COPPE/UFRJ

IMPLEMENTAÇÃO DA INFORMAÇÃO MÚTUA USANDO ENTROPIA DE RENYI E APLICAÇÃO NA SELEÇÃO DE GENES EM EXPERIMENTOS DE MICROARRANJOS

Claudia Teixeira de Araujo

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Biomédica.

Orientadores: Flávio Fonseca Nobre

Carlos Eduardo Pedreira

Rio de Janeiro

Novembro de 2008

IMPLEMENTAÇÃO DA INFORMAÇÃO MÚTUA USANDO ENTROPIA DE RENYI E APLICAÇÃO NA SELEÇÃO DE GENES EM EXPERIMENTOS DE MICROARRANJOS

Claudia Teixeira de Araujo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

Aprovada por:

_____
Prof. Flávio Fonseca Nobre, Ph.D.

_____
Prof. Carlos Eduardo Pedreira., Ph.D.

_____
Dra. Elaine Sobral da Costa, D.Sc.

_____
Prof. Márcio Nogueira de Souza, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
NOVEMBRO DE 2008

Araujo, Claudia Teixeira de

Implementaçao da informação mútua usando entropia de Renyi e aplicação na seleção de genes em experimentos de microarranjos / Claudia Teixeira de Araujo. – Rio de Janeiro: UFRJ/COPPE, 2008.

IX, 64 p.: 29,7 cm.

Orientadores: Flávio Fonseca Nobre

Carlos Eduardo Pedreira

Dissertação (Mestrado) – UFRJ/ COPPE/ Programa de Engenharia Biomédica, 2008.

Referências Bibliográficas: p. 58-61

1. Microarranjos de DNA. 2. Informação mútua. 3. Entropia de Renyi. 4. Seleção de atributos. 5. Leucemia. I. Nobre, Flavio Fonseca, *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Titulo.

Agradecimentos

Aos meus pais, Olívia e José Carlos, por todo amor e confiança. Estes sempre serão meus melhores mestres e exemplo a ser seguido.

Ao professor Flávio, cuja orientação foi essencial para o desenvolvimento desta dissertação e por todo apoio e amizade.

Ao professor Carlos Pedreira e sua valiosa colaboração desde o início do projeto até a revisão final.

Aos amigos e colegas do laboratório de Engenharia de Sistemas de Saúde, tão importantes nos momentos de dificuldades e nas horas de lazer, tornando muito mais agradáveis os dias de trabalho.

Aos demais professores e funcionários do Programa de Engenharia Biomédica, responsáveis pela excelente formação a mim oferecida.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

IMPLEMENTAÇÃO DA INFORMAÇÃO MÚTUA USANDO ENTROPIA DE RENYI E APLICAÇÃO NA SELEÇÃO DE GENES EM EXPERIMENTOS DE MICROARRANJOS

Claudia Teixeira de Araujo

Novembro/2008

Orientadores: Flávio Fonseca Nobre
Carlos Eduardo Pedreira

Programa: Engenharia Biomédica

Este trabalho desenvolve um modelo de seleção de atributos em experimentos de microarranjos de DNA, fundamentais para a análise de expressão gênica global. Os microarranjos podem ser de grande utilidade na classificação de subtipos de neoplasias a partir da análise da expressão dos milhares de genes extraídos de biópsias. Este processo tem como principais dificuldades a grande dimensão da matriz de expressão gênica e as amostras disponíveis para análise apresentarem reduzido número de observações. Nesse estudo, propomos a seleção de genes informativos através de um algoritmo baseado na informação mútua substituindo-se a entropia de Shannon pela entropia de Renyi, obtendo-se um estimador diretamente das amostras. O método proposto foi aplicado ao conjunto de dados de leucemia aguda e como classificador, utilizou-se máquina de vetores de suporte (SVM), com mapeamento não-linear obtido pelo emprego de núcleos Gaussianos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

GENE SELECTION FOR MICROARRAY EXPERIMENTS BY MUTUAL INFORMATION BASED ON RENYI'S ENTROPY

Claudia Teixeira de Araujo

November/2008

Advisors: Flávio Fonseca Nobre
Carlos Eduardo Pedreira

Department: Biomedical Engineering

This work presents a feature selection model in microarray experiments, one of the most promising technologies for global gene expression analysis. Mycroarrays can be very useful in classification of subtypes of cancer based on gene expression from biopsies. The main difficulties in this process are the large dimension of gene expression matrix and the small number of samples in database. The objective of this work is to develop and test a mutual information feature selection approach based on Renyi´s entropy which estimates probability density distribuitions directly from data. This algorithm was used to rank informative genes from a DNA microarray expression of acute leukemia. To evaluate the selected subset of informative genes a support vector machine classifier (SVM) with a Gaussian kernel was used.

Índice

Índice das Figuras

Índice das Tabelas

Capítulo 1

1.1 Introdução

Esta dissertação apresenta uma nova metodologia para identificação de genes capazes de diferenciar duas formas de leucemia, a mielóide aguda (AML) e a linfóide aguda (ALL), em conjuntos de dados de microarranjos de DNA.

A metodologia aqui proposta não se encerra em problemas de diferenciação de subtipos de câncer, nem em problemas de origem biológica. Pode ser aplicada às mais diversas áreas onde são estudados problemas envolvendo conjuntos de dados com grande número de variáveis de entrada e amostra reduzida, desde que os rótulos das classes sejam conhecidos.

Devido ao caráter multidisciplinar do presente estudo, os capítulos a seguir focarão tanto em assuntos de interesse daqueles cuja formação tem como base as ciências exatas, quanto daqueles cuja formação tem como base as ciências da saúde. Sendo assim, alguns itens aqui discutidos parecerão triviais, enquanto outros se mostrarão absolutamente novos.

A seção reservada aos microarranjos inicia-se com um enfoque biológico seguindo posteriormente para a discussão e comparação de algumas ferramentas matemáticas utilizadas na análise de dados oriundos do estudo de microarranjos de DNA.

Informação mútua e máquinas de vetores de suporte recebem um tratamento matemático com o objetivo de facilitar a replicação ou implementação da metodologia aqui apresentada, caso haja interesse.

Finalmente, a seção reservada à leucemia visa contextualizar o leitor não familiarizado com os aspectos principais desta enfermidade, tais como: onde as leucemias mielóide e linfóide se originam, quais as características que compartilham, como se desenvolvem, quais as formas de diagnósticos usuais e suas limitações, visando uma melhor compreensão do problema aqui estudado e de sua relevância.

Microarrajos de DNA é uma tecnologia fundamental para a análise de expressão gênica global. Tem sido demonstrado que os microarranjos podem ser de

grande utilidade na classificação de subtipos de neoplasias a partir da análise da expressão dos milhares de genes extraídos de biópsias [1, 2, 3].

Este processo tem como principais dificuldades a grande dimensão da matriz de expressão gênica, o que resulta em aumento da complexidade computacional e, tipicamente, as amostras disponíveis para análise apresentam reduzido número de observações, devido, em geral, a raridade da enfermidade investigada.

No contexto de classificação de padrões, em particular, na classificação dos subtipos de câncer, diversas ferramentas estão sendo desenvolvidas visando superar estas dificuldades técnicas e viabilizar a identificação dos genes (atributos) relevantes para a discriminação das diferentes classes (subtipos de câncer).

Os diferentes métodos para seleção de variáveis em ambiente supervisionado, isto é, quando são conhecidos os rótulos das amostras pesquisadas (classes) podem ser agrupados em dois tipos:

a) métodos de filtragem – Em microarranjos de DNA são chamados de abordagem de ordenação individual de genes
b) métodos envelopados – Em microarranjos de DNA são chamados de abordagem de ordenação de subconjuntos de genes.

Os métodos envelopados, popularizados por Kohavi e John [4], utilizam algoritmo treinado para a discriminação das classes e o desempenho do classificador é avaliado, via uma função de erro, para cada subconjunto do conjunto original. Os subconjuntos são obtidos através da inclusão de atributos um a um e o processo é finalizado quando o desempenho ao se adicionar qualquer novo atributo é menor que o apresentado pelo subconjunto de atributos anteriormente selecionados.

Segundo Xiaohua Hu e Yi Pan [5] os métodos envelopados são geralmente criticados devido ao custo computacional, principalmente em conjuntos de grande dimensão e o baixo poder de generalização. O primeiro porque o algoritmo de aprendizagem treinado deve ser uma sub-rotina que necessita ser repetida a cada nova interação e o segundo deve-se ao fato de que os candidatos a subconjuntos são determinados com base na acurácia do classificador predeterminado e assim, o subconjunto de atributos ótimos se ajusta apenas para este classificador especificamente. Para o mesmo conjunto de dados, diferentes subconjuntos ótimos de atributos podem ser selecionados ao serem aplicados diferentes classificadores. Kohavi e John [4] baseados em seus experimentos, afirmam que a superespecialização é um problema principalmente quando o número de observações é pequeno.

Os métodos de filtragem ordenam os genes de acordo com critérios pré-estabelecidos (medidas), por exemplo através da medida da informação mútua ou testes estatísticos tais como T-teste e F-Teste. Estas medidas e testes estimam a relevância de um atributo com a classe alvo. Vale notar que o conjunto de atributos selecionados depende de características intrínsecas ao conjunto de treinamento e não de um classificador em particular, uma vez que a seleção dos atributos relevantes ocorre como uma etapa de pré-processamento, sem envolver um algoritmo de aprendizagem específico. Em particular, este método é computacionalmente vantajoso, pois requer apenas o cálculo da relevância entre dos **N** atributos de entrada e a classe-alvo.

Especificamente em microarranjos os métodos de filtragem são chamados de abordagem de ordenação individual dos genes. O número de atributos necessários é geralmente determinado por tentativa e erro e sua ordenação ocorre de acordo com a correlação de cada gene com o desfecho (classe), segundo algum critério de medida (distância Euclidiana, distância de Bhattacharyya, índice de correlação de Pearson, informação mútua etc.). O custo computacional é significativamente inferior ao método de ordenação dos subconjuntos de genes, mas por outro lado, apresenta como desvantagem o fato de a correlação entre os atributos não ser avaliada. Sendo assim, genes correlacionados individualmente com as distribuições de classe podem não apresentar conjuntamente o melhor desempenho no processo de classificação. Além disso, alguns atributos podem conter a mesma informação, introduzindo redundância ao sistema. Por outro lado, atributos que possuem informações complementares entre si para a discriminação das classes não serão selecionados caso não apresentem alta correlação individual com o desfecho.

Um bom exemplo de ordenação individual de genes é o método GS apresentado por Golub *et al.* [6] que propõe a utilização de uma métrica de correlação que mede a separação relativa entre as classes, produzida pelos valores dos níveis de expressão de um gene.

Uma proposta inovadora para filtragem de genes informativos é a utilização da informação mútua que, apesar de pouco utilizada na área biológica [7, 8], foi largamente aplicada em engenharia de comunicações desde a década de 60. A informação mutua (IM) é uma medida de independência estatística entre variáveis aleatórias e, portanto, pode ser utilizada como um indicador de relevância entre atributos e o desfecho. Pode também mensurar a redundância quando calculada entre atributos. Entretanto, a estimativa da IM para variáveis contínuas tradicionalmente é feita utilizando-se a entropia de Shannon, o que não é tarefa simples, pois requer a estimativa da função densidade de probabilidade (pdf) e a integração destas funções.

Buscando contornar este problema, Príncipe *et al.* [9] desenvolveram um método para calcular a informação mútua utilizando a entropia quadrática de Renyi, obtendo um estimador de informação mútua baseado diretamente nas observações.

Nesta dissertação propomos a seleção de genes informativos através de um algoritmo de ordenação de genes baseado no algoritmo apresentado por Kwak e Choi [10] chamado de MIFS-U (*Mutual Information Feature Selector under Uniform Distribuition*) que introduz um parâmetro que considera a redundância entre as variáveis de entrada, com o objetivo de minorar os problemas relacionados aos métodos de filtragem. Este método utiliza como critério de medida a informação mútua estimada pela entropia de Shannon e a contribuição metodológica inovadora que introduzimos é a substituição, da entropia de Shannon pela entropia de Renyi, batizando o novo algoritmo de MIFS-U-Renyi.

Como classificador, utilizou-se máquina de vetores de suporte (SVM), que vem sendo empregada com sucesso em múltiplas áreas de análise de dados biológicos, incluindo análise de microarranjos de DNA [3].

O método proposto foi testado no conjunto de dados de Leucemia, apresentada pela primeira vez por Golub *et al.* [6], uma vez que este conjunto já possui vários estudos publicados facilitando a tarefa de comparação e validação do método proposto.

As duas formas predominantes de leucemia aguda são a Leucemia Mielóide Aguda (AML) e a Leucemia Linfóide Aguda (ALL), que apesar de morfologicamente semelhantes e de compartilharem muitas características clínicas (anemias, infecções, hemorragias, etc.), diferem quanto à história natural, prognóstico e resposta aos diversos agentes terapêuticos. O sucesso do tratamento depende fundamentalmente de um diagnóstico correto e precoce.

As estimativas de incidência de câncer publicadas pelo Instituto Nacional do Câncer [11], prevêem que no Brasil ocorrerão 9.540 novos casos de leucemias este ano, sendo 5.220 homens e 4320 mulheres. A tabela 1 mostra a estimativa de novos casos de Leucemia, por estado, para o ano de 2008.

Segundo Golub *et al.* [6], não há teste único, que seja suficiente para discriminar entre AML e ALL. São necessários hematopatologistas experientes para interpretação da morfologia celular em esfregaços de medula óssea, análise histoquímica, imunofenotípica e citogenética, cada uma delas realizadas separadamente em laboratórios especializados. Essas etapas fazem com que a classificação da leucemia ainda seja imperfeita.

A correta identificação dos genes diferencialmente expressos e o desenvolvimento de novas ferramentas de diagnóstico trazem grandes benefícios aos

pacientes, uma vez que o diagnóstico preciso é fundamental para o sucesso terapêutico. Permite também elucidar os mecanismos de funcionamento da doença ao nível molecular, o que auxilia no desenvolvimento de novos fármacos de ação especifica minorando os efeitos colaterais.

Tabela 1: Estimativas, para o ano 2008, de número de casos novos de câncer, por Estado.*

Estados	Leucemias	Estados	Leucemias
Acre	20	Paraíba	150
Alagoas	110	Paraná	630
Amapá	30	Pernambuco	350
Amazonas	120	Piauí	100
Bahia	410	Rio de Janeiro	960
Ceará	360	Rio Grande do Norte	170
Distrito Federal	120	Rio Grande do Sul	810
Espírito Santo	190	Rondônia	50
Goiás	270	Roraima	20
Maranhão	180	Santa Catarina	380
Mato Grosso	130	São Paulo	2.530
Mato Grosso do Sul	130	Sergipe	70
Minas Gerais	960	Tocantins	50
Pará	240	**Brasil**	**9.540**

* Números arredondados para 10 ou múltiplo de 10.

1.2 Objetivos

Nesta dissertação será desenvolvido e implementado um algoritmo para seleção de atributos, usando a informação mútua estimada através da entropia de Renyi, com a finalidade de realizar seleção de genes informativos em experimentos de microarranjos relacionados com leucemia. O objetivo central é a classificação das duas formas predominantes de leucemia aguda, AML e ALL, através de um classificador de Máquina de Vetores de Suporte (SVM). A eficácia desse modelo foi comparada com outros estudos, em particular com os que utilizaram informação mútua estimada através da entropia de Shannon.

1.3 Estrutura do Trabalho

No capítulo 2 são apresentados os fundamentos teóricos de microarranjos de DNA, de informação mútua, de Máquinas de Vetores de suporte e de câncer e Leucemia.

No capítulo 3 é apresentado o algoritmo proposto na dissertação, denominado MIFS_U_Renyi e o resultado da implementação testada em bancos de dados para avaliação da metodologia, através da comparação com os resultados obtidos ao se implementar o algoritmo cuja informação mútua é estimada pela entropia de Shannon, conhecido como MIFS_U.

No capitulo 4 os resultados são discutidos e comparados com outros estudos que utilizaram o mesmo banco de dados para classificação das duas formas predominantes de leucemia aguda.

No capítulo 5 a conclusão é apresentada.

Capítulo 2

Fundamentação Teórica

Neste capítulo serão descritas brevemente as bases teóricas para a compreensão das ferramentas utilizadas na construção da metodologia proposta nesta dissertação.

2.1 Microarranjos de DNA

Nesta seção são apresentados os microarranjos de DNA, permitindo o conhecimento da tecnologia envolvida em sua fabricação, os principais problemas decorrentes deste processo e alguns métodos aplicados à análise de dados oriundos desta técnica.

Microarranjos de DNA podem prover importante informação para o diagnóstico do câncer através do nível de expressão gênica, devido à habilidade de medir a quantidade de RNA transcrito para milhares de genes simultaneamente [12, 13], visando compreender as variações moleculares nos processos relacionados às doenças celulares.

A molécula de DNA contém a informação genética que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus. Consiste de duas fitas antiparalelas entrelaçadas na forma de dupla-hélice [14]. Cada fita é composta por uma seqüência de nucleotídeos, 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 fita homóloga, de maneira específica: Adenina com Timina, Citosina com Guanina. Os dois primeiros nucleotídeos (A e T) ligam-se por duas pontes de hidrogênio, enquanto os dois últimos (C e G) ligam-se por meio de três pontes de hidrogênio (Fig. 2) que estabilizam a estrutura da dupla hélice.

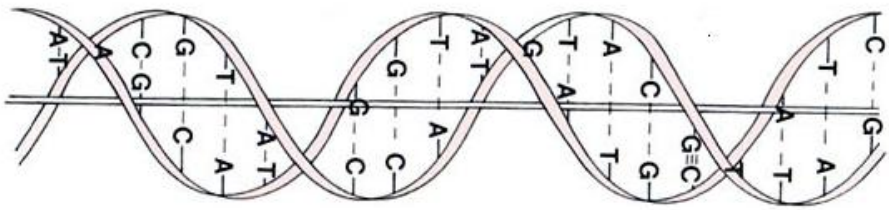

Figura 1: Estrutura do DNA

Uma das principais funções do DNA é armazenar as informações necessárias para a construção das proteínas. A transferência de informação do DNA para a proteína acontece através da síntese de uma molécula intermediária conhecida como RNA. O RNA, assim como o DNA, é constituído de uma seqüência de nucleotídeos composta por quatro bases complementares. No RNA a base timina é trocada pela uracila (U).

Moléculas de RNA são sintetizadas a partir do DNA por um processo conhecido como transcrição, que utiliza uma fita do DNA como molde. Vários tipos de RNA são transcritos, mas somente o chamado mRNA (RNA mensageiro) serve como molde para a síntese de proteínas [15].

Após o processo de desnaturação (enzimático ou por calor), onde há a separação entre as fitas homólogas, a fita simples de DNA tem a capacidade de se religar a seqüências homólogas em um processo chamado de hibridização.

Os microarranjo de DNA complementar, ou cDNA [16], também chamado de Chip de DNA, opera através do princípio de hibridização de moléculas com seqüências homólogas ou complementares. Nessa técnica (figura 3), são impressos ordenadamente em uma lâmina milhares de fragmentos de DNA, sejam oligonucleotídeos, cDNA ou ESTs (*Expressed Sequence Tags*) representativos de genes. A célula é estudada em diferentes condições (por exemplo, amostra e controle) e seu mRNA é extraído. Esses conjuntos de mRNAs são reversamente transcritos a cDNA na presença de nucleotídeos marcados com diferentes substâncias fluorescentes, por exemplo, corantes de cianina (Cy3 e Cy5), ou menos freqüentemente usados, os corantes fluoresceína e rodamina. Feito isso, esses dois tipos de cDNAs marcados são misturados e hibridizados contra a sonda de DNA impressa na lâmina.

As condições necessárias à hibridização (concentração da amostra, força iônica, temperatura, etc.) são fortemente dependentes das dimensões das sondas de DNA presentes no arranjo e, por isso, devem ser determinadas individualmente em

cada experimento. As seqüências dos genes para as quais existem cDNA marcado vão se ligar a essa sonda, resultando em pontos fluorescentes na lâmina [17].

A variação natural do processo de hibridização, a possibilidade de depósito de poeira ou espalhamento de corantes nas lâminas produz resultados ruidosos (artefatos) e uma das formas de minorar esses efeitos negativos é a replicação de cada experimento.

Para a visualização, é feita a lavagem do material não hibridizado e a varredura eletrônica utilizando-se câmeras CCD *(Charge Coupled Device)*, *scaners* a laser não-confocais e *scaners* a laser confocais. Os dados gerados durante essa varredura formecem imagens formadas por pontos de diferentes intensidades. Utilizando os valores das intensidades, a imagem é pseudocolorida, onde a cor de cada ponto indica a condição fisiológica na qual aquele gene foi expresso e a intensidade da cor será proporcional à sua intensidade de expressão (figura2).

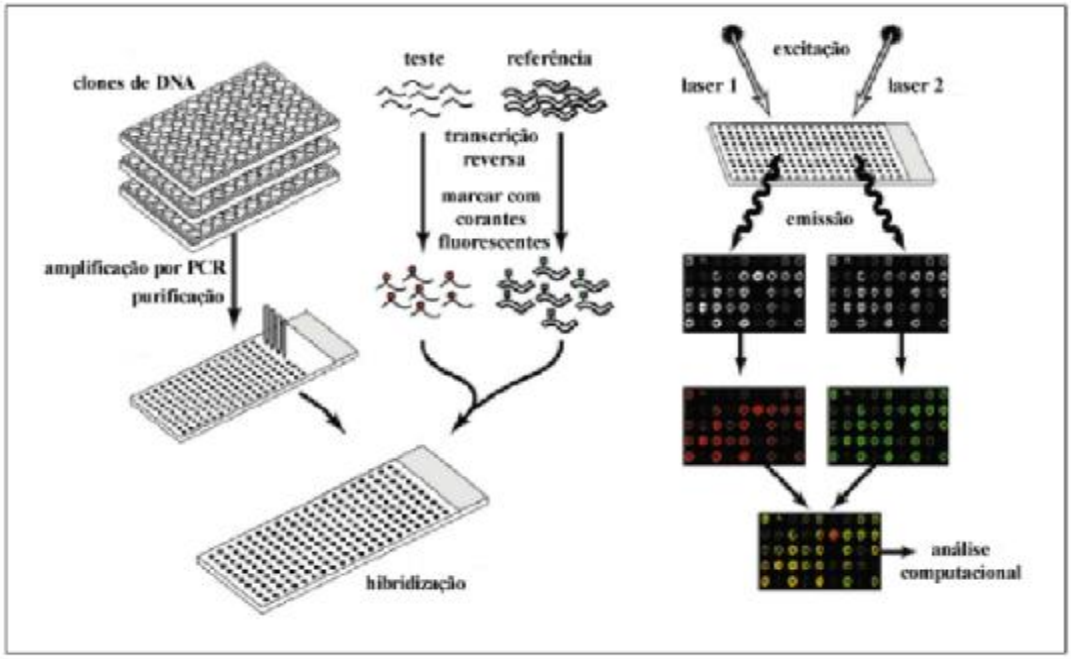

Figura 2: Arranjos de cDNA. Sondas são transferidas por um robô para uma lâmina de vidro. O mRNA das células são extraídos e convertidos a cDNA (transcriptase reversa) e marcados com corantes fluorescentes. Em geral, a amostra de controle é marcada em verde e a de interesse em vermelho. Após serem misturadas, são hibridizadas às sondas na lâmina de vidro. Após lavagem do material não-hibridizado, a lâmina é escaneada com um microscópio confocal a laser e a imagem é computacionalmente analisada. [18]

Pontos vermelhos indicam que um dado gene está induzido na amostra de teste em relação à amostra controle; pontos verdes indicam que um dado gene está

reprimido na amostra de teste em relação à amostra controle; e pontos amarelos indicam que não há diferença significante no nível de expressão de determinado gene entre as duas amostras [19].

Os dados de microarranjos são geralmente organizados como uma matriz M(n x m), onde as linhas representam os genes e as colunas representam as amostras biológicas cujo nível de expressão foi monitorado.

Matematicamente, o nível de expressão gênica é dado pela razão entre as intensidades observadas para a amostra e o controle [20] e para a normalização dos dados, isto é, para que estes se aproximem de uma distribuição normal, é comum transformar-se esse valor pelo logaritmo.

Usualmente usa-se o corante Cy5 para a amostra e o Cy3 para o controle. Representando-se as intensidades medidas em cada condição como Int(Cy5) e Int(Cy3), o nível de expressão gênica é dado por:

$$Expressao\_Genica = \log_2 \frac{Int(Cy5)}{Int(Cy3)} \quad (1)$$

2.1.1 Seleção de genes discriminantes em dados de Microarranjos.

Os microarranjos de DNA são ferramentas importantes na busca de alterações fenotípicas e genotípicas ocorridas durante o progresso de diversas enfermidades.

Na pesquisa do câncer, é comum utilizarem-se microarranjos na classificação de padrões amostrais, como por exemplo, na discriminação entre células saudáveis de células malignas ou mesmo entre duas formas tumorais de difícil diagnóstico. Este tipo de problema pode ser inserido no ambiente de reconhecimento de padrões (ex. célula saudável x célula maligna, leucemia mielóide x leucemia linfóide etc.) e cada padrão é representado por um conjunto de atributos (genes). O objetivo central é selecionar os genes que discriminem o padrão de expressão amostral característico a cada uma dessas classes, uma vez que a maioria é irrelevante, apenas introduzindo ruído no processo de classificação e ainda podem diminuir a contribuição dos genes preditivos.

Tipicamente duas abordagens são utilizadas na redução de dimensionalidade em problemas de reconhecimento de padrões: extração de atributos e seleção de atributo [21]. Na extração de atributos novos atributos são criados baseados em

combinação ou transformação dos atributos do conjunto original. No método de seleção de atributos é selecionado o melhor subconjunto de atributos do conjunto original.

Particularmente, métodos de seleção de atributos supervisionados, isto é aqueles em que são conhecidos os rótulos de classes associados às amostras pesquisadas, são agrupados em duas abordagens: métodos envelopados e métodos de filtragem [4].

Nos métodos envelopados, avaliam-se subconjuntos de atributos até que se encontre o mais adequado, o que é computacionalmente custoso. Para a avaliação são utilizados algoritmos de aprendizagem de máquinas. Assim, no método envelopado utiliza-se o algoritmo treinado, classificador ou função discriminante na seleção de atributos. No método de filtragem a seleção de variáveis independe do algoritmo treinado.

Os métodos envelopados buscam o melhor subconjunto de atributos através da avaliação de uma função de custo de uma máquina de aprendizagem. Na prática, é necessário definir: (i) Como realizar a busca do melhor subconjunto no espaço de todos os possíveis subconjuntos de atributos; (ii) Como avaliar o desempenho de predição de uma máquina de aprendizagem para orientar a busca e os critérios de parada; e (iii) Qual preditor utilizar. Várias estratégias já bastante exploradas em outras aplicações têm sido utilizadas, como, por exemplo, *best first*, *branch-and-bound*, *simulated annealing* e algoritmos genéticos [4]. O desempenho é geralmente avaliado através de validação cruzada e os preditores usuais incluem árvores de decisão, Naïve Bayes, máquinas de vetores de suporte e mínimos quadrados linear [22].

Nos métodos de filtragem supõe-se que se tem uma medida de avaliação de cada atributo (por exemplo, informação mútua e teste qui-quadrado), que determinará sua relevância em relação à classe alvo. Uma hierarquia dos atributos é estabelecida baseada na relevância de predição com a classe alvo separadamente, o que resulta em um custo computacional menor. A partir dessa hierarquia, decide-se quantos atributos serão eliminados.

O método de filtragem seleciona subconjuntos de atributos de forma independente do classificador escolhido. Esta metodologia, também conhecida como Abordagem de Ordenação Individual de Genes, consiste em se medir para cada gene (atributo) a associação com a classe resposta segundo algum critério de medida (distância Euclidiana, índice de correlação de Pearson ou informação mútua) e em seguida, ordená-los seguindo a ordem decrescente do valor de associação individual de cada gene com as distribuições de classe. Dessa forma, a associação entre os

perfis gênicos não é considerada e a mesma informação pode ser introduzida por diversos genes que contribuem com a mesma informação em relação à classe de saída, o que aumentaria a redundância do sistema. Por outro lado, genes que individualmente tenham pouca associação com a classe de saída, mas que sejam complementares entre si podem não ser selecionados. Por isso, é necessário conhecer não somente a relevância dos genes em relação às classes consideradas, mas também a redundância entre os genes, uma vez que conjuntos que apresentam variáveis altamente redundantes têm seu poder de predição reduzido.

Um exemplo de método de seleção de atributo aplicado a microarranjos de DNA foi o apresentado por *Golub et al*. [6], pioneiro no estudo do conjunto de dados conhecido como Leucemia. Seu método verificou quais padrões de expressão gênica eram fortemente correlacionados com as duas classes de leucemia presentes no conjunto de dados. Para determinar quando essa correlação era maior do que o esperado ao acaso, *Golub et al.* [6] desenvolveram o método chamado "Análise de vizinhança" que, definido um padrão de expressão ideal (correspondente a um gene com expressão uniformemente alta em uma classe e uniformemente baixa em outra), busca por outros genes próximos ao padrão idealizado, verificando a existência de distribuições de genes não usuais em torno deste padrão.

Com esses resultados, Golub *et al.* [6] construíram um preditor, utilizando os genes selecionados na primeira etapa, capaz de classificar novas observações em uma das duas classes de leucemia. A partir dos genes informativos selecionados através da correlação com a classe de saída e baseando-se no nível de expressão desses genes, uma nova observação pode ser classificada. A predição de uma nova amostra é baseada em votos ponderados (figura 3). Cada gene informativo corresponde a um voto para a classe AML (leucemia mielóide aguda) ou ALL (leucemia linfóide aguda), dependendo se seu nível de expressão x_i está mais próximo do nível de expressão médio da classe AML ou ALL (μ_{AML} ou μ_{ALL}**)** num conjunto de observações de referência. A magnitude de cada voto é $\mathbf{w_i v_i}$, onde w_i é o fator de peso que reflete o quanto este gene é correlacionado com a classe e $v_i = x_i - (\mu_{AML} + \mu_{ALL})/2$ reflete o desvio do nível de expressão na observação da média μ_{AML} ou μ_{ALL}.

O total de votos para a classe 1 é obtido somando-se os valores absolutos de todos os votos positivos, enquanto que o total de votos para classe 2 é obtido somando-se os valores absolutos de todos os votos negativos. Para que uma nova observação seja incluída em determinada classe, Golub *et al.* [6] apresentaram o conceito de "prediction strength" (PS) ou medida da margem de vitória, cujo intervalo de definição está entre 0 e 1 e é definido por $PS = (V_{Venc} - V_{Perd}) / (V_{Venc} + V_{Perd})$, onde V_{Venc} e V_{Perd} representam os totais de votos para a classe vencedora e perdedora,

respectivamente. Finalmente, uma nova observação é associada à classe vencedora se o valor PS excede um limiar predeterminado, que depende do número de genes no preditor. Caso contrário, sua classificação é considerada como incerta. O limiar utilizado por Golub *et al.* [6] foi de 0.3.

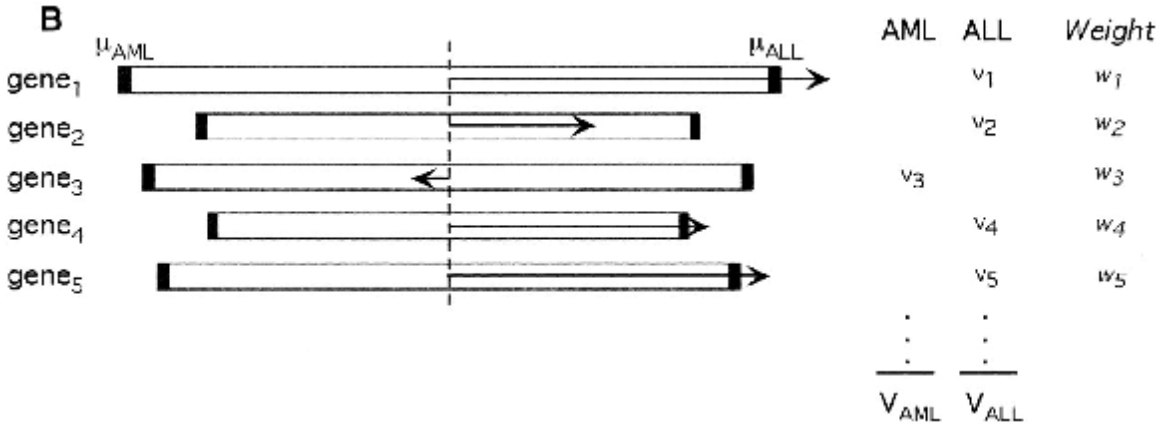

Figura 3: Predição de classe. A predição da classe de uma nova observação é baseada em "votos ponderados" de um conjunto de genes informativos. Golub

Uma vez classificadas as amostras, Golub *et al.* [6] utilizaram validação cruzada no conjunto de treinamento e acurácia no conjunto de teste para avaliar a acurácia do método de classificação proposto. Para a construção do preditor foram selecionados 50 genes que classificaram corretamente 36 das 38 observações e 2 foram classificadas como incertas (PS < 0,3).

O estudo desenvolvido por Golub *et al.* [6] motivou diversos pesquisadores a utilizarem este conjunto de dados. De uma forma geral, os genes discriminantes encontrados por cada pesquisador, não coincidem com o encontrado por Golub *et al.* [6]

A tabela 2 exemplifica alguns autores que estudaram o conjunto de dados de leucemia e a metodologia empregada.

Na tabela 3 são exibidos os genes mais relevantes na classificação da Leucemia encontrados pelos autores descritos na tabela 2. Nota-se que há pouca interseção entre os conjuntos de genes selecionados.

Tabela 2: Alguns estudos realizados e suas metodologias

Autor	Metodologia
Golub *et al.* (1999) [6]	Análise de Vizinhança
Marohnic, Debeljak e Bogunovié (2004) [7]	Mifs/SVM híbrido
Cho e Won (2003) [1]	Distância Euclidiana, Coeficiente de Correlação de Pearson e Coeficiente de Cosine
Huerta, Duval, Hão (2006) [23]	Lógica Fuzzy
Li, Zhu e Ruan (2004) [12]	SVM com núcleo linear

Tabela 3: ID dos genes discriminantes encontrados.

ID dos genes mais relevantes (Ordem crescente de ID)			
Golub	Marohnic	Huerta	Li
1249	804	235	1779
1630	1574	461	1796
1704	1779	804	1834
1745	4847	1144	1882
1834	4951	1260	2242
1882		1615	
1909		1745	
1928		1829	
2020		1834	
2043		1882	
2111		2020	
2121		2111	
2186		2121	
2242		2354	
2288		2363	
2348		2642	
2354		3252	
2402		3320	
2441		4196	

ID dos genes mais relevantes (Ordem crescente de ID)			
2642		4229	
2909		4328	
3056		4366	
3096		4847	
3258		4951	
3320		6005	
3847		6041	
4052		6185	
4177		6225	
4196		6281	
4328		6855	
4389			
4535			
4847			
4973			
5093			
5191			
5254			
5501			
5593			
5772			
6200			
6201			
6281			
6373			
6376			
6515			
6539			
6803			
6855			
7119			

Suspeita-se que a observação #66 tenha sido, em algum momento, rotulada erroneamente como AML, pois nem Golub *et al.* [6], nem os participantes do "*Critical Assessment of Techiniques for Microarray Data Mining"* (CAMDA'00, Dezembro 2000), uma conferência analisando o banco de dados de Leucemia, foram capazes de acertar sua classificação.

Nota-se também que os diferentes métodos encontram diferentes conjuntos de genes discriminantes, podendo acontecer de genes descritos na literatura como informativos não constarem na lista dos selecionados pelo preditor. Por exemplo, o gene Zyxin (4847) considerado informativo por Golub *et al.* [6] não figura em todos os conjuntos de genes mais relevantes apresentados na tabela 3.

2. 3 informação mútua

Nesta seção é apresentado um resumo da teoria da informação mútua, com descrição dos aspectos matemáticos envolvidos e alguns métodos que a utilizam no processo de seleção de atributos. O objetivo principal é ressaltar a dificuldade inerente à sua aplicação no conjunto de dados selecionado para o presente estudo, bem como sua solução.

A Teoria da Informação é uma formalização matemática do conceito de informação contida em mensagens. Se uma mensagem é perfeitamente conhecida *a priori*, a informação nela contida é nula. Entretanto, quanto menos previsível é uma mensagem, maior a quantidade de informação nela contida [9].

Um conceito primordial na compreensão da Teoria da informação é o de entropia. O conceito de entropia foi introduzido por SHANNON (1948), sendo o mais conhecida e empregado aquele que recebeu o seu nome. A entropia de Shannon é um funcional de funções de massa de probabilidade, p(x), sendo usualmente descrita como:

$$H_S(X) = H(p(x_1), p(x_2), \ldots, p(x_n)), \qquad (2)$$

onde, x é uma variável discreta aleatória de X, em um conjunto discreto X = {x_1, ... , x_n}, com função de massa de probabilidade p(x) = Pr(X = x). A entropia de Shannon de X, H(X), é definida como:

$$H_S(X) = -\sum_{x \in X} p(x) \log_b p(x) = \sum_{x \in X} p(x) \log_b \frac{1}{p(x)}, \quad (3)$$

onde, a base *b* do logaritmo é usualmente 2 e, neste caso, a entropia é medida em "bits". Se o logaritmo estiver na base neperiana, a unidade de H(X) é "nats". Pode-se escrever a entropia de Shannon como uma esperança matemática:

$$H_S(X) = E[-\log_b p(x)], \quad (4)$$

onde, o valor $-\log_b p(x)$ é interpretado como a informação contida na variável discreta $x \in X$ e é também chamada de informação de Hartley de x [9].

A entropia de Shannon representa a quantidade de informação contida numa variável aleatória x, isto é, o valor de x com a sua incerteza removida. A entropia de Shannon satisfaz às seguintes propriedades:

(1) Para qualquer X, $H(p(x_1),...,p(x_n))$ é uma função contínua e simétrica das variáveis $x_1,..., x_n$;
(2) Evento de probabilidade zero não contribui para a entropia, isto é, $H(p(x_1),...,p(x_n), 0) = H(p(x_1),...,p(x_n))$; e
(3) A entropia é maximizada quando a distribuição de probabilidade p(x) é uniforme, ou seja, para todo n, tem-se:

$$H_S(p(x_1),..., p(x_n)) \leq H\left(\frac{1}{p(x_1)},...,\frac{1}{p(x_n)}\right). \quad (5)$$

Isto é devido à inequação de Jensen [9]:

$$H(X) = E\left[\log_b \frac{1}{p(x)}\right] \leq \log_b\left(E\left[\frac{1}{p(x)}\right]\right) = \log_b n. \quad (6)$$

Renyi (1976) propôs uma definição mais geral de entropia na qual a entropia de Shannon (3) aparece como um caso particular [9]. Para uma variável aleatória discreta a entropia de Renyi é dada por:

$$H_{Ra}(X) = \frac{1}{1-a} \log \sum p(x)^a , \tag{7}$$

onde, a é um parâmetro livre, denominado ordem e *p(x)*, assim como em (3) é uma função de massa de probabilidade. Demonstra-se que quando $a \to 1$, $\lim_{a \to 1} H_{Ra} = H_S$, a entropia de Renyi (7) tende para entropia de Shannon (3) [9, 24].

Estendendo-se o conceito de entropia para conjuntos discretos diferentes e considerando que x e y são variáveis aleatórias dos conjuntos X e Y, respectivamente, a entropia associada a X e Y é dada por:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_b p(x,y), \tag{8}$$

onde, p(x,y) indica a função de massa de probabilidade conjunta associada aos conjuntos X e Y. Note que a entropia de X não depende dos valores atuais de X, apenas de p(x). Ou equivalentemente,

$$H(X,Y) = -E[\log_b p(x,y)]. \tag{9}$$

A entropia conjunta de n variáveis aleatórias é definida de forma análoga.

Outro conceito importante na compreensão da informação mútua é o de entropia relativa, D(p||q), divergência de Kulleback-Leibler [9, 25], ou ainda, divergente de p em relação a q. A entropia relativa mede a ineficiência de assumir que uma função de massa de probabilidade é q(.) quando a função verdadeira é p(.). Então, sejam p(.) e q(.) duas funções de massa de probabilidade de X e Y, respectivamente, onde $X \subset Y$, pode-se medir a entropia relativa entre as duas distribuições de massa de probabilidade como:

$$D(p \,||\, q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}. \tag{10}$$

Apesar de D(p||q) ser geralmente chamada de distância, esta não é verdadeiramente uma métrica, pois não é simétrica e não satisfaz a propriedade triangular. Entretanto, tem-se $D(p||q) \geq 0$, com a igualdade ocorrendo quando p(x) = q(x).

Estendendo-se o conceito de entropia, agora para conjuntos discretos não-independentes e considerando-se que variáveis discretas x e y, aleatoriamente amostradas dos conjuntos X e Y, respectivamente, tem distribuição p(x,y); então a entropia condicional H(Y|X) é definida como:

$$H(Y \mid X) = \sum_{x \in X} \sum_{y \in Y} p(x) H(Y \mid X = x) \tag{11}$$

$$H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x) \log_b p(y \mid x) \tag{12}$$

$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_b p(y \mid x), \tag{13}$$

ou, equivalentemente,

$$H(Y \mid X) = -E[\log_b \ p(y \mid x)] \tag{14}$$

Uma vez definida a entropia de Shannon para variáveis discretas, tanto assumindo dependência como a independência, pode-se agora definir a entropia de Shannon no caso contínuo, então chamada de entropia Diferencial, como:

$$H_S(x) = \int_{-\infty}^{+\infty} f(x) \log_b \ f(x) dx, \tag{15}$$

onde, *x* é uma variável contínua de função de densidade de probabilidade *f(x)*. A integral da equação (15) pode ser aproximada, por Riemann, como:

$$\int_{-\infty}^{\infty} f(x) dx = \lim \sum_{i=-\infty}^{+\infty} f(\mathrm{x_i}) \Delta, \tag{16}$$

onde, x_i é um valor discreto de *x* (figura 5), obtido pelo teorema do valor médio, em uma janela de tamanho Δ da função de densidade de probabilidade *f(x)*.

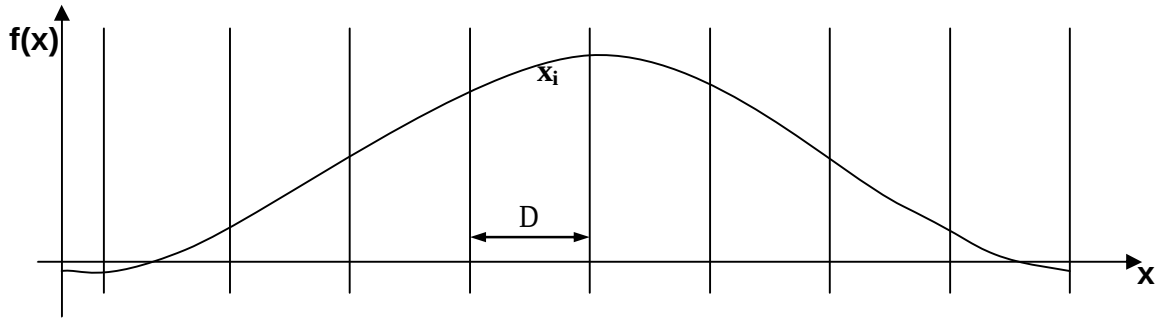

Figura 4: Discretização da função f em janelas de largura D

Assim, como para variáveis discretas, a entropia de Renyi (7) pode ser estendida para o caso contínuo, tendo a seguinte forma:

$$H_{Ra}(X) = \frac{1}{1-a}\log\left(\int_{-\infty}^{+\infty} p(x)^a\, dx\right), a \neq 1, \qquad (17)$$

onde, $p(x)^a$ é uma função de densidade de probabilidade. A entropia de Shannon é um caso particular da entropia de Renyi, onde $a \to 1$, isto é, $\lim_{a\to 1} H_{Ra} = H_S$. Um caso particularmente interessante é quando a =2, chamado de entropia quadrática de Renyi.

A entropia Relativa também pode ser escrita para o caso contínuo. Para duas funções de densidade de probabilidade f(.) e g(.), tem-se:

$$D(p \,\|\, q) := \int_S f \log \frac{f}{g}, \qquad (18)$$

onde, S é o suporte de f(.).

Uma vez definidos os conceitos de entropia (Shannon e Renyi) e divergência de Kulleback-Leibler, ou entropia relativa, para os casos discretos e contínuos, pode-se então definir a informação mútua I[X,Y], utilizando-se da entropia de Shannon (3), para dois conjuntos discretos X e Y, como:

$$I[X;Y] = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right), \tag{19}$$

ou na forma de Esperança, como:

$$I[X,Y] = E\left[\log \frac{p(x,y)}{p(x)p(y)} \right], \tag{20}$$

ou ainda, usando-se o divergente de Kulleback-Leibler D(p||q) = $\sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$ para casos discretos:

$$I[X;Y] = D(p(x,y) \,||\, p(x)p(y)). \tag{20}$$

Um número considerável de expressões relaciona informação mútua com a entropia de Shannon:

$$\begin{aligned}
&0 \le I[X;Y] \le H[X]; \\
&I[X;Y] = H[X] - H[X \mid Y]; \\
&I[X;Y] = H[X] + H[Y] - H[X,Y]; \\
&I[X;X] = H[X].
\end{aligned}$$

lembrando que a entropia H(X) quantifica a incerteza sobre X. A última linha justifica a descrição de entropia como auto-informação. Essas relações são ilustradas a seguir (figura 6), onde H(X) = H(X/Y) U I(X,Y); H(Y) = H(Y/X) U I(X,Y); H(X,Y) = H(X/Y) U H(Y/X).

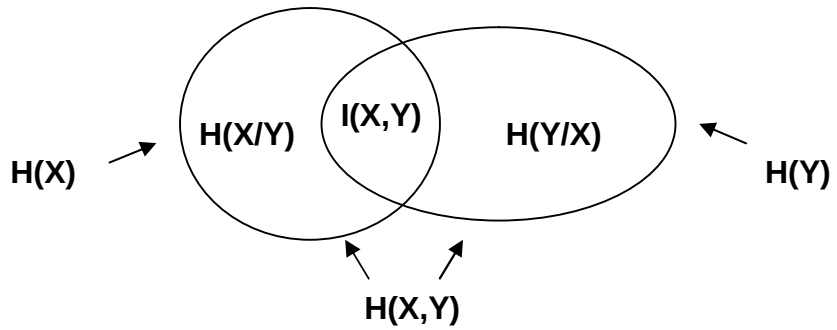

Figura 5: Relação entre entropia e informação mútua

A característica mais óbvia da informação mútua é que ela depende simultaneamente de X e Y. Não há informação no vácuo, pois informação é sempre a respeito de alguma coisa. Para o caso em que X = Y, I[X,Y] é a informação em X sobre X. Como o próprio nome sugere, a informação mútua é simétrica, isto é, I[X, Y] = I[Y,X]. Portanto, qualquer informação em X sobre Y é a mesma que em Y sobre X. Ao se definir I[X, Y] em termos de entropia relativa, esta pode ser interpretada como um tipo de "distância" entre a distribuição conjunta $p(x,y)$ e o produto das distribuições $p(x)p(y)$.

Para X e Y independentes, $p(x,y) = p(x)p(y)$, então a divergência de Kulleback-Leibler tende a zero e I[X,Y]=0, como esperado. Entretanto, quando não se assume independência entre as variáveis, pode-se, por exemplo, definir a informação mútua condicional entre as variáveis aleatórias X e Y dado Z, como:

$$I[X,Y \mid Z] = H(X \mid Z) - H(X \mid Y,Z), \text{ `} \tag{21}$$

ou na forma de esperança, como:

$$I[X,Y \mid Z] = E\left[\log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)}\right]. \tag{22}$$

Usando-se a medida de entropia de Renyi (17), a informação mútua (19) é escrita da seguinte maneira, no caso contínuo:

$$I_{Ra}[X,Y] = \frac{1}{a-1}\log \int_x \int_y \frac{p(x,y)^a}{(p(x)p(y))^{a-1}} dxdy, \tag{23}$$

e a informação mútua é maximizada quando a entropia relativa (divergência de Kulleback-Leibler) entre X e Y é minimizada.

Estimar a função densidade de probabilidade geralmente não é uma tarefa simples, sendo assim PRÍNCIPE *et al.* [9] desenvolveram um método de se calcular a entropia diretamente das amostras, utilizando-se a entropia quadrática de Renyi combinada com o estimador de função de densidade de probabilidade não-paramétrico de Parzen. Para o cálculo da informação mútua duas novas medidas de dependência entre funções de densidade de probabilidade baseadas na desigualdade

de Cauchy-Schwartz e distância Euclidiana são introduzidas. O estimador de Parzen para uma variável x é definido como:

$$\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} k(x - x_i), \tag{24}$$

onde, N é o número de amostras da variável aleatória x, x_i são as observações da variável aleatória, e k é uma função de núcleo (*kernel*) Gaussiano, descrito por:

$$k(x) = G(x, s^2 I) = \frac{1}{(2ps^2)^{\frac{M}{2}}} e^{-\frac{x^T x}{(2s^2)}}. \tag{25}$$

Tem-se então que:

$$\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} G(x - x_i, s^2) \tag{26}$$

Dadas duas funções Gaussianas $G(x - a_i, \Sigma_1)$ e $G(x - a_j, \Sigma_2)$ com médias a_i e a_j, onde Σ_1 e Σ_2 são as matrizes de covariância, tais que $x, a_i, a_j \in \Re^m$ e Σ_1, $\Sigma_2 \in \Re^{mxm}$, pode-se concluir que:

$$\int_{-\infty}^{+\infty} G(x - a_i, \Sigma_1) G(x - a_j, \Sigma_2) dx = G(a_i - a_j, (\Sigma_1 + \Sigma_2)), \tag{27}$$

Assim, substituindo a estimativa da função densidade de probabilidade apresentada em (26) na equação (17) para a =2 e utilizando a relação (27) encontra-se o estimador da entropia quadrática de Renyi para a variável aleatória x com as observações x_i :

$$HR(X) = -\log V(X) \quad (28)$$

onde, V(X) = $\sum_{i=1}^{N}\sum_{j=1}^{N} G(x_i - x_j, 2s^2)$ é chamado de potencial de informação [9].

A relação (28) mostra que a combinação do estimador de função de densidade de probabilidade de Parzen combinado com a entropia quadrática de Renyi fornece um estimador de informação mútua baseado diretamente nas amostras.

Dessa forma, a informação mútua pode ser estimada com base nas relações com a entropia.

2.3.1 Seleção de Atributos via Informação Mútua

Nesta seção será descrita a evolução do algoritmo de seleção de atributos baseado na informação mútua (MIFS) utilizado como base no presente estudo até a sua forma atual, rebatizada de MIFS-U-Renyi.

Batiti [26] propôs um algoritmo para resolver o problema de seleção de variáveis através da maximização da informação mútua entre as variáveis de entrada e a classe de saída, isto é, minimizando a entropia condicional entre elas. Este algoritmo é baseado no chamado “greedy selection’’ e partindo de um conjunto vazio, os melhores atributos são adicionados um a um, como descrito a seguir:

1) Seja F o conjunto com n atributos e S um conjunto vazio.
2) $\forall f \in F$, calcula-se I(C;f), onde C é o conjunto das classes.
3) Encontra-se o atributo que maximiza I(C;f), subtraindo-o do conjunto F e adicionando-o ao conjunto S.
4) Repetem-se os passos anteriores até que o número de atributos desejado seja atingido.

i. $\forall f \in F$, calcula-se I(C;f;S)

ii. Seleciona-se o atributo $f \in F$ que maximiza I(C;f;S), subtraindo-o do conjunto F e adicionando-o ao conjunto S.

O conjunto S contém os atributos selecionados.

O problema reside em calcular I(C;f;S), a informação mútua entre o novo candidato a ser selecionado f e o vetor de atributos já selecionados e a classe C. A solução apresentada por Batiti [26], denominada MIFS (*Mutual Information Feature Selector*), usa apenas o cálculo I(C;f) e I(f;f'), onde f e f' são atributos individuais.

Assim, MIFS tem a mesma estrutura do "greedy selection" a menos do passo 4, como descrito a seguir:

4) Repetem-se os passos anteriores até que o número de atributos desejado seja atingido:

i. Para todo par (f,s) com $f \in F$ e $s \in S$, calcula-se I(f;s)

ii. Seleciona-se o atributo $f \in F$ que maximiza

$$I(C;f) - \beta \sum_{s \in S} I(f;s) \tag{29}$$

subtraindo-o do conjunto F e adicionando-o ao conjunto S.

O parâmetro β regula a importância relativa da informação mútua entre o candidato a atributo e os atributos já selecionados com respeito a informação mútua com a classe de saída. Se β é zero, apenas a informação mútua com a classe de saída é considerada para cada atributo. Conforme o valor de β aumenta há um desconto proporcional da IM total com respeito aos atributos já selecionados. Na prática, um valor de β entre 0.5 e 1 é apropriado para muitos problemas de classificação [26].

Kwak e Choi [10] propuseram um algoritmo de seleção de variáveis, também baseado na teoria da informação, bastante próximo ao "greedy selection" e ao MIFS, denominado *Mutual Information Feature Selector under Uniform Information Distribution* (MIFS-U). Em seu estudo, mostraram que estar condicionado por uma classe C não altera a razão de entropia de f_s e a informação mútua entre f_s e f_i, isto é:

$$\frac{H(f_s \mid C)}{H(f_s)} = \frac{I(f_s; f_i \mid C)}{I(f_s; f_i)}. \tag{30}$$

Consequentemente, $I(f_s;f_i|C)$ pode ser representada por:

$$I(f_s;f_i \mid C) = \frac{H(f_s \mid C)}{H(f_s)} I(f_s;f_i) \tag{31}$$

Usando a equação (31) e o fato de que a Informação mútua condicional pode ser escrita como:

$$I(C;f_i|f_s) = I(C;f_i) - \{I(f_s;f_i) - I(f_s;f_i|C)\} \tag{32}$$

Tem-se que:

$$\begin{aligned} I(C;f_i \mid f_s) &= I(C;f_i) - \left(1 - \frac{H(f_s \mid C)}{H(f_s)}\right) I(f_s;f_i) \\ &= I(C;f_i) - \frac{I(C;f_s)}{H(f_s)} I(f_s;f_i) \end{aligned} \tag{33}$$

Sendo assim, o passo 4 do algoritmo "greedy selection" pode ser reescrito como:

4. Repetem-se os passos anteriores até que o número de atributos desejado seja atingido

i. Para todo par (f,s) com $f \in F$ e $s \in S$, calcula-se I(f;s)

ii. Seleciona-se o atributo $f \in F$ que maximiza $I(C;f) - \beta \sum_{s \in S} \frac{I(C;s)}{H(s)} . I(f,s)$, subtraindo-o do conjunto F e adicionando-o ao conjunto S.

Outra forma de estimar a pdf apresentada por Kwak e Choi [10] é o método denominado de seleção de atributos através de janelas Parzen (PWFS). Neste, a estimativa das pdfs utlilizadas da entropia de Shannon, também é feita através de Janelas Parzen com núcleo gaussiano, como o próprio nome sugere. Assim, sabendo-se que a probabilidade condicional p(c|x) é

$$p(c \mid x) = \frac{p(c \mid x)}{\sum_{k=1}^{Nc} p(k \mid x)} = \frac{p(c)p(x \mid c)}{\sum_{k=1}^{Nc} p(k)p(x \mid k)} . \tag{34}$$

e segundo termo da igualdade é dado pela regra de Bayes, usando (26) obtém-se a estimativa da pdf

$$\hat{p}(c \mid x) = \frac{\sum_{i \in I_c} G(x - x_i, h_c)}{\sum_{k=1}^{N_c} \sum_{i \in I_k} G(x - x_i, h_k)}, \tag{35}$$

onde h_c e h_k são os parâmetros de largura da janela específico para cada classe.

Se a janela gaussiana for utilizada com os mesmos parâmetros de largura e a mesma matriz de covariância para cada classe, utilizando (25) a equação (35) pode ser reescrita como:

$$\hat{p}(c \mid x) = \frac{\sum_{i \in I_c} \exp\left(-\frac{(x - x_i)^T \Sigma^{-1} (x - x_i)}{2h^2} \right)}{\sum_{k=1}^{N} \sum_{i \in I_k}^{n} \exp\left(-\frac{(x - x_i)^T \Sigma^{-1} (x - x_i)}{2h^2} \right)} \tag{36}$$

onde, N é o número de classes, I_c é o conjunto de índices dos exemplos de treinamento pertencentes a classe c, Σ é a matriz de covariância e h é o parâmetro utilizado para dimensionar a janela.

A título de comparação, a matriz de covariância foi utilizada de duas formas diferentes; primeiro utilizando-se apenas os elementos da diagonal (PWFS – Tipo I) e depois utilizando todos os elementos da matriz (PWFS – Tipo II); chegando-se a conclusão que as performances não apresentavam diferenças consideráveis.

Para o cálculo da entropia condicional com n observações e supondo que cada uma destas tem a mesma probabilidade, obteve-se :

$$\hat{H}(C / X) = -\sum_{j=1}^{n} \frac{1}{n} \sum_{c=1}^{N} \hat{p}(c \mid x_j) \log \hat{p}(c \mid x_j), \tag{37}$$

onde, x_j é o j-ésimo registro do conjunto de treino, X é o conjunto de variáveis de entrada e C é o conjunto das classes.

A escolha correta do parâmetro h e da função janela é muito importante. Normalmente, são utilizadas janelas retangulares ou Gaussianas para função de janela Parzen, sendo que a função Gaussiana apresenta a vantagem de não possuir

descontinuidades. Para o parâmetro h, Kwak e Choi [10] o estimaram utilizando 1/log n, onde n é o tamanho da amostra. Esta escolha satisfaz às seguintes condições necessárias:

a) $\lim_{n\to\infty} h(n) = 0$

b) $\lim_{n\to\infty} nh^d(n) = \infty$, onde d é a dimensão dos vetores de entrada.

Em microarranjos de DNA a informação mútua é utilizada para ordenar os genes de acordo com a quantidade de informação simultânea entre os genes e a classe de saída e além disso, para minimizar a redundância entre os atributos.

2.4 Máquinas de Vetores de suporte

Nesta seção é definido o conceito de máquina de vetor de suporte, suas vantagens e limitações no ambiente de classificação.

O algoritmo de máquinas de vetores de suporte (SVM[♣]) foi desenvolvido, no início da década de 60, como uma ferramenta para classificação de padrões. A SVM se utiliza de um mapeamento do espaço de vetores de entrada x em um espaço de maior dimensão, através de um hiperplano de separação ótima Esta metodologia está exposta em detalhes em "Statistical Leaning Theory – Vapnik, 1998" e em "The Nature of Statistical Learning Theory, Springer", New York,1995" [27, 28] , e será apresentada aqui de forma resumida.

Em problemas de classificação dicotômicos, o objetivo é utilizar um conjunto de treinamento constituído por pares "entrada-saída" desejada $\{(x_1, y_1),...,(x_l,y_l)\}$ para construir uma função que discrimine as duas classes. Aqui, $x_k \in IR^n$ e y_k assumem os valores -1 ou + 1, dependendo da classe associada à observação x_k. Se o problema for linearmente separável, o objetivo das SVM é encontrar um hiperplano definido a partir de uma função $f:IR^n \rightarrow IR^n$ tal que sign(f(x)) seja capaz de predizer a classe de

[♣] SVM, como a ferramenta é mais conhecida, se refere a abreviatura da nomenclatura em língua inglesa 'Support Vector Machines'

uma observação x não utilizada na fase de treinamento. Caso o problema não seja linearmente separável a dimensão é aumentada até que esta condição se verifique.

SVMs buscam hiperplanos que estejam o mais distante possível de todas as observações de cada uma das classes a ser aprendida pelo algoritmo. Isto é feito encontrando-se aquele que maximiza uma magnitude chamada de margem, que é a distância entre hiperplano e os pontos mais próximos de cada classe (figura 07).

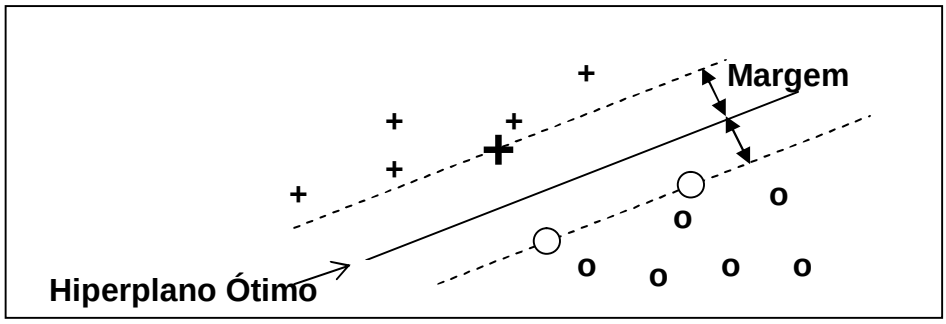

Figura 6: Hiperplano de separação ótima é o único que separa os dados com margem máxima.

Para que dois subconjuntos finitos de vetores x do conjunto de treinamento $(x_1, y_1),...,(x_l, y_l)$, $x \in \mathbb{R}^n$, $y \in \{-1,1\}$ sejam separáveis (em subconjunto I para y = 1 e subconjunto II para y = -1) é necessário que exista um vetor unitário w e uma constante b tais que as desigualdades abaixo sejam verificadas:

$$\begin{aligned} &< x_i . w > -b > 0 \text{ , se } x_i \in \text{ I} \\ &< x_i . w > -b < 0 \text{ , se } x_i \in \text{ II} \end{aligned} \qquad (38)$$

Para qualquer vetor unitário w, define-se:

$$\begin{aligned} c_1(w) &= min_{x_i \in I}(< x_i , w >) \\ c_2(w) &= max_{x_j \in II}(< x_j , w >) \end{aligned} \qquad (39)$$

Chamando de w_o, o vetor unitário que maximiza a função abaixo, sob a condição de que as inequações (34) sejam satisfeitas, temos:

$$\rho(w) = \frac{c_1(w) - c_2(w)}{2}, \quad |w| = 1 \tag{40}$$

O vetor w e a constante $c_o = \frac{c_1(w_o) - c_2(w_o)}{2}$ determinam o hiperplano que separa os vetores $x_1,...,x_\alpha$ do subconjunto I dos vetores $x_1,...,x_\beta$ do subconjunto II ($\alpha + \beta = l$) e tem margem máxima (37). Este hiperplano é chamado de hiperplano de margem máxima ou hiperplano ótimo. Demonstra-se que o hiperplano ótimo é único [27, 28].

Estes tipos de máquinas possuem duas camadas e durante o processo de aprendizagem são encontrados, na primeira camada, os pesos $w = [w_1,..., w_n]$ e o bias b enquanto que na segunda camada constrói-se a função de decisão ou função discriminante:

$$f_0 = \text{sign}(\rho(x_i))$$

Novos padrões desconhecidos são classificados de acordo com o sinal da função de decisão [2]:

$$\begin{aligned} f_0 > 0 &\Rightarrow x \in \text{ classe 1 } (y_1 = +1) \\ f_0 < 0 &\Rightarrow x \in \text{ classe 2 } (y_1 = -1) \\ f_0 = 0 &\Rightarrow \text{Margem de decisão} \end{aligned} \tag{41}$$

Geometricamente, a margem de decisão M a ser maximizada é a projeção no plano normal de pesos da distância entre dois vetores de suporte quaisquer. Pode também ser definida como a norma do vetor peso w. Assim sendo, considerando x_1 e x_2 dois pontos de diferentes classes tais que $wx_1 + b = 1$ e $wx_2 + b = -1$, então, a margem é dada pela distância perpendicular desses dois pontos ao hiperplano:

$$M = \frac{D_1 - D_2}{2}.$$

Escrevendo:

$$D_1 = ||x_1|| \cos\alpha;$$

$$D_2 = ||x_2|| \cos\beta;$$

$$\cos\alpha = \frac{x_1^T w}{\| x_1 \| \| w \|};$$

$$\cos\beta = \frac{x_2^T w}{\| x_2 \| \| w \|}.$$

Verifica-se que:

$$M = \frac{x_1^T w - x_2^T w}{2. \| w \|}$$

Substituindo $w^T x_1 = 1 - b$ e $w^T x_2 = -1 - b$, obtém-se:

$$M = \frac{1}{\| w \|} \tag{42}$$

Dessa forma, a máquina de aprendizado encontra a margem máxima através da minimização de $||w||^2$, sob a condição:

$$yi\ |w^T x_i + b| \geq 1 \tag{43}$$

O vetor ψ_o com a menor norma satisfazendo a condição (39) define o hiperplano ótimo. Se b = 0, o hiperplano ótimo passa pela origem.

A margem ρ_o entre o hiperplano ótimo e os vetores é igual a:

$$\rho(w_o) = sup\frac{1}{2}\left(\underset{i\in I}{min}(< x_i, w_o >) - \underset{j\in II}{max}(< x_j, w_o >)\right) = \frac{1}{\psi_o} \tag{44}$$

Minimizar $||w||^2 = <w,w>$ sob a condição (39) consiste de um problema de otimização quadrática e Vapnik [27, 28] sugere que ao invés de encontrar solução no espaço primal, esses resultados sejam obtidos no espaço dual (Espaço dos multiplicadores de Lagrange)

Assim, deve-se encontrar o ponto estacionário (ponto de sela) da função de Lagrange:

$$L(\psi,b,a) = \frac{1}{2} < \psi,\psi > -\sum_{i=1}^{l} \alpha_i \left(y_i \left[< x_i,\psi > +b \right] - 1 \right) \qquad (45)$$

onde $\alpha_i \geq 0$ são os multiplicadores de Lagrange.

Para isto, deve-se minimizar a função em relação a ψ e b e maximizá-la em relação aos multiplicadores de Lagrange $\alpha_i \geq 0$.

Sabe-se que, segundo o Teorema de Fermat [27, 28], os pontos de mínimo satisfazem as seguintes condições:

$$\begin{aligned} &(a)\ \frac{\partial L(\psi,b,\alpha)}{\partial \psi} = \psi - \sum_{i=1}^{l} y_i \alpha_i x_i = 0 \\ &(b)\ \frac{\partial L(\psi,b,\alpha)}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0 \end{aligned} \qquad (46)$$

De acordo com as condições acima, para o vetor ψ que define o hiperplano ótimo, as seguintes igualdades são verdadeiras:

$$\begin{aligned} &\psi - \sum_{i=1}^{l} y_i \alpha_i x_i = 0 \\ &\sum_{i=1}^{l} y_i \alpha_i = 0 \end{aligned} \qquad (47)$$

Chamando $L(\psi, b, a)$ de $W(\alpha)$ para enfatizar a transformação do espaço primal para o espaço dual, a equação (45) pode ser reescrita como:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j < x_i, x_j >, \qquad \alpha_i \geq 0 \text{ e } i = 1,...,l \qquad (48)$$

Dessa forma, o problema dual a ser resolvido é, considerando um conjunto de treinamento $\{(x_1, y_1),...,(x_l,y_l)\}$, determinar os multiplicadores de Lagrange ótimos α_i para que W seja maximizado e que satisfaçam (46).

As soluções ótimas α^*, b^* e ψ^* devem satisfazer a igualdade:

$$\alpha^* \left[y_j \left(\left(\psi^* . x_j \right) + b^* \right) - 1 \right] = 0 \text{, para j = 1,..., N} \qquad (49)$$

Assim, pela equação, vê-se que os pontos em que $y_j \left(\left(\psi . x_j \right) + b \right) - 1 \neq 1$ deve necessariamente ter $\alpha = 0$. Só os pontos de margem 1 podem ter os correspondentes $\alpha \neq 0$. Esses pontos são os vetores de suporte.

O hiperplano ótimo definido pela margem de separação ótima ao se resolver o problema de otimização do dual, é expresso somente em termos deste conjunto de vetores suporte e é descrito pela função de decisão não-linear abaixo:

$$f\left(x, \alpha, b\right) = sgn \mid \sum_{i=1}^{N_{SV}} \alpha_i y_i x_i^T x + b \mid, \qquad (50)$$

onde N_{SV} representa o número de vetores suporte, que pode ser bem menor que o número de dados de treinamento. Tanto o hiperplano ótimo definido pela equação (47) quanto a função objetivo do problema de otimização definida pela equação (48) não dependem da dimensão do vetor x, mas sim do produto interno entre dois vetores e isto permite que hiperplanos de separação sejam construídos mesmo em espaços de grande dimensão (no espaço de dimensão infinita de Hilbert). Além disso, apenas esses vetores são importantes na predição da classe de um novo ponto.

Separar classes por meio de hiperplanos só é possível para problemas linearmente separáveis, mas utilizando-se funções matemáticas, pode-se mapear (Φ) o espaço das entradas, projetando-o em um novo espaço, de maior dimensão, chamado espaço de atributos (*F*), no qual os dados provavelmente serão separáveis.

Pode-se provar que, para qualquer conjunto de dados com rótulos de classe consistentes (onde consistente significa que o conjunto de dados não contém dois objetos idênticos com rótulos opostos) existe uma função de núcleo que permitirá uma separação linear dos dados [29]. Este novo espaço pode ter mais dimensões que o espaço original, o que é bastante útil na classificação de dados não separáveis, mas aumenta o risco de especialização (*overfitting*) [2].

Assim,

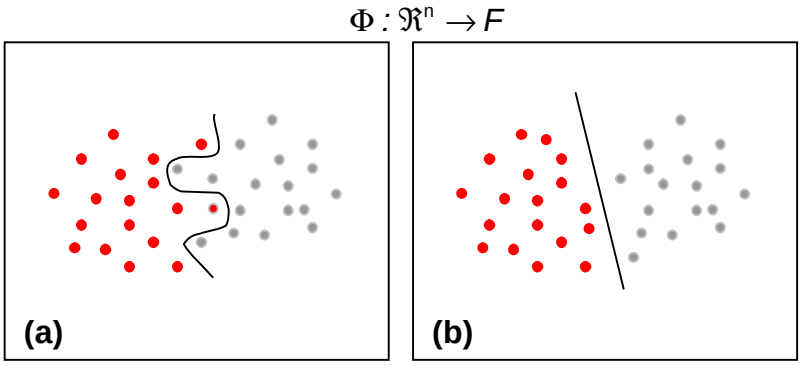

Figura 7: (a) Espaço das entradas não linearmente separáveis; (b) Espaço dos atributos.

O algoritmo de treinamento dependerá apenas dos dados e de seus produtos internos em *F*, isto é, de funções da forma $\Phi(x_i).\Phi(x_j)$. Definindo-se uma função (positiva-semidefinida) de núcleo K tal que para todo $x_i,x_j \in$ X, tem-se $K(x_i,x_j)$ igual ao produto interno entre as funções $\Phi(x_i)$ e $\Phi(x_j)$, isto é, $K(x_i,x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$, não será necessário explicitar ou mesmo conhecer o mapeamento Φ. O algoritmo de treinamento utilizará apenas a função k [30].

Há restrições quanto às funções de núcleo admissíveis. As funções de núcleo devem ser contínuas e simétricas e atender ao Teorema de Mercer [27, 28] que estabelece a condição suficiente para que uma função contínua e simétrica K(x,y) seja um núcleo.

Teorema (Mercer): Existe um mapeamento Φ e uma expansão $K(x,y) = \sum_i \Phi(x)_i \Phi(y)_i$ se e somente se, para qualquer g(x) tal que $\int g(x)^2 dx$ é finita, então:

$$\int k(x,y)g(x)g(y)dxdy \geq 0 \qquad (51)$$

Mostra-se que esta condição é satisfeita para integrais positiva das potências do produto escalar $K(x,y) = (x.y)^p$ [30].

As funções de núcleo usuais são [31]:

Tabela 4: Funções de Núcleo ususais

01. Função Linear	$K(x,y)=\langle x,y\rangle$
01. Função de Base Radial (RBF)	$K(x,y)=exp\left(\frac{-\|x-y\|^2}{c}\right)$
02. Polinomial	$K(x,y)=\left((x.y)+\theta\right)^d$
03. Sigmoidal	$K(x,y)=tanh\left(\kappa(x.y)+\theta\right)$
04. Multiquadrática inversa	$K(x,y)=\frac{1}{\sqrt{\|x-y\|^2+c^2}}$

Pode-se ainda criar núcleos mais complexos a partir de outros mais simples.

Para classificadores SVM não-lineares, uma vez que $\langle x_i,x_j\rangle$ é substituído pelo mapeamento $\langle\Phi(x_i),\Phi(x_j)\rangle$ e este pode ter dimensão infinita, o problema quadrático dual é, então, reescrito utilizando-se o núcleo de produto interno $K(x_i,x_j) = \langle\Phi(x_i),\Phi(x_j)\rangle$ da seguinte maneira:

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_i y_j\alpha_i\alpha_j K<x_i,x_j>$$

$$\text{para } \sum_{i=1}^{l}y_i\alpha_i=0,\ \alpha_i\geq 0 \text{ e } i=1,...,l \qquad (52)$$

Finalmente, o classificador SVM não-linear assume a forma

$$f(x,\alpha,b)=sgn\left[\sum_{i=1}^{N_{SV}}\alpha_i y_i k(x,x_i)+b\right], \qquad (53)$$

com as constantes positivas reais α_i sendo a solução do problema de otimização convexa.

Uma vez determinados os vetores de suporte, pode-se calcular o limite superior do valor esperado de se cometer um erro de classificação em uma amostra independente.

$$E_l[P(\text{erro})] \leq \frac{E[N_{VS}]}{l} \tag{54}$$

É importante notar que o limite independe da dimensão do espaço de entrada, e por isso, SVMs com pequeno número de vetores suporte apresentarão boa capacidade de generalização mesmo em espaços de entrada de alta dimensão.

Para verificar a qualidade das estimativas vindas dos dados é necessário definir um critério. A diferença entre o valor dado por $f(x_i)$ e o valor y_i do arranjo dos dados de treinamento é chamado resíduo de saída e é uma indicação da precisão do ajuste no ponto x_i.

Para dados de treinamento que apresentam ruídos são estabelecidas medidas apropriadas (funções de erro, perda ou custo) que penalizem tais singularidades. Essas funções determinam a importância da precisão. No caso de classificação binária, será do tipo:

$$\begin{aligned} L(y, f_o(x,y)) &= 0, \text{ se } f_o(x,y) = y \\ L(y, f_o(x,y)) &\neq 0, \text{ se } f_o(x,y) \neq y \end{aligned} \tag{55}$$

Em regressão, as funções de erro mais usuais são: o erro quadrático $\left(L(y, f_o(x,y)) = (y - f_o)^2\right)$; e o erro absoluto $\left(L\left(y, f_o(x,y)\right) = | y - f_o |\right)$ O objetivo é encontrar a função f que minimiza o valor esperado da perda, fornecido pelo risco funcional:

$$R[f] = E[\left(y - f(x)\right)^2] = \int L(y, f_o(x,y))P(x,y)DxDy\,, \tag{56}$$

que mede o erro para todos os padrões de entrada-saída que são gerados por uma função caracterizada pela distribuição de probabilidade P(x,y).

No caso de a distribuição P(x,y) ser desconhecida deve-se utilizar apenas os dados de treinamento para estimar a função f_o, aproximando-a daquela que minimiza o risco esperado R[f].

Uma aproximação possível para o risco funcional é o chamado Risco Funcional Empírico:

$$R_{emp} = \frac{1}{l}\sum_{i=1}^{l}\left(y_i - f(x_i, w)\right)^2 \qquad (57)$$

Vapnik [27, 28] sugere dois princípios para minimização do risco empírico diretamente das amostras: O primeiro utiliza o conceito de minimização de risco empírico ("Empirical Risk Minimization" - ERM). Pela lei dos grandes números, se o conjunto de treinamento tende para infinito então o risco empírico converge para o risco esperado, i.e.,

$$\lim_{l\to\infty} (\text{Rreal}[f] - \text{Remp}[f]) = 0. \qquad (58)$$

Assintoticamente, o risco empírico converge para o risco esperado, mas para amostras pequenas, grandes desvios podem ocorrer aumentando o risco de especialização (*overfitting*)

O segundo é um refinamento do primeiro princípio. O ERM geralmente apresenta bons resultados com amostras grandes, entretanto em pequenas amostras, onde a razão entre o número de observações e a dimensão VC do conjunto de funções da máquina de aprendizado, que será definida abaixo, um valor baixo do risco empírico não garante um risco esperado pequeno. Este princípio, chamado de minimização do risco estrutural ("Structural Risk Minimization" - SRM) estabelece que para minimizar o risco esperado deve-se procurar a relação ótima entre o número de dados amostrais, a qualidade da aproximação dos dados pela função (máquina de aprendizado) escolhida dentre um conjunto de funções com diferentes graus de liberdade e o valor que caracteriza a capacidade do conjunto de funções (dimensão VC).

Para compreender o conceito de dimensão VC, considere apenas as funções que correspondem aos casos de reconhecimento de padrões com duas classes. Um dado conjunto com l pontos possui 2^l possibilidades de rótulos de classe e para cada possibilidade, um elemento do conjunto de funções fará a distribuição correta dos rótulos. Diz-se que este conjunto de pontos pode ser separado por esse conjunto de funções [30].

SRM encontra um conjunto de funções que, para uma amostra de tamanho fixo, minimiza simultaneamente a soma do risco empírico com o intervalo de confiança

VC referente a ele. Isto pode ser feito treinando-se uma série de máquinas, uma para cada subconjunto, onde para cada subconjunto dado o objetivo é minimizar o risco empírico.

Vapnik [27, 28] mostra que para qualquer vetor de parâmetros σ do conjunto de funções $f(x;\sigma)$, $\sigma \in \Lambda$, para $k<l$ e para dados identicamente distribuídos (i.i.d.), o limite:

$$R(\sigma) \leq R_{emp}(\sigma) + \sqrt{\frac{h(ln(l/h)+1) - ln(\eta/4)}{l}} \qquad (59)$$

é verdadeiro, com probabilidade $(1 - \eta)$ e o segundo termo é chamado *termo de confiança* ou *intervalo de confiança VC,* que depende da dimensão VC e de h (que caracteriza a capacidade do conjunto de funções)

Observando a equação (59), fazendo com que o número de observações do conjunto de treino tenda a infinito ($l \to \infty$), então o intervalo de confiança VC tende para zero, fazendo com que o risco esperado $R(\alpha)$ tenda para o risco empírico $R_{emp}(\alpha)$. Sendo assim, se existe um número limitado de dados de treinamento, então o intervalo de confiança VC arbitrado não deve ser alto.

A maior limitação da SVM provavelmente reside na escolha do núcleo, uma vez que se deseja uma função que seja capaz de separar adequadamente os dados sem introduzir muitas dimensões irrelevantes e não existe regra para determiná-la.

Diferentemente das redes neurais, o treinamento de uma SVM encontra o mínimo global, uma vez que treinar uma SVM significa resolver um problema de otimização quadrática convexa [30].

2.1 Câncer e Leucemia

Esta seção visa apresentar brevemente os mecanismos de desenvolvimento das leucemias, mais especificamente, da leucemia mielóide aguda e da leucemia linfóide aguda, as estruturas envolvidas e a importância do diagnóstico preciso e precoce para o sucesso terapêutico, sem o objetivo de realizar um estudo profundo, mas permitir a familiarização do leitor de diversas áreas com a relevância em se desenvolver ferramentas precisas de classificação, especificamente para as duas formas predominantes de leucemia aguda, a leucemia mielóide aguda e leucemia linfóide aguda e a compreensão de que tal ferramenta poderá ser utilizada para classificação de outras malignidades.

Câncer é o nome popularmente usado para designar um conjunto de mais de 100 doenças que têm em comum o crescimento rápido e desordenado de células clonais [32]. Mutações em determinados genes alteram os comandos de divisão, diferenciação e morte celular, permitindo sua multiplicação desenfreada. Com seus mecanismos de controle da divisão inoperantes passam a se multiplicar independentemente das necessidades do organismo.

Por meio de sucessivas divisões, a célula, agora chamada de maligna, acaba formando um agrupamento de células praticamente idênticas que recebe o nome de tumor. Diante dessa perda de controle intrínseca da multiplicação celular, só resta ao organismo tentar identificar e destruir essas células anormais por intermédio do seu Sistema Imunológico. Se esse sistema mostrar-se ineficaz, a doença passará a ter condições de evoluir, podendo espalhar-se (metástase) para outras regiões do corpo.

Dividindo-se rapidamente, estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de tumores (acúmulo de células cancerosas) ou neoplasias malignas. Outras características que diferenciam os diversos tipos de câncer entre si são a velocidade de multiplicação das células e a capacidade de invadir tecidos e órgãos vizinhos ou distantes (metástases).

A leucemia é uma neoplasia originária da transformação maligna das células hematopoiéticas (formadoras de sangue), mais especificamente, dos leucócitos, que são células especializadas na defesa do organismo (figura 8). O mecanismo neoplásico desses tipos celulares é mal compreendido, mas sabe-se que envolve

alterações fundamentais no ácido desoxirribonucléico (DNA), conferindo características malignas e hereditárias às células transformadas e suas descendentes.

As células leucêmicas proliferam primariamente na medula óssea (tecido esponjoso que fica no interior dos ossos e onde todas as células do sangue são produzidas) e tecidos linfóides (constituinte principal do sistema imunológico, encontrado em várias partes do corpo, entre elas os gânglios linfáticos, o timo, baço, amígdalas e adenóides e medula óssea), onde interferem na hematopoiese normal e imunidade. Posteriormente, emigram para o sangue periférico e infiltram outros tecidos [33].

Até o momento, as causas para o surgimento da leucemia não são completamente conhecidas. Pesquisas indicam fatores de risco tais como: exposição à radiação e a substâncias químicas (ex.: tabagismo e benzeno), condições genéticas (ex.: Síndrome de Down), certos tipos de infecções virais (ex.:HTLV), entre outros [33].

A leucemia pode se apresentar de forma aguda ou crônica. Na primeira, células anormais, denominadas blastos, permanecem imaturas e são incapazes de desempenhar as funções de células normais. O número de blastos aumenta com grande velocidade e a doença apresenta evolução rapidamente fatal dos pacientes que não são tratados. Já na forma crônica há um número mais reduzido de blastos, que, em geral, apresentam-se mais maduros e podem desempenhar algumas das funções de uma célula normal. Além disso, na forma crônica, o número de blastos não aumenta tão rapidamente quanto na forma aguda, resultando num avanço mais gradual da doença [32].

São reconhecidas duas variantes principais das leucemias crônicas e agudas: linfocíticas e mielocíticas. Quando a doença atinge as células linfóides, é chamada de leucemia linfóide, linfoblástica ou linfocítica; enquanto ao atingir as células mielóides, é chamada de leucemia mielóide ou mielógena [33].

As células sanguíneas formam-se originalmente, das chamadas células-tronco pluripotentes da medula óssea vermelha que, em ativa proliferação, podem produzir as duas diferentes linhagens celulares, a linfóide e a mielóide. As células linfóides vão originar a linhagem dos linfócitos; enquanto as mielóides produzirão hemácias, os outros leucócitos e as plaquetas. A leucemia acontece quando o desenvolvimento dessas células passa a ser descontrolado e muito rápido, ocupando espaço na medula óssea e impedindo o desenvolvimento de outras células (hemácias, leucócitos e plaquetas).

Uma característica fundamental das células malignas na leucemia aguda é a incapacidade em amadurecer além do nível de mieloblasto ou pro-mielócito, na leucemia mielóide aguda (AML), e do nível de linfoblasto, na leucemia linfóide aguda

(ALL). As células leucêmicas acumulam-se na medula óssea devido à proliferação excessiva e a um defeito na diferenciação terminal celular (maturação). A incapacidade de amadurecimento terminal das células não-replicantes é a principal razão para o acúmulo de células leucêmicas na AML. As células leucêmicas proliferam primariamente na medula óssea, circulam no sangue e podem infiltrar outros tecidos, como linfonodos, fígado, baço, pele, vísceras e sistema nervoso central [33]. Os principais sintomas da leucemia decorrem do acúmulo dessas células na medula óssea, prejudicando, ou até impedindo, a produção dos glóbulos vermelhos (hemácias), causando anemia; dos glóbulos brancos, causando infecções; e, das plaquetas, causando hemorragias. Depois de instalada, a doença progride rapidamente, exigindo que o tratamento seja iniciado logo após o diagnóstico e a classificação da leucemia.

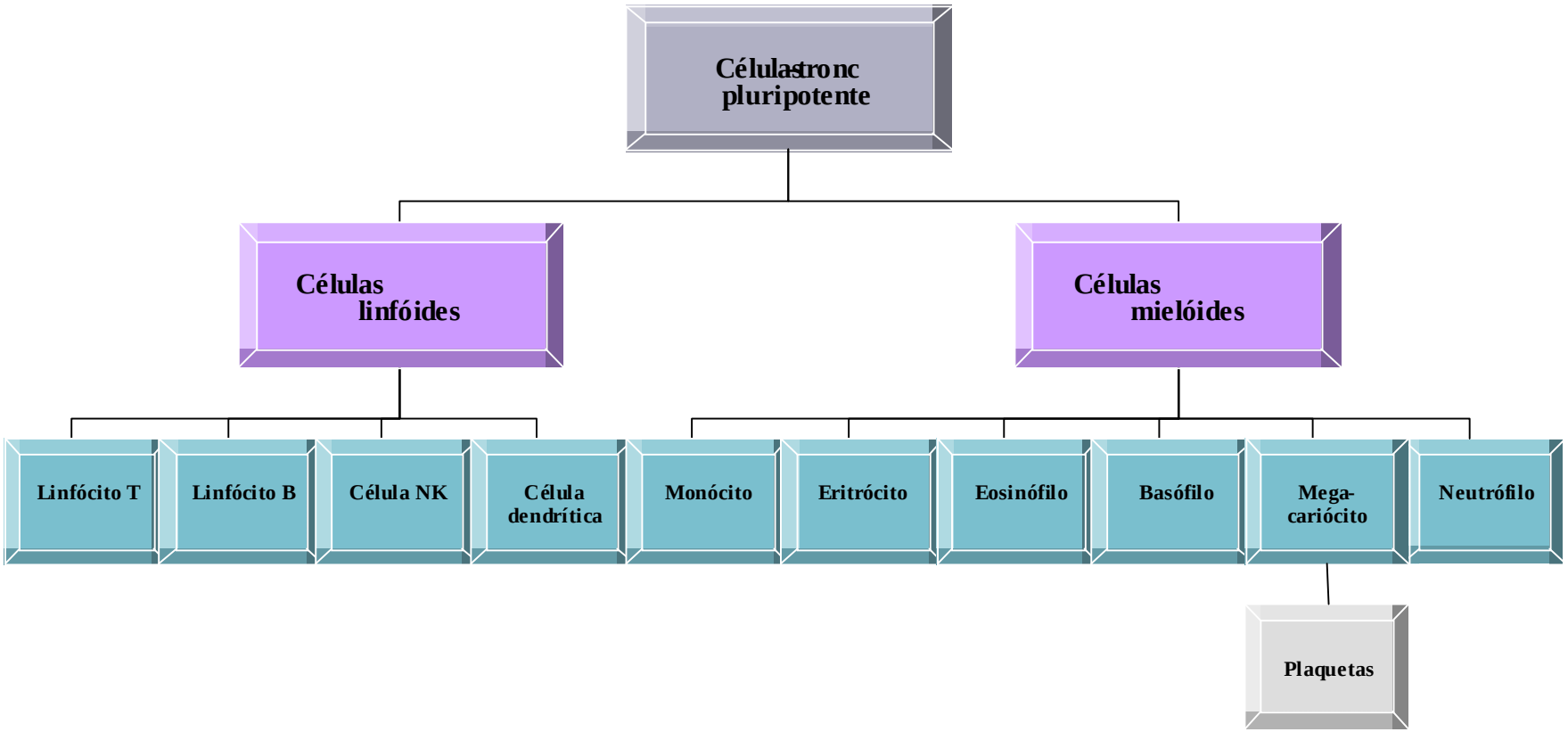

Figura 8: Hematopoiese - processo de formação, maturação e liberação na corrente sanguínea das células do sangue. O tecido conjuntivo hemocitopoético, ou tecido reticular, é produtor das duas linhagens de glóbulos: leucócitos e hemácias. Esse tecido aparece no baço, no timo e nos nódulos linfáticos, recebendo o nome de tecido linfóide. No interior da medula óssea vermelha, esse tecido é chamado mielóide, ocupando os espaços entre lâminas ósseas que formam o osso esponjoso.

Capítulo 3

3.1 Avaliação Inicial

Para avaliação da viabilidade do método proposto na seção 1.2, o algoritmo MIFS-U utilizando a entropia de Renyi (MIFS-U-Renyi), foi testado em dois bancos de dados e seus resultados comparados com o MIFS-U tradicional, que calcula a IM através da entropia de Shannon.

Os bancos de dados utilizados, que vamos denominar Doenças Cardíacas e Sonar, são largamente utilizados como base de comparação e estão disponíveis em http://mlearn.uci.edu/MLSummary.html e ftp://ftp.ics.uci.edu/pub/machine-learning-databases, respectivamente têm suas características resumidas na tabela a seguir:

Tabela 5: Descrição resumida dos dados para estudos preliminares

Conjunto de dados	Dimensão do espaço de entrada	Classes	Tamanho da Amostra Treino/Teste
Doenças Cardíacas	76	Presence/absence	540/200
Sonar	60	Rock/mine	104/104

3.1.1 Conjunto de dados Doenças Cardíacas

O conjunto de dados Doenças Cardíacas, disponível em *http://mlearn.uci.edu/MLSummary.html*, consta do diagnóstico de doenças cardíacas. Todos os atributos são numéricos e os dados foram coletados dos quatro seguintes locais: Cleveland Clinic Foundation [http://www.clevelandclinic.org/] ; Hungarian Institute of Cardiology, Budapest [http://www.kardio.hu/]; V.A, Medical Center, Long Beach [http://www.va.gov/]; University Hospital, Zurich [http://www.en.usz.ch/german].

Este conjunto de dados originalmente continha o diagnóstico de 740 pacientes com 76 atributos, sendo 540 pacientes reservados ao conjunto de treinamento e 200 ao conjunto de teste.

Em particular, o banco de dados de Cleveland é o único que foi até o momento utilizado pelos pesquisadores e apenas 14 dos 76 atributos são considerados. Será também o único utilizado neste estudo tendo ainda 4 atributos eliminados uma vez que continham dados faltosos.

Tabela 6: Descrição dos atributos do banco de dados Doenças Cardíacas

Atributos	Nome	Descrição
#3	age	Idade em anos
#4	sex	Sexo (1 = “male”, 0 = “Female”)
#9	cp	Tipo de dor no peito (Chest Pain Type)
#10	trestbps	Pressão sanguínea em repouso (em mm Hg na admissão ao hospital)
#12	chol	Colesterol sérico em mg/dl
#16	fbs	Açucar no sangue em jejum > 120 mg/dl (1 = ”true; 0 =”false”)
#19	restecg	Eletrocardiograma em repouso (0 = normal; 1 = anormalidade ST-T (inversão da onda T, elevação ou depressão da onda ST > 0,05 mV); 2 = hipertrofia ventricular esquerda provável ou definitiva segundo o critério Estes.
#32	thalach	Freqüência cardíaca máxima
#38	exang	Angina induzida por exercícios (1 = “yes"; 0 = “no”)
#40	oldpeak	Depressão da onda ST induzida por exercícios medida em repouso.
#41	slope	Curvatura do pico da curva ST (1 = “upsloping”; 2 = “flat”; 3 = “downsloping”)
#44	ca	Número de grandes vasos (0-3) coloridos por fluoroscopia
#51	thal	3 = “normal”; 6 = “fixed defect”; 7 = “reversable defect”
#58	num	Diagnóstico de doença cardíaca (0 = < 50% de estreitamento do diâmetro; 1 = > 50% estreitamento do diâmetro).

A variável de saída deste banco se refere à presença ou ausência de doença cardíaca e é representada por um valor inteiro que varia de 0 (ausência) até 4 (1 para angina típica, 2 para angina atípica, 3 para dor não-anginal e 4 para assintomático). Experimentos com o banco de dados de Cleveland se propõem a simplesmente distinguir a presença (valores 1, 2, 3 e 4) da ausência (valor 0), ficando assim reduzido a um problema dicotômico.

Para a nossa análise, os atributos foram selecionados utilizando-se informação mútua, calculada tanto pela entropia de Shannon quanto pela entropia de Renyi e a acurácia do classificador foi avaliada utilizando-se SVM não-linear com função de núcleo de base radial (RBF). No pré-processamento os dados foram normalizados retirando-se a média e dividindo-se pelo desvio padrão com o objetivo de homogeneizar a variabilidade das variáveis, criando um intervalo de amplitude similar onde todas as variáveis residirão.

A RBF utilizada no modelo requer a determinação de dois parâmetros: um parâmetro de custo (c); e um parâmetro da função de núcleo (g). O parâmetro c é um parâmetro de penalidade para vetores de treinamento classificados incorretamente, enquanto que o parâmetro g é um parâmetro da função de núcleo tal que:

$$K(x,y) = exp\left(g \cdot \frac{-||x-y||^2}{c} \right) \tag{60}$$

As acurácias de treino e teste foram obtidas nos conjuntos de treino e teste para os atributos que foram ordenados por ordem de seleção tanto pelo MIFS-U Renyi quanto pelo MIFS-U (tabela 4). Utilizou-se o parâmetro de regularização c igual a 0,1 e g da função gama igual a 0,1.

Utilizando a entropia de Shannon, o melhor valor de acurácia no conjunto de teste foi atingido ao serem selecionados 5 atributos (acurácia de 81%). Para a entropia de Renyi, o valor máximo de acurácia no conjunto de teste foi atingido ao serem selecionados 3 atributos (82,5%). Ao serem incluídos novos genes ao conjunto de preditores, nota-se uma queda na acurácia no conjunto de teste que pode ser atribuída à inclusão de informação irrelevante ao modelo sendo assim consideradas como variáveis de confusão, mostrando que o método desenvolvido pode ser altamente influenciado pela quantidade de informação irrelevante.

Tabela 7: Acurácia de treino e teste para o banco Heart Disease utilizando informação mútua via Shannon e via Renyi para seleção de variáveis e classificador SVM.

Atributos	Treino Shannon %	Treino Renyi %	Teste Shannon %	Teste Renyi %
1	59,44	72,41	65,00	77,00
2	66,11	72,41	65,50	78,00
3	75,18	77,04	78,50	82,50
4	77,41	78,15	79,00	77,00
5	78,52	81,30	81,00	80,00
6	78,15	81,48	80,50	80,00
7	77,41	82,22	79,00	82,00
8	78,89	82,04	77,50	81,00
9	78,70	82,04	82,00	80,50
10	82,22	82,22	81,50	81,50

3.1.2 Conjunto de dados Sonar

O conjunto de dados Sonar, disponível em f*tp://ftp.ics.uci.edu/pub/machine-learning-databases,* contém 208 padrões obtidos pelo sinal emitido por um sonar ao atingir dois diferentes tipos de materiais. Do total de observações, 111 padrões representam o sinal emitido ao atingir por cilindro de metal, em diferentes ângulos e sob diversas condições e 97 padrões representam o sinal emitido ao atingir pedras sob condições similares. Sendo assim, cada registro está associado a um rótulo contendo as letras “R”, se o objeto for uma pedra ou “M” se o objeto for um cilindro de metal.

Possui ao todo 60 atributos, que variam de 0 até 1, cada um representando a energia de uma banda de freqüência particular em um determinado intervalo de tempo.

Assim, como no dados de doenças cardíacas, os dados foram normalizados retirando-se a média e dividindo-se pelo desvio padrão e as acurácias de treino, teste e validação foram obtidas para os conjuntos de treino e teste para os 30 atributos melhor ordenados pela seleção do MIFS-U Renyi e pelo MIFS-U utilizando-se SVM

não – linear com parâmetro de regularização c igual a 5 e g da função gama igual a 3 para MIFS-U e 13 para MIFS-U-Renyi.

A tabela 6 mostra a acurácia de treino e de teste segundo a entropia de Shannon e de Renyi obtida variando-se o número de atributos.

Tabela 8: Acurácia de teste para o banco Sonar utilizando informação mútua via Shannon para seleção de variáveis e classificador SVM.

Número de atributos	Treino Shannon %	Treino Renyi %	Teste Shannon %	Teste Renyi %
1	54.81	69.23	44.23	60.58
2	58.65	72.12	50,00	78.85
3	57.69	74.04	50,96	77.89
4	59,62	79,81	50,96	76,92
5	60,58	82,69	53,85	81,73
6	63,46	84,62	54,81	82,69
7	66,35	88,46	54,81	83,65
8	70,19	90,39	58,65	85,58
9	72,12	92,31	71,15	84,62
10	73,08	95,19	69,23	84,62
11	72,12	96,15	75,96	88,46
12	76,92	99,04	79,81	91,35
13	75,96	99,04	78,85	91,35
14	78,85	100	76,92	89,42
15	81,73	100	75,00	88,46
16	86,52	100	77,89	89,42
17	90,39	100	82,69	88,46
18	96,15	100	82,69	88,46
19	97,12	100	84,62	87,50
20	98,08	100	87,50	88,46
21	99,04	100	90,39	88,46
22	99,04	100	88,46	89,42
23	99,04	100	91,35	93,27
24	100	100	89,42	90,39
25	100	100	89,42	90,39
26	100	100	88,46	85,58
27	100	100	87,50	82,69
28	100	100	90,39	80,77
29	100	100	91,35	82,69
30	100	100	90,39	81,73
31	100	100	89,42	81,73
32	100	100	89,42	81,73
33	100	100	89,42	81,73
34	100	100	88,46	80,77
35	100	100	86,54	81,73
36	100	100	86,54	81,73

Número de atributos	Treino Shannon %	Treino Renyi %	Teste Shannon %	Teste Renyi %
37	100	100	86,54	81,73
38	100	100	86,54	78,85
39	100	100	87,50	78,85
40	100	100	87,50	78,85
41	100	100	87,50	76,92
42	100	100	86,54	75,96
43	100	100	86,54	75,96
44	100	100	87,50	75,00
45	100	100	87,50	75.00
46	100	100	87,50	74,04
47	100	100	87,50	71,15
48	100	100	87,50	71,15
49	100	100	87,50	70,19
50	100	100	87,50	70,19
51	100	100	87,50	68,27
52	100	100	87,50	67,31
53	100	100	87,50	66,35
54	100	100	87,50	66,35
55	100	100	87,50	65,39
56	100	100	87,50	64,42
57	100	100	87,50	64,42
58	100	100	87,50	63,46
59	100	100	87,50	62,50
60	100	100	87,50	62,50

Utilizando a entropia de Shannon, o melhor valor de acurácia no conjunto de teste foi atingido ao serem selecionados 23 atributos (acurácia de 91,35%). Para a entropia de Renyi, esta acurácia foi obtida com apenas 12 atributos e o valor máximo no conjunto de teste também foi atingido ao serem selecionados 23 atributos (93,27%). Ao serem incluídos novos genes ao conjunto de preditores, nota-se uma queda na acurácia no conjunto de teste que pode ser atribuída à inclusão de informação irrelevante ao modelo, sendo assim, consideradas como variáveis de confusão, mostrando que o método desenvolvido pode ser altamente influenciado pela quantidade de informação irrelevante.

3.2 Leucemia Aguda

A seleção de atributos via MIFS-U-Renyi proposta neste trabalho foi aplicada no conjunto de dados de Leucemia Aguda. Esse conjunto possui uma dimensão maior quanto ao número de variáveis e tamanho de amostra pequeno, sendo dados oriundos de experimentos de microarranjos de DNA, disponível em http://broad.mit.edu/cgi-bin/datasets.cgi.

O conjunto de dados de Leucemia aguda contém o perfil de expressão gênico de 72 pacientes com leucemia aguda, extraído de biópsias de medula óssea (62 casos) e sangue periférico (10 casos) medidos em 7129 sondas de 6187 genes humanos. O conjunto de treinamento é constituído de 27 amostras de leucemia linfoblástica aguda (ALL) e 11 amostras de leucemia mielóide aguda (AML) enquanto que o conjunto de teste independente (validação) possui 20 amostras ALL e 14 amostras AML. Esta divisão corresponde a utilizada por Golub *et al. [6]*

O pré-processamento dos dados foi feito seguindo a publicação de DEB e REDDY [34] e PAUL e IBA [35]. com as seguintes etapas: (1) Uma vez que valores de expressão gênica negativos ou muito pequenos geralmente ocorrem devido à problemas experimentais, valores negativos ou menores que 20 foram substituído por 20. Da mesma forma, o limite superior de 16000 foi imposto visando adequação à sensibilidade do scanner. (2) Análise dos perfis de expressão gênicos e exclusão daqueles com razão entre níveis de expressão máximo e mínimo ≤ 5 e diferença ≤ 500, já que é esperado que os genes necessários para a diferenciação dos subtipos de leucemia tenham ampla variação. Após esta etapa restaram 3859 genes; (3) Normalização das amostras de treinamento e de teste, após transformação logarítmica de base 10 dos valores de expressão, pela subtração das médias dos perfis de expressão e divisão pelos desvios-padrão obtidos da amostra de treino, seguindo a sugestão de Golu *et al*. [6]

Para o parâmetro de controle da redundância equação (29), foi atribuído o valor 1 baseado na literatura [26, 36].

Na etapa de classificação utilizou-se a SVM não-linear e cálculo da entropia de Renyi cujos parâmetros estão resumidos na tabela 9.

As acurácias de treino, teste e validação foram obtidas para os conjuntos de treino e teste para os 30 atributos melhor ordenados pela seleção do MIFS-U-Renyi e pelo MIFS-U utilizando-se SVM não – linear

Tabela 9: Parâmetros SVM para o conjunto de dados de Leucemia

Parâmetro	Descrição	MIFS-U (SHANNON)	MIFS-U (RENNYI)
T	Tipo de núcleo	2 (Núcleo Gaussiano)	2 (Núcleo Gaussiano)
C	Parâmetro de regularização	10	0.3
S	Tipo de SVM	0 (C-SVC)	0 (C-SVC)
G	Gama da função de núcleo	0.02	0.2
Beta	parâmetro de controle da redundância	1	1

Tabela 10: Acurácia de treino e teste para o conjunto de dados Leucemia, utilizando MIFS-U-Renyi.

Número de atributos	Treino Renyi %	Teste Renyi %
1	92,11	97.06
2	92,11	94.12
3	97,37	97.06
4	97,37	97.06
5	100	97.06
6	100	97.06
7	100	97.06
8	100	94.12
9	100	94.12
10	100	97.06
11	100	97.06
12	100	97.06

13	100	94.12
14	100	94.12
15	100	94.12
16	100	94.12
17	100	94.12
18	100	94.12
19	100	94.12
20	100	94.12
21	100	91.18
22	100	91.18
23	100	88.23
24	100	88.23
25	100	85.29

.

Tabela 11: Acurácia de treino e teste para o conjunto de dados Leucemia, utilizando MIFS-U.

Número de atributos[2]	Acurácia de Treino Shannon %	Acurácia Teste Shannon %
1	71,05	58.82
7	84,21	82.35
11	89.47	85.29
12	89.47	85.29
15	92,11	73.52
20	97,37	70.58
25	100	76.47
30	100	82.35
31	100	82.35
32	100	82.35
33	100	82.35
34	100	82.35
35	100	82.35
36	100	82.35

[2] Os valores omitidos apresentaram acurácia idêntica ou muito próxima aos valores vizinhos.

37	100	82.35
38	100	85.29
39	100	85.29
40	100	88.23
50	100	82.35
55	100	91.17
56	100	91.17
57	100	94.11
58	100	91.17
59	100	91.17
63	100	94.11

Tabela 12: Ordem de seleção dos atributos para o conjunto de dados Leucemia utilizando MIFS-U e MIFS-U-Renyi

Ordem de seleção Via MIFS-U-Renyi	Ordem de seleção Via MIFS-U
1272	289
3473	2455
2781	3179
1627	2015
592	290
1417	2730
783	3415
1414	3727
1314	1910
2731	3205
986	128
833	3283
1016	3016
1061	2103
2463	2539

A melhor acurácia obtida até 100 atributos para a MIfs_U foi encontrada com 57 atributos e acurácia de treino foi de 92%. Pode-se observar que, tanto para a

entropia de Shannon quanto para a entropia de Renyi, o MIFS_U apresentou melhores resultados com menor número de atributos. Por apresentar boa acurácia com número pequeno de atributos, este desempenho justificou a escolha do MIFS_U_Renyi como o método de seleção utilizado no conjunto de dados de Leucemia aguda.

Para todos os estudos realizados utilizou-se o software MATLAB® (The MathWorks, Inc.; HTTP://www.mathworks.com/), o pacote LIBSVM [37], implementado em C com interface para MATLAB® e os algoritmos desenvolvidos para cálculo da informação mútua via entropia de Shannon e de Renyi (MIFS-U e MIFS-U-Renyi) e implementados em MATLAB® .

Capítulo 4

Discussão

Neste trabalho foi utilizado para a seleção de atributos (genes) o algoritmo MIFS-U; mas, diferentemente do original apresentado por Kwak e Choi [10] e de vários outros trabalhos que o sucederam, a entropia de Shannon foi substituída pela entropia de Renyi, passando então a ser chamado de MIFS-U-Renyi.

Alguns pesquisadores optam por incluir a amostra de teste no treinamento do algoritmo de seleção, visando evitar a especialização do classificador, uma vez que a amostra de treinamento em dados de microarrando de DNA tem tamanho reduzido. Assim, apresentam somente a acurácia obtida durante a fase de treino, geralmente advinda de validação cruzada tipo LOOCV. Este enfoque pode acarretar em um viés e quando testado em amostras independentes, favorece predições com baixo desempenho. Para evitar este tipo de viés, a acurácia de teste foi priorizada neste estudo, mas para permitir a comparação com outros trabalhos, a acurácia de treino também foi determinada seguindo este modelo.

Quando comparado ao estudo desenvolvido por Marohinie *et al*. [7] no mesmo conjunto de dados e com metodologia semelhante à aqui empregada, obtivemos resultados superiores, mesmo com número reduzido de genes utilizado no treinamento do classificador.

Marohinic *et al*. [7] utilizaram a informação mútua combinada com SVM para filtrar e classificar os atributos. Para diferentes números de genes selecionados, a largura da função de núcleo foi variada e o resultado aqui apresentado foi aquele que obteve melhor performance, diferentemente do nosso estudo, que utilizou uma largura fixa para todos os valores selecionados. Os seus resultados foram pouco inferior aos obtidos através do MIFS-U-Renyi, como pode ser visto na tabela a seguir:

Tabela 13: Comparação dos resultados do método proposto (MIFS-U-Renyi) com o método MIFS associado a o classificador SVM em função da acurácia de validação cruzada no conjunto de treino (LOOCV), de teste e do número de atributos (genes) selecionados.

Número de atributos[3]	Acurácia de Treino MIFS/SVM (LOOCV) %	Acurácia Proposto (LOOCV) %	Acurácia Teste MIFS/SVM %	Acurácia Teste Proposto %
1	97,37	89,47	91,43	97.06
2	100	89,47	91,43	94.12
3	100	94,74	91,43	97.06
4	100	97,37	68,57	97.06
5	(*)	97,37	(*)	97.06
6	(*)	97,37	(*)	97.06
7	(*)	100	(*)	97.06

Os resultados acima mostram que o método proposto apesar de apresentar menor resultado na acurácia de validação cruzada (LOOCV), teve melhor desempenho ao classificar o conjunto de dados independentes, tendo apenas uma observação classificada erroneamente.

Em outro estudo utilizando informação mútua, Guo *et al.* [38] organizou os atributos de acordo com a informação mútua entre o nível de expressão e a classe de saída para todas as observações e aqueles que apresentaram os maiores valores foram selecionados e o método de vizinhos próximos foi aplicado. Esta metodologia foi chamada de GSMI.

A acurácia mais alta (LOOCV utilizando as 72 observações) foi encontrada quando 19 genes foram selecionados, 98,61%. Estando pouco abaixo do atingido pelo método proposto no presente estudo. A acurácia no conjunto de teste não foi informada.

Além dos métodos baseados em informação mútua, muitos outros vêm sendo testados exaustivamente na seleção de atributos em dados de microarranjo.

Li et al. [12] utilizaram o método “Sequential Floating Foward Search” (SFFS) associado à SVM, tendo a distância de Bhattacharyya como função de avaliação dentro da SFFS para selecionar um conjunto de candidatos a atributos. Em linhas

(*) Valores não apresentados pelo autor.

gerais, o algorítimo SFFS seleciona subconjuntos de candidatos a atributos maximizando a distância de Bhattacharyya.

O único resultado discriminado consta de 5 genes; onde, tanto o erro LOOCV quanto o erro no conjunto de teste foram nulos. Tamanha precisão na classificação dos dois subtipos de leucemia é discordante de outros trabalhos anteriores, uma vez que a observação #66 pode ter sido rotulada erroneamente como AML, como citado na seção 2.1.1.

Como observado na seção 2.1, os conjuntos de genes preditores variam de acordo com o método de seleção empregado, havendo pequena ou nenhuma interseção entre eles. Essas múltiplas soluções não têm sido investigadas na literatura, não apresentando justificativa para essas diferenças. Uma análise básica de correlação entre o subconjunto de atributos obtido através do método aqui proposto e as demais metodologias apresentadas não obteve resultados relevantes, pois um número muito pequeno de genes apresentaram correlação em torno de 0,7.

Dos métodos estudados, os subconjuntos apresentados por Li *et al.* e Golub foram os que apresentaram maiores correlações. Os genes 1834 e 2121 obtiveram a maior correlação de 0,81, assim como 1882 e 4177. Ao se comparar o método proposto com o de Golub *et al.* os genes 592 e 5255 apresentaram a correlação mais alta igual a 0,72 e ao se comparar o subconjunto de atributos selecionados por Marohnic *et al.* com o método de Golub *et al.* a maior correlação foi verificada entre os genes 4847 e 6373. Sendo assim, a busca para uma justificativa a essas diferentes soluções requer um estudo relacionado às vias de ativação entre os distintos genes, pois esta poderia, enventualmente, gerar alguma informação, fornecendo subsídios para os diferentes métodos, o que não foi objeto específico deste trabalho.

Tabela 14: Correlação entre os conjuntos de genes

Genes (Golub/Li)	Correlação
2121 e 1834	0,81
4177 e 1882	0,81
5501 e 2241	0,77
6376 e 1882	0,80
6376 e 1834	0,77
6803 e 1882	0,78

Genes (Golub/Marohnic)	Correlação
6376 e 4847	0,70
6373 e 4847	0,74

Genes (Golub/Proposto)	Correlação
5254 e 592	0,73

A desvantagem do método proposto reside em não haver um processo automático de seleção de atributos para a SVM, tendo esta de ser feita por tentativa e erro.

Capítulo 5

Conclusão

Neste estudo propomos um algoritmo de seleção de atributos para problemas de classificação, em particular, de dois subtipos predominantes de leucemia. Algoritmos tais como, MIFS_U, baseados em teoria da informação têm a vantagem de dispenderem um custo computacional relativamente pequeno [10], mas ao se trabalhar com conjunto de dados contínuos, a estimativa da pdf se torna uma limitação.

O método proposto é uma evolução do MIFS-U onde a entropia de Shannon foi substituída pela entropia de Renyi, permitindo ser aplicado com bom desempenho tanto em dados discretos quanto contínuos, sem comprometimento do custo computacional. Combinado com o classificador tipo SVM, apresentou acurácia igual ou superior aos métodos anteriormente desenvolvidos, utilizando um subconjunto de atributos com número reduzido de genes.

Como trabalho futuro, deseja-se testar a metodologia aqui apresentada em outros bancos de dados, verificando sua capacidade de generalização e desenvolver um método que automatize o processo de determinação dos parâmetros da SVM, uma vez que ainda não existe um método formal para tal seleção. Um caminho a ser estudado é a utilização de algoritmos genéticos para este fim.

Bibliografia

[1] CHO, S.-B.,H.-H. WON, *Machine Learning in DNA Microarray Analysis for Cancer Classification*, in *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003 - Volume 19*. 2003, Australian Computer Society, Inc.: Adelaide, Australia.

[2] GUYON, I., J. WESTON, S. BARNHILL, et al., "Gene Selection for Cancer Classification Using Support Vector Machines"*. Machine Learning*, v. 46, n. 1, pp. 389-422,2002.

[3] FUREY, T.S., N. CRISTIANINI, N. DUFFY, et al., "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data"*. Bioinformatics*, v. 16, n. 10, pp. 906-914,2000.

[4] KOHAVI, R.,G.H. JOHN, "Wrappers for Feature Subset Selection"*. Artif. Intell.*, v. 97, n. 1-2, pp. 273-324,1997.

[5] HU, X.,Y. PAN, *Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications* 1v. 1 ed. New York, ed. Wiley, 2007.

[6] GOLUB, T., D. SLONIM, P. TAMAYO, et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring"*. Science*, v. 286, n. 5439, pp. 531-537,1999.

[7] MAROHNIC, V., Z. DEBELJAK, N. BOGUNOVIC. *Mutual Information Based Reduction of Data Mining Dimensionality in Gene Expression Analysis*. in *Information Technology Interfaces, 2004. 26th International Conference on*. 2004.

[8] PEDREIRA, C.E., L. MACRINI, M.G. LAND, et al., "A New Decision Support Tool for Treatment Intensity Choice in Childhood Acute Lymphoblastic Leukemia"*. IEEE Transactions on Information Technology in Biomedicine, (no prelo)*, v. n. pp.,2008.

[9] PRINCIPE, J.C.,D. XU, "Information-Theoretic Learning Using Renyi's Quadratic Entropy", In: Editor^Editors, *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation*, 1 v. 407--412, USA, Aussois. 1999.

[10] KWAK, N.,C. CHONG-HO, "Input Feature Selection by Mutual Information Based on Parzen Window"*. Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 24, n. 12, pp. 1667-1671,2002.

[11] INSTITUTO NACIONAL DO CÂNCER, *Estimativa 2008: Incidência Do Câncer No Brasil,* Brasil, 2008

[12] LI, Y.-X., Y.-H. ZHU, X.-G. RUAN. *Gene Selection for Leukemia Subtype Classification from Gene Expression Profile*. in *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*. 2004.

[13] WANG, Y., I.V. TETKO, M.A. HALL, et al., "Gene Selection from Microarray Data for Cancer Classification--a Machine Learning Approach"*. Computational Biology and Chemistry*, v. 29, n. 1, pp. 37-46,2005.

[14] WATSON, J.D.,F.H.C. CRICK, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid"*. Nature*, v. 171, n. 4356, pp. 737-738,1953.

[15] TOWNSEND, M.C., D.C.S. JUNIOR, D. BEUCHAMP, *Sabiston: Tratado De Cirurgia*, 1 v. 17 ed., Elsevier 2005.

[16] SCHENA, M., D. SHALON, R.W. DAVIS, et al., "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray"*. Science*, v. 270, n. 5235, pp. 467-470,1995.

[17] BROWN, P.O.,D. BOTSTEIN, "Exploring the New World of the Genome with DNA Microarrays"*. Nat Genet*, v. n. pp.

[18] DUGGAN, D.J., M. BITTNER, Y. CHEN, et al., "Expression Profiling Using Cdna Microarrays"*. Nat Genet*, v. n. pp.

[19] BROWN, P.O.,D. BOTSTEIN, "Exploring the New World of the Genome with DNA Microarrays". *Nat Genet*, v. 21, n. 1 Suppl, pp. 33-37,1999.

[20] KIM, K.-J.,S.-B. CHO, "Ensemble Classifiers Based on Correlation Analysis for DNA Microarray Classification". *Neurocomputing*, v. 70, n. 1-3, pp. 187-199,2006.

[21] JAIN, A.K., R.P.W. DUIN, M. JIANCHANG, "Statistical Pattern Recognition: A Review". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 22, n. 1, pp. 4-37,2000.

[22] GUYON, I.,A. ELISSEEFF, "An Introduction to Variable and Feature Selection". *Journal of Machine Learning Research*, v. 3, n. pp. 1157-1182,2003.

[23] HUERTA, E., B. DUVAL, J.-K. HAO, "A Hybrid Ga/Svm Approach for Gene Selection and Classification of Microarray Data", In: Editor^Editors, *Applications of Evolutionary Computing*, v. 34-44. 2006.

[24] KAROL YCZKOWSKI, "Renyi Extrapolation of Shannon Entropy". *Open Systems & Information Dynamics*, v. 10, n. 3, pp. 297-310,2003.

[25] PRINCIPE, J.C., D. XU, J.W.F. III, *Chapter 7 Information-Theoretic Learning*.

[26] BATTITI, R., "Using Mutual Information for Selecting Features in Supervised Neural Net Learning". *Neural Networks, IEEE Transactions on*, v. 5, n. 4, pp. 537-550,1994.

[27] VAPNIK, V., *Statistical Learning Theory*, New York, Wiley-Interscience, 1998.

[28] VAPNIK, V., *The Nature of Statistical Learning Theory*, New York, Springer-Verlag, 1995.

[29] NOBLE, W.S., "What Is a Support Vector Machine?" *Nat Biotech*, v. 24, n. 12, pp. 1565-1567,2006.

[30] BURGES, C., "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Min. Knowl. Discov.*, v. 2, n. 2, pp. 121-167,1998.

[31] MULLER, K.R., S. MIKA, G. RATSCH, et al., "An Introduction to Kernel-Based Learning Algorithms". *Neural Networks, IEEE Transactions on*, v. 12, n. 2, pp. 181-201,2001.

[32] COTRAN, R.S., V. KUMAR, S.L. ROBBINS, *Robins: Patologia Estrutural E Funcional*, 1 v. 4 ed. Brasil, Guanabara Koogan, 1991.

[33] BRAUNWALD, E., K.J. ISSELBACHER, R.G. PETERSDORF, et al., *Harrisson: Medicina Interna*, 2 v. Brasil, 1988.

[34] DEB, K.,A. RAJI REDDY, "Reliable Classification of Two-Class Cancer Data Using Evolutionary Algorithms". *Biosystems*, v. 72, n. 1-2, pp. 111-129,2003.

[35] PAUL, T.K.,H. IBA. *Selection of the Most Useful Subset of Genes for Gene Expression-Based Classification*. in *Congress on Evolutionary Computation, 2004. {CEC2004}*. 2004: IEEE.

[36] NOJUN, K.,C. CHONG-HO. *Improved Mutual Information Feature Selector for Neural Networks in Supervised Learning*. 1999.

[37] CHANG, C.,C. LIN, Libsvm: A Library for Support Vector Machines, 2001. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[38] GUO, S.-B., M. LYU, T.-M. LOK, "Gene Selection Based on Mutual Information for the Classification of Multi-Class Cancer", In: Editor^Editors, *Computational Intelligence and Bioinformatics*, v. 454-463. 2006.

Anexo

Descrição dos genes preditores nos estudos avaliados.

Golub *et al.*

Num	Gene Description	Gene Accession Number
1249	INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1	L08246_at
1630	Inducible protein mRNA	L47738_at
1704	ADA Adenosine deaminase	M13792_at
1745	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog	M16038_at
1834	CD33 CD33 antigen (differentiation antigen)	M23197_at
1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891_at
1909	IL7R Interleukin 7 receptor	M29696_at
1928	Oncoprotein 18 (Op18) gene	M31303_rna1_at
2020	FAH Fumarylacetoacetate	M55150_at
2043	LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)	M57710_at
2111	ATP6C Vacuolar H+ ATPase proton channel subunit	M62762_at
2121	CTSD Cathepsin D (lysosomal aspartyl protease)	M63138_at
2186	MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3	M69043_at
2242	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR	M80254_at
2288	DF D component of complement (adipsin)	M84526_at
2348	ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain	M91432_at
2354	CCND3 Cyclin D3	M92287_at
2402	Azurocidin gene	M96326_rna1_at
2441	HKR-T1	S50223_at
2642	MB-1 gene	U05259_rna1_at
2909	SRP9 Signal recognition particle 9 kD protein	U20998_at
3056	Cytoplasmic dynein light chain 1 (hdlc1) mRNA	U32944_at
3096	Heterochromatin protein p25 mRNA	U35451_at

3258	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA	U46751_at
3320	Leukotriene C4 synthase (LTC4S) gene	U50136_rna1_at
3847	GB DEF = Homeodomain protein HoxA9 mRNA	U82759_at
4052	Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)	X04085_rna1_at
4177	IRF2 Interferon regulatory factor 2	X15949_at
4196	PRG1 Proteoglycan 1, secretory granule	X17042_at
4328	PROTEASOME IOTA CHAIN	X59417_at
4389	GTF2E2 General transcription factor TFIIE beta subunit, 34 kD	X63469_at
4535	RETINOBLASTOMA BINDING PROTEIN P48	X74262_at
4847	Zyxin	X95735_at
4973	RABAPTIN-5 protein	Y08612_at
5093	Mitogen inducible gene mig-2	Z24725_at
5191	Adenosine triphosphatase, calcium	Z69881_at
5254	MCM3 Minichromosome maintenance deficient (S. cerevisiae) 3	D38073_at
5501	TOP2B Topoisomerase (DNA) II beta (180kD)	Z15115_at
5593	Transcriptional activator hSNF2b	D26156_s_at
5772	C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds	U22376_cds2_s_at
6200	Interleukin 8 (IL8) gene	M28130_rna1_s_at
6201	INTERLEUKIN-8 PRECURSOR	Y00787_s_at
6281	MYL1 Myosin light chain (alkali)	M31211_s_at
6373	ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide)	M81695_s_at
6376	PFC Properdin P factor, complement	M83652_s_at
6515	DHPS Deoxyhypusine synthase	U26266_s_at
6539	Epb72 gene exon 1	X85116_rna1_s_at
6803	LYZ Lysozyme	M19045_f_at
6855	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	M31523_at
7119	Transcriptional activator hSNF2b	U29175_at

Marohnic *et al.*

Num	Gene Description	Gene Accession Number
804	Macmarcks	HG1612-HT1612_at
1574	(clone S171) mRNA	L40393_at
1779	MPO Myeloperoxidase	M19507_at
4847	Zyxin	X95735_at
4951	Nucleoside-diphosphate kinase	Y07604_at

Li *et al.*

Num	Gene Description	Gene Accession Number
1779	MPO Myeloperoxidase	M19507_at
1796	APOC1 Apolipoprotein CI	M20902_at
1834	CD33 CD33 antigen (differentiation antigen)	M23197_at
1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891_at
2242	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR	M80254_at

Método Proposto

Num	Gene Description	Gene Accession Number
1272	TGM3 Transglutaminase 3 (E polypeptide, protein-glutamine-gamma-glutamyltransferase)	L10386_at
3473	FEZ1 mRNA	U60060_at
2781	Homolog of Drosophila discs large protein, isoform 2 (hdlg-2) mRNA	U13896_at
1627	PSEN2 Presenilin 2 (Alzheimer disease 4)	L43964_at
592	KIAA0186 gene	D80008_at