	0,00%	0,00%	0,00%	

Figura 4.7: Resultados: base de dados Iris

Leukemia

Nome da Base de Dados			Leukemia	
Nº de Amostras			72	
Nº de Amostras de Treinamento			54	
Nº de Amostras de Teste			18	
Aspecto	Inicial	Final	Ganho	Média
Nº de Características	7129	8	99,89%	99,89%
Desempenho NSC	23	0	100,00%	97,17%
Desempenho NaiveBayes	260	0	100,00%	
Desempenho SMO	200	20	90,00%	
Desempenho Multilayer Perceptron	ND	70	ND	
Desempenho J48	340	0	100,00%	
Desempenho RandomForest	240	10	95,83%	
Erro NSC	0,00%	5,56%	-5,56%	2,22%
Erro NaiveBayes	0,00%	5,56%	-5,56%	
Erro SMO	5,56%	11,11%	-5,56%	
Erro Multilayer Perceptron	ND	11,11%	ND	
Erro J48	22,22%	5,56%	16,67%	
Erro RandomForest	16,67%	5,56%	11,11%	

Figura 4.8: Resultados: base de dados *Leukemia*

Lymphoma

Nome da Base de Dados			Lymphoma	
Nº de Amostras			66	
Nº de Amostras de Treinamento			50	
Nº de Amostras de Teste			16	
Aspecto	Inicial	Final	Ganho	Média
Nº de Características	4026	3026	24,84%	24,84%
Desempenho NSC	12	10	16,67%	36,85%
Desempenho NaiveBayes	90	80	11,11%	
Desempenho SMO	240	220	8,33%	
Desempenho Multilayer Perceptron	ND	ND	ND	
Desempenho J48	320	70	78,13%	
Desempenho RandomForest	200	60	70,00%	
Erro NSC	6,25%	6,25%	0,00%	0,00%
Erro NaiveBayes	25,00%	25,00%	0,00%	
Erro SMO	0,00%	0,00%	0,00%	
Erro Multilayer Perceptron	ND	ND	ND	
Erro J48	12,50%	12,50%	0,00%	
Erro RandomForest	12,50%	12,50%	0,00%	

Figura 4.9: Resultados: base de dados *Lymphoma*

Prostate

Nome da Base de Dados			Prostate	
Nº de Amostras			102	
Nº de Amostras de Treinamento			77	
Nº de Amostras de Teste			25	
Aspecto	Inicial	Final	Ganho	Média
Nº de Características	12600	2	99,98%	99,98%
Desempenho NSC	60	0	100,00%	96,35%
Desempenho NaiveBayes	20	0	100,00%	
Desempenho SMO	490	50	89,80%	
Desempenho Multilayer Perceptron	ND	830	ND	
Desempenho J48	560	10	98,21%	
Desempenho RandomForest	320	20	93,75%	
Erro NSC	44,00%	28,00%	16,00%	-2,40%
Erro NaiveBayes	44,00%	28,00%	16,00%	
Erro SMO	4,00%	20,00%	-16,00%	
Erro Multilayer Perceptron	ND	24,00%	ND	
Erro J48	16,00%	28,00%	-12,00%	
Erro RandomForest	12,00%	28,00%	-16,00%	

Figura 4.10: Resultados: base de dados *Prostate*

4.3.2 Resultados por classificador

NSC

Classificador		NSC			
Aspecto	Base de Dados	Inicial	Final	Ganho	Média
Desempenho	Breast	20	0	100,00%	45,24%
	Colon	0	0	0,00%	
	Glasses	0	0	0,00%	
	Iris	0	0	0,00%	
	Leukemia	23	0	100,00%	
	Lymphoma	12	10	16,67%	
	Prostate	60	0	100,00%	
Erro	Breast	20,00%	40,00%	-20,00%	0,41%
	Colon	33,33%	26,67%	6,67%	
	Glasses	44,23%	38,46%	5,77%	
	Iris	2,78%	2,78%	0,00%	
	Leukemia	0,00%	5,56%	-5,56%	
	Lymphoma	6,25%	6,25%	0,00%	
	Prostate	44,00%	28,00%	16,00%	

Figura 4.11: Resultados: classificador NSC

Naive-Bayes

Classificador		NaiveBayes			
Aspecto	Base de Dados	Inicial	Final	Ganho	Média
Desempenho	Breast	330	0	100,00%	44,44%
	Colon	80	0	0,00%	
	Glasses	10	0	0,00%	
	Iris	0	0	0,00%	
	Leukemia	260	0	100,00%	
	Lymphoma	90	80	11,11%	
	Prostate	20	0	100,00%	
Erro	Breast	20,00%	40,00%	-20,00%	-0,41%
	Colon	33,33%	26,67%	6,67%	
	Glasses	50,00%	50,00%	0,00%	
	Iris	0,00%	0,00%	0,00%	
	Leukemia	0,00%	5,56%	-5,56%	
	Lymphoma	25,00%	25,00%	0,00%	
	Prostate	44,00%	28,00%	16,00%	

Figura 4.12: Resultados: classificador *Naive-Bayes*

SMO

Classificador		SMO			
Aspecto	Base de Dados	Inicial	Final	Ganho	Média
Desempenho	Breast	290	20	93,10%	63,05%
	Colon	190	20	89,47%	
	Glasses	270	90	66,67%	
	Iris	30	30	0,00%	
	Leukemia	200	20	90,00%	
	Lymphoma	240	220	8,33%	
	Prostate	320	20	93,75%	
Erro	Breast	60,00%	40,00%	20,00%	-0,77%
	Colon	26,67%	26,67%	0,00%	
	Glasses	32,69%	36,54%	-3,85%	
	Iris	2,78%	2,78%	0,00%	
	Leukemia	5,56%	11,11%	-5,56%	
	Lymphoma	0,00%	0,00%	0,00%	
	Prostate	4,00%	20,00%	-16,00%	

Figura 4.13: Resultados: classificador SMO

Multilayer Perceptron

Classificador		Multilayer Perceptron			
Aspecto	Base de Dados	Inicial	Final	Ganho	Média
Desempenho	Breast	ND	17	ND	37,61%
	Colon	ND	60	ND	
	Glasses	1170	290	75,21%	
	Iris	110	110	0,00%	
	Leukemia	ND	70	ND	
	Lymphoma	ND	ND	ND	
	Prostate	ND	830	ND	
Erro	Breast	ND	60,00%	ND	0,96%
	Colon	ND	33,33%	ND	
	Glasses	42,31%	40,38%	1,92%	
	Iris	2,78%	2,78%	0,00%	
	Leukemia	ND	11,11%	ND	
	Lymphoma	ND	ND	ND	
	Prostate	ND	24,00%	ND	

Figura 4.14: Resultados: classificador *Multilayer Perceptron*

J48

Classificador		J48			
Aspecto	Base de Dados	Inicial	Final	Ganho	Média
Desempenho	Breast	180	10	94,44%	72,02%
	Colon	210	0	100,00%	
	Glasses	30	20	33,33%	
	Iris	0	0	0,00%	
	Leukemia	340	0	100,00%	
	Lymphoma	320	70	78,13%	
	Prostate	560	10	98,21%	
Erro	Breast	40,00%	60,00%	-20,00%	-3,29%
	Colon	46,67%	46,67%	0,00%	
	Glasses	44,23%	51,92%	-7,69%	
	Iris	2,78%	2,78%	0,00%	
	Leukemia	22,22%	5,56%	16,67%	
	Lymphoma	12,50%	12,50%	0,00%	
	Prostate	16,00%	28,00%	-12,00%	

Figura 4.15: Resultados: classificador J48

Random Forest

Classificador		RandomForest			
Aspecto	Base de Dados	Inicial	Final	Ganho	Média
Desempenho	Breast	300	30	90,00%	69,23%
	Colon	200	10	95,00%	
	Glasses	50	30	40,00%	
	Iris	10	10	0,00%	
	Leukemia	240	10	95,83%	
	Lymphoma	200	60	70,00%	
	Prostate	320	20	93,75%	
Erro	Breast	40,00%	40,00%	0,00%	-3,43%
	Colon	26,67%	40,00%	-13,33%	
	Glasses	25,00%	30,77%	-5,77%	
	Iris	0,00%	0,00%	0,00%	
	Leukemia	16,67%	5,56%	11,11%	
	Lymphoma	12,50%	12,50%	0,00%	
	Prostate	12,00%	28,00%	-16,00%	

Figura 4.16: Resultados: classificador *Random Forest*

5 Considerações Finais

A seguir é apresentada uma síntese dos resultados médios obtidos nos testes comparativos entre as sete bases de dados (Figura 5.1) e entre os seis classificadores (Figura 5.2).

Essa síntese retrata claramente os efeitos produzidos pelo processo de seleção de características no que diz respeito à redução obtida no número de atributos em cada base de dados, bem como a variação de desempenho e de acurácia.

Base de Dados	Breast	Colon	Glasses	Iris	Leukemia	Lymphoma	Prostate	Média
Redução nº de Carasterísticas	99,87%	99,65%	44,44%	0,00%	99,89%	24,84%	99,98%	**66,95%**
Ganho Médio Desempenho	95,51%	76,89%	52,54%	0,00%	97,17%	36,85%	76,35%	**62,19%**
Ganho Médio Erro	-8,00%	1,33%	-1,60%	0,00%	2,22%	0,00%	-2,40%	**-1,21%**

Figura 5.1: Variações de redução de atributos, de desempenho e de erro por base de dados

Classificador	NSC	NaiveBayes	SMO	Multilayer Perceptron	J48	RandomForest	Média
Ganho Médio Desempenho	45,24%	44,44%	63,05%	37,61%	72,02%	69,23%	55,26%
Ganho Médio Erro	0,41%	-0,41%	-0,77%	0,96%	-3,29%	-3,43%	-1,09%

Figura 5.2: Variações de desempenho e de erro por classificador

Através da análise dos resultados obtidos pelos testes de comparativos, pode-se destacar os impactos causados pela seleção de características sob os seguintes aspectos:

- **Número de Atributos × Desempenho**

  Em linhas gerais, os testes realizados apresentaram uma grande redução do número de atributos dos modelos de dados analisados, acarretando assim, em um ganho considerável de desempenho.

  Os maiores índices de redução de atributos e ganho de desempenho ocorreram em bases de dados de alto nível dimensional, como *Breast* (12625 atributos), *Colon* (2000 atributos), *Leukemia* (7129 atributos) e *Prostate* (12600 atributos). Bases de dados com poucos atributos, como *Glasses* (9 atributos) e *Iris* (4 atributos), sofreram pouco ou nenhum impacto no número de atributos eliminados.

- **Acurácia**

  Alguns classificadores se ajustaram melhor ao processo de seleção de características realizado, apresentando ganho ou pouca redução de acurácia. São os casos dos classificadores *NSC* (+0,41%), *Multilayer Perceptron* (+0,96%), *Naive-Bayes* (-0,41%) e *SMO* (-0,77%). Os classificadores baseados em ávores de decisão (*J48* e *Random Forest*) obtiveram os piores resultados.

  De forma análoga, as bases de dados também apresentaram conportamentos diferentes em relação à variação da acurácia. A grande maioria das bases se comportou bem diante da redução de atributos, embora a base *Breast* tenha apresentado o maior aumento de erro (8%).

- **Desempenho × Acurácia**

  Em alguns casos, os índices de ganho de desempenho e de acurácia caminharam positivamente em paralelo, conduzindo a um cenário de considerável redução do número de atributos acompanhado de melhoria do processo de classificação. Deles podemos citar as bases de dados *Colon*, *Leukemia* e *Lymphoma*, todas de alto nível dimensional.

A grande maioria dos casos presentes neste estudo apresentaram melhora considerável de desempenho e perda pouco significativa de acurácia dos classificadores e bases de dados envolvidos.

Desta forma, conclui-se que o método *Nearest Shrunken Centroids* é bastante eficiente e pode ser utilizado para seleção de características, principalmente em problemas cujos modelos de dados apresentam alto nível dimensional.

5.1 Disponibilização do Sistema Inteligente

O código fonte do sistema inteligente implementado está disponível sob a lincença de *software* livre *GNU GPL v3* (http://www.gnu.org/licenses/gpl.html) e pode ser acessado a partir de um repositório *Subversion (SVN)* das seguintes formas:

- Através do endereço:
  - http://nscc.googlecode.com/svn/trunk/
- Por meio de um cliente SVN
  - svn checkout http://nscc.googlecode.com/svn/trunk/

5.2 Trabalhos Futuros

Como forma de continuidade do estudo a respeito do processo de seleção de características e classificação em bases de dados, tem-se em mente a realização de estudos comparativos entre o método *Nearest Shrunken Centroids* e outras técnicas de seleção de características disponíveis atualmente.

Referências Bibliográficas

Dietterich, T. G. **Machine Learning**. Annual Review of Computer Science, 1990.

Usama Fayyad, G. P.-S.; Smyth, P. **From Data Mining to Knowledge Discovery: An Overview, In: Advances in Knowledge Discovery and Data Mining**. AAAI/MIT Press, 1996.

Goldschmidt, R.; Passos, E. **Data Mining: Um Guia Prático - Conceitos, Técnicas, Ferramentas, Orientações e Aplicações**. Editora Campus, 2005.

Han, J.; Kamber, M. **Data Mining: Concepts and Techniques**. 2nd. ed., Diane Cerra, 2006.

Klassen, M.; Kim, N. Nearest shrunken centroid as feature selection of microarray data. **California Lutheran University**, 2009.

da Motta, C. G. L. **Sistema inteligente para avaliação de riscos em vias de transporte terrestre**. Universidade Federal do Rio de Janeiro, COPPE, 2004.

Reich, Y.; Barai, S. V. **Evaluating Machine Learning Models for Engineering Problems**. Atificial Intelligence in Engineering, 1999.

S. O. Rezende, J. B. Pugliese, E. A. M.; Paula, M. F. **Mineração de Dados: Sistemas Inteligentes - Fundamentos e Aplicações**. Editora Manole Ltda, 2003.

Santoro, D. M. **Sobre o processo de seleção de subconjuntos de atributos - as abordagens filtro e wrapper**. São Carlos: UFSCar, 2005.

Silver, D. L. **Knowledge Discovery and Data Mining**. MBA course notes of Dalhousie University, 1998.

Robert Tibshirani, Trevor Hastie, B. N.; Chu, G. Class prediction by nearest shrunken centroids, with applications to dna microarrays. **PNAS**, 2002.

Robert Tibshirani, Trevor Hastie, B. N.; Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. **Stanford University**, 2002.

Ian H. Witten, E. F.; Hall, M. A. **Data Mining: Practical Machine learning Tools and Techniques**. 3rd. ed., Elsevier Inc., 2011.