Redução de dimensionalidade utilizando entropia condicional média aplicada a problemas de bioinformática e de processamento de imagens

David Correa Martins Junior

DISSERTAÇÃO APRESENTADA AO INSTITUTO DE MATEMÁTICA E ESTATÍSTICA DA UNIVERSIDADE DE SÃO PAULO PARA OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO

Área de Concentração: Ciência da Computação
Orientador: Prof. Dr. Roberto Marcondes Cesar Junior

Durante a elaboração deste trabalho, o autor recebeu apoio financeiro da FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo (proc. 02/04611-0)

São Paulo, dezembro de 2004

Redução de dimensionalidade utilizando entropia condicional média aplicada a problemas de bioinformática e de processamento de imagens

Este exemplar corresponde à redação final
da dissertação devidamente corrigida
e apresentada por David Correa Martins Junior
e aprovada pela Comissão Julgadora.

São Paulo, 06 de dezembro de 2004

Banca Examinadora:

- Prof. Dr. Roberto Marcondes Cesar Junior (orientador) - MAC-IME-USP
- Prof. Dr. Junior Barrera - MAC-IME-USP
- Profa. Dra. Maria Carolina Monard - DCCE-ICMC-USP

Agradecimentos

Em primeiro lugar, gostaria de agradecer ao meu grande amigo, o Prof. Dr. Roberto Marcondes Cesar Jr., pela sua energia e disposição em orientar este mestrado. Sempre cativante, ele sabe transmitir sua energia positiva e sua sabedoria aos alunos.

Não poderia deixar de agradecer também ao Prof. Dr. Junior Barrera pela intensa colaboração com esta pesquisa. Sempre disposto a ajudar com todo o seu conhecimento, sua co-orientação foi fundamental para o andamento e o enriquecimento de todo trabalho.

Agradeço à FAPESP pelo apoio financeiro concedido a essa pesquisa.

Meus agradecimentos aos colaboradores científicos: Prof. Dr. Paulo J. S. Silva e Ricardo Vêncio do IME-USP, Helena Brentani do Ludwig Institute for Cancer Research e Prof. Dr. Hernando Del Portillo do ICB-USP.

Ao Prof. Dr. Ronaldo F. Hashimoto e ao Yossi Zana pelas importantes observações e considerações sobre o texto de qualificação.

A todo o pessoal do laboratório de processamento de imagens do IME-USP, pelos auxílios prestados, seja com relação a rede ou sobre peculiaridades do Latex.

A todos os meus amigos, especialmente aos Imescos (TM) pelo sólido círculo de amizades construído ao longo dos últimos 7 anos.

Finalmente, dedico o presente texto aos meus pais, Beth e David, e à minha irmã, Vanessa, pelo apoio, pelo carinho e pela paciência fundamentais para o desenvolvimento desta pesquisa.

Resumo

Redução de dimensionalidade é um problema muito importante da área de reconhecimento de padrões com aplicação em diversos campos do conhecimento. Dentre as técnicas de redução de dimensionalidade, a de seleção de características foi o principal foco desta pesquisa. De uma forma geral, a maioria dos métodos de redução de dimensionalidade presentes na literatura costumam privilegiar casos nos quais os dados sejam linearmente separáveis e só existam duas classes distintas. No intuito de tratar casos mais genéricos, este trabalho propõe uma função critério, baseada em sólidos princípios de teoria estatística como entropia e informação mútua, a ser embutida nos algoritmos de seleção de características existentes. A proposta dessa abordagem é tornar possível classificar os dados, linearmente separáveis ou não, em duas ou mais classes levando em conta um pequeno subespaço de características. Alguns resultados com dados sintéticos e dados reais foram obtidos confirmando a utilidade dessa técnica.

Este trabalho tratou dois problemas de bioinformática. O primeiro trata de distinguir dois fenômenos biológicos através de seleção de um subconjunto apropriado de genes. Foi estudada uma técnica de seleção de genes fortes utilizando máquinas de suporte vetorial (MSV) que já vinha sendo aplicada para este fim em dados de SAGE do genoma humano. Grande parte dos genes fortes encontrados por esta técnica para distinguir tumores de cérebro (glioblastoma e astrocytoma), foram validados pela metodologia apresentada neste trabalho. O segundo problema que foi tratado neste trabalho é o de identificação de redes de regulação gênica, utilizando a metodologia proposta, em dados produzidos pelo trabalho de DeRisi *et al* sobre *microarray* do genoma do *Plasmodium falciparum*, agente causador da malária, durante as 48 horas de seu ciclo de vida. O presente texto apresenta evidências de que a utilização da entropia condicional média para estimar redes genéticas probabilísticas (PGN) pode ser uma abordagem bastante promissora nesse tipo de aplicação.

No contexto de processamento de imagens, tal técnica pôde ser aplicada com sucesso em obter W-operadores minimais para realização de filtragem de imagens e reconhecimento de texturas.

Abstract

Dimensionality reduction is a very important pattern recognition problem with many applications. Among the dimensionality reduction techniques, feature selection was the main focus of this research. In general, most dimensionality reduction methods that may be found in the literature privilegiate cases in which the data is linearly separable and with only two distinct classes. Aiming at covering more generic cases, this work proposes a criterion function, based on the statistical theory principles of entropy and mutual information, to be embedded in the existing feature selection algorithms. This approach allows to classify the data, linearly separable or not, in two or more classes, taking into account a small feature subspace. Results with synthetic and real data were obtained corroborating the utility of this technique.

This work addressed two bioinformatics problems. The first is about distinguishing two biological fenomena through the selection of an appropriate subset of genes. We studied a strong genes selection technique using support vector machines (SVM) which has been applied to SAGE data of human genome. Most of the strong genes found by this technique to distinguish brain tumors (glioblastoma and astrocytoma) were validated by the proposed methodology presented in this work. The second problem covered in this work is the identification of genetic network regulation, using our proposed methodology, from data produced by work of DeRisi *et al* about *microarray* of the *Plasmodium falciparum* genome, malaria agent, during 48 hours of its life cycle. This text presents evidences that using mean conditional entropy to estimate a probabilistic genetic network (PGN) may be very promising.

In the image processing context, it is shown that this technique can be applied to obtain minimal W-operators that perform image filtering and texture recognition.

Sumário

Lista de símbolos

$\mathbf{X}$	Vetor de características
X	Característica; variável aleatória (seção 5.2)
$\mathbf{x}$	Padrão ou instância observada de $\mathbf{X}$
x	Valor da característica X; variável de uma função densidade de probabilidades (seção 6.2.3); valor da variável aleatória X (seção 5.2)
Y	Variável aleatória correspondente aos rótulos das classes
y	Rótulo da classe; valor da variável aleatória Y (seção 5.2)
n	Dimensionalidade (tamanho) do espaço de características
c	Número de classes
$\mathcal{F}(\cdot)$	Função critério
T	Conjunto de amostras de treinamento
$\psi(\cdot)$	Classificador
$\mathcal{Z}$	Conjunto de índices de um subespaço de características
$\mathcal{I}$	Conjunto de índices do espaço total de características
$\mathbf{X}_{\mathcal{Z}}$	Subespaço de $\mathbf{X}$ onde os índices das características são dados por $\mathcal{Z}$
$\mathbf{x}_{\mathcal{Z}}$	Instância de $\mathbf{X}_{\mathcal{Z}}$
Δ	Constante de condição de parada do algoritmo SFFS
P	Probabilidade
$\hat{P}$	Probabilidade estimada
$\beta(\cdot)$	Função beta de probabilidade
a, b	Parâmetros da função beta
$\mathcal{C}$	Índice de credibilidade
t_1, t_2	Extremos do intervalo de credibilidade
$H(\cdot)$	Entropia
$M(\cdot)$	Informação mútua

i, j, h, q	Índices
m	Número de instâncias possíveis de um espaço de características
$E[\cdot]$	Esperança
o	Número de ocorrências de uma determinada instância no conjunto de treinamento
t	Número de amostras; Variável da função beta (seção 6.2.3); Instante de tempo (seção **??**)
p	Número de valores discretos que cada característica pode assumir
d	Dimensão de um subespaço do espaço total de características
α	Constante de ponderação positiva da entropia condicional média
$\bar{d}$	Dimensão de uma janela selecionada para o W-operador
F, G	Espaços de características (seção 5.2)
f	Característica de F (seção 5.2)
g	Característica de G (seção 5.2)
$\mathcal{Z}^*$	Conjunto de índices do melhor subespaço de características
$\eta[\cdot]$	Transformação normal
D_{min}	Dimensão onde ocorre o ponto de mínimo da entropia condicional média
μ	Média
σ	Desvio padrão
$\mathbf{X_r}$	Vetor de um subespaço de características relacionadas aos rótulos
$\mathbf{X_{nr}}$	Vetor de um subespaço de características não relacionadas aos rótulos
$\mathbf{x_r}$	Instância de $\mathbf{X_r}$
$\mathbf{x_{nr}}$	Instância de $\mathbf{X_{nr}}$
W	Janela onde um W-operador está definido
E	Plano dos inteiros ($Z \times Z$)
o	Origem de E
$\mathcal{P}(E)$	Conjunto potência de E
Ψ	Operador
Ψ_{opt}	Operador ideal (ótimo)
Ψ_{est}	Operador estimado do operador ideal
$\mathbf{S}$	Imagens de entrada
S	Realização de $\mathbf{S}$
$\mathbf{I}$	Imagens de saída

I	Realização de $\mathbf{I}$
$\mathcal{O}$	Espaço de operadores
$\ell(\cdot)$	Função perda
$nBib_i$	Número total de observações ds tags na biblioteca i em dados de SAGE

Lista de abreviaturas e termos

Característica	Tradução adotada para "feature" no contexto de reconhecimento de padrões
ECM	Entropia Condicional Média
DFT	Transformada Discreta de Fourier ("Discrete Fourier Transform")
MCI	Melhores Características Individuais
MSV	Máquina de Suporte Vetorial ("Support Vector Machines")
PGN	Rede Gênica Probabilística ("Probabilistic Genetic Network")
SAGE	Serial Analysis of Gene Expression
SFS	Sequential Forward Search - Busca Seqüencial para Frente
SFFS	Sequential Floating Forward Search - Busca Seqüencial Flutuante para Frente

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Comentários iniciais

A área de reconhecimento de padrões visa resolver problemas de classificação de objetos ou padrões em um número de categorias ou classes [70]. Dado um conjunto de c classes, $y_1, ..., y_c$, e um padrão desconhecido $\mathbf{x}$, um sistema de reconhecimento de padrões tem como finalidade associar $\mathbf{x}$ a uma classe y_i com base em medidas definidas em um espaço de características. Em diversas aplicações, a dimensão do espaço de características dos objetos tende a ser relativamente grande, tornando a tarefa de classificação bastante complexa e sujeita a erros. Deve-se a esse fato a importância do estudo do problema de redução de dimensionalidade em reconhecimento de padrões.

Redução de dimensionalidade é um problema genérico onde se deseja identificar um subespaço suficientemente reduzido de características que seja capaz de representar qualquer padrão conhecido de acordo com um determinado critério. Existem diversas abordagens para tratar este problema. Dentre elas, enfatizamos a técnica de seleção de características.

Seleção de características pode ser aplicada em várias situações onde verifica-se um grande espaço de características e deseja-se selecionar um subespaço adequado. Aplicações em bioinformática e processamento de imagens figuram entre elas, tendo sido os principais alvos de nosso estudo. Com o decorrer da pesquisa, idealizamos um critério para seleção de características, baseado em conceitos da teoria de informação, de propósito geral, ou seja, não restrito a apenas algum problema específico.

Após a implementação da nossa técnica, concentramos então a pesquisa em verificar os efeitos dela em problemas de ambas as áreas. Temos aplicado nosso critério para seleção de características em dados simulados, dados reais de bioinformática e em problemas de processamento de imagens envolvendo W-operadores com resultados bastante promissores. Além disso, estudamos alguns métodos propostos na literatura para reconhecimento de padrões na análise de expressões gênicas e percebemos que não há uma técnica bem conhecida que combine seleção de características e teoria da informação que tenha sido aplicada especificamente para resolver problemas dessa área. A mesma observação vale para a construção de W-operadores, ou seja, não se conhece uma técnica de seleção de características utilizando um critério baseado em conceitos de teoria da informação que tenha sido aplicada para este fim.

1.2 Aplicações em bioinformática

Pesquisas em bioinformática, de um modo geral, exigem que sejam aplicadas técnicas de reconhecimento de padrões e seleção de características em diversos contextos como em análise de seqüências de DNA, classificação estrutural de proteínas, diagnóstico de tumores de tecido através da análise de expressões gênicas, identificação de redes de regulação gênica e análise filogenética.

Pode-se destacar a importância de métodos de redução de dimensionalidade [11, 31, 46, 47, 70] para tratar esses problemas. A questão de bioinformática de principal interesse nesse contexto é justamente o de encontrar um conjunto de genes diferencialmente expressos através da análise de expressões gênicas. Dois tipos de problemas são bastante estudados. O primeiro é o de identificação de genes que mudam de estado (por exemplo, super-regulado para sub-regulado e vice-versa). O segundo é o de identificação de padrões de genes que sejam suficientes para caracterizar um determinado fenômeno biológico (por exemplo, câncer em uma determinada região cerebral) ou que sejam suficientes para separar diferentes classes (por exemplo, tecido normal e tecido com tumor).

O primeiro problema é muito mais simples que o segundo, sendo freqüentemente estudado por testes de hipótese estatística ou análise de variância (ver Capítulo 3 em [48]). Algumas repetições do experimento biológico são exigidas para aplicação correta dos procedimentos estatísticos.

O segundo problema foi tratado neste trabalho, sendo usualmente estudado através de técnicas de redução de dimensionalidade. Um procedimento de redução de dimensionalidade tenta encontrar um número mínimo de características que sejam suficientes para separar duas ou mais classes. As características empregadas neste caso são as taxas de expressão gênica requeridas para escolher pequenos subconjuntos de genes (tipicamente três ou quatro) que sejam suficientes para separar classes com fenômenos biológicos distintos (por exemplo, separar tecidos com diferentes tipos de câncer). Normalmente, existe um grande volume de dados de expressão para poucas observações, sendo este fato a principal dificuldade para lidar com este problema.

A identificação de redes de regulação gênica baseada na evolução das expressões dos genes no tempo é uma outra preocupação crucial em bioinformática. A aplicação de técnicas de redução de dimensionalidade neste contexto não é trivial. Porém, um passo anterior ao de identificar redes de regulação gênica é verificar a dependência do valor de um gene com relação aos valores de alguns genes em um instante de tempo anterior. Este passo pode ser modelado como um problema de seleção de características, onde se quer descobrir quais genes são responsáveis pelo valor de um determinado gene num instante posterior.

1.3 Aplicações em processamento de imagens digitais

Considerando uma máscara a ser aplicada em algum processamento de imagens digitais como sendo uma matriz $M \times N$ de valores discretos, seu número de características é justamente o produto de M por N. Devido a essa elevada quantidade de características, em determinados problemas de análise e processamento de imagens, a seleção de características é importante para obter um W-operador [10] com um conjunto ideal de características (pixels) necessárias para realizar uma determinada operação em uma imagem.

W-operadores são usados em praticamente qualquer aplicação de processamento de imagens binárias para redução de ruído, extração de formas e reconhecimento de texturas. Um problema prático e importante é construir W-operadores que realizam essas tarefas em contextos específicos. Há diversas abordagens heurísticas para fazer isso. Uma abordagem formal consiste em estimar o operador ideal a partir de conjuntos de pares de imagens de entrada-saída. Esses pares de imagens compõem os dados de treinamento, que descrevem o

resultado da transformação desejada em algumas imagens típicas do domínio considerado. Tecnicamente, este problema é equivalente à construção de classificadores supervisionados, em reconhecimento estatístico de padrões, ou ao aprendizado de funções Booleanas, em aprendizagem computacional [25]. Este tipo de técnica foi aplicado com sucesso, por exemplo, na indústria de documentos digitais.

Neste trabalho, foram tratadas duas aplicações de processamento de imagens: filtragem de imagens ruidosas e reconhecimento de texturas. Em ambas, foi utilizado o critério proposto nesta dissertação para selecionar os atributos (pixels) da janela W e obter o W-operador construído a partir de W para a realização dessas operações.

1.4 Objetivos

Concentramos a nossa pesquisa em técnicas de redução de dimensionalidade com foco em seleção de características buscando propor uma nova técnica genérica o bastante que pudesse ser aplicada tanto no âmbito da área de bioinformática como na área de processamento de imagens. As heurísticas e as funções critério mais conhecidas e utilizadas para seleção de características também foram estudadas.

Devido ao fato de termos trabalhado com o reconhecimento estatístico de padrões, no qual tanto o vetor de características como os rótulos dos padrões são vistos como variáveis aleatórias, estudamos alguns conceitos da teoria da informação que tem como objetivo principal medir a informação que as variáveis aleatórias fornecem sobre si mesmas ou sobre outras variáveis aleatórias. Apesar de existirem diversos trabalhos propostos na literatura para seleção de características através de alguns desses conceitos, não há nenhum trabalho que tenha explorado tais idéias da maneira desenvolvida nesta dissertação. Consideramos também original a aplicação nos problemas de bioinformática e de processamento de imagens focalizados neste trabalho.

Foi visado também um estudo empírico bastante aprofundado das propriedades da função critério proposta para seleção de características, através de testes exaustivos com dados simulados. Tais propriedades serviram como fundamento para que a aplicação dessa abordagem pudesse ser estendida aos problemas reais tratados nessa pesquisa.

1.5 Contribuições

Realizamos uma revisão bibliográfica sobre reconhecimento de padrões e redução de dimensionalidade com ênfase em seleção de características, técnicas de análise de expressões gênicas, e conceitos básicos da teoria de projeto de W-operadores. Além disso, as principais contribuições deste trabalho foram:

- Implementação de um sistema de identificação e seleção de genes fortes. Esse sistema cuida da formatação e normalização dos dados de expressões gênicas de entrada, os quais podem ter sido produzidos por *microarray* ou por SAGE, para que a etapa principal do sistema, implementada pelo Prof. Dr. Paulo J. S. Silva do IME-USP, analise os dados através de técnicas de máquinas de suporte vetorial (SVM) e de programação linear. Esta etapa devolve uma tabela com os melhores subconjuntos obtidos e informações adicionais como o erro, a distância entre as classes e a freqüência de cada gene nesses subconjuntos. Para os casos em que os subconjuntos considerados são trincas, o sistema exibe a informação do intervalo de credibilidade de cada ponto do espaço e gera um gráfico tridimensional desses pontos para cada uma das trincas selecionadas. Esse trabalho originou-se de uma colaboração com a pesquisadora Helena Brentani do Ludwig Institute for Cancer Research com o objetivo de identificar genes que melhor separam dois tipos distintos de tumores de células com câncer a partir de dados de SAGE. Um artigo em fase de preparação mostrará alguns resultados dessa colaboração (seção 6.2.3 e [8]).

- Introdução do conceito de índice de credibilidade como um critério para avaliar a dispersão das expressões dos genes em dados de SAGE. Tal conceito foi aplicado como um critério de avaliação adicional das trincas de genes devolvidas como solução pela técnica de MSV (seção 6.2.3 e [8]). Esse trabalho foi desenvolvido em conjunto com Ricardo Z. N. Vêncio, aluno de mestrado em estatística do IME-USP.

- Proposta de uma função critério para seleção de características baseada na entropia condicional. Esse critério não privilegia apenas os subespaços de características que separam linearmente as classes como ocorre com a maior parte das funções critérios existentes na literatura. Além disso, a aplicação deste critério é apropriada para os casos nos quais há mais de duas classes distintas possíveis (capítulos 5 e 6).

- Validação dos resultados obtidos pela técnica de MSV em dados de SAGE através do critério de entropia condicional média proposto neste trabalho (seção 6.1.2).

- Metodologia proposta para identificação de redes de regulação gênica a partir dos dados de expressões de *microarray* durante as 48 horas do ciclo de vida do agente causador da malária (*Plasmodium falciparum*) produzidos por Derisi *et al* [15], utilizando entropia condicional média para identificar os possíveis genes preditores do comportamento das expressões de um determinado gene. Alguns resultados preliminares mostram que tal metodologia possui potencial de produzir redes que refletem o conhecimento biológico preexistente e até mesmo gerar conhecimento novo (seção 6.1.2 e [7]). Esse trabalho originou-se de uma colaboração com a equipe do Prof. Hernando Del Portillo do Instituto de Ciências Biomédicas - USP.

- Proposta de uma nova abordagem para construção de W-operadores minimais pela análise da entropia condicional média (seção 6.1.3 e Martins *et al* [53]). Resultados dessa metodologia aplicada em reconhecimento de texturas (seção 6.1.3) serão publicados em um artigo futuro [59].

1.6 Organização do texto

O texto da dissertação divide-se em duas partes principais. A primeira parte versa sobre revisão bibliográfica e conceitos básicos sobre reconhecimento de padrões e redução de dimensionalidade, análise de expressões gênicas e teoria de W-operadores. A segunda parte mostra a técnica proposta para seleção de características, resultados experimentais e conclusões.

O capítulo 2 apresenta uma visão geral sobre a área de reconhecimento de padrões, bem como a importância da redução de dimensionalidade nesse contexto. A seção 2.2 formula o problema de seleção de características, além de introduzir algoritmos clássicos que buscam resolvê-lo. Ainda nessa mesma seção, podem ser encontradas algumas das funções critérios mais popularmente utilizadas. Esse capítulo encerra-se discutindo alguns trabalhos de seleção de características produzidos na literatura nos quais há aplicação de conceitos da teoria da informação (seção 2.3).

No capítulo 3, é apresentada uma revisão sobre as técnicas mais conhecidas e utilizadas

na análise de expressões gênicas. Em particular, a finalização deste capítulo ocorre com a seção 3.4, que faz uma revisão sobre redes de regulação gênica.

O capítulo 4 introduz os conceitos e as propriedades dos W-operadores, bem como os principais desafios encontrados no processo de sua construção.

O capítulo 5 discute em profundidade a nossa proposta de um critério para seleção de características, introduzindo os conceitos necessários para compreendê-lo.

O capítulo 6 mostra experimentos e resultados que ilustram algumas propriedades adicionais interessantes e bastante importantes do critério formulado. Na seção 6.1.1, experimentos aplicados sobre dados sintéticos foram realizados. A seção 6.1.2 mostra experimentos que utilizam entropia condicional média para validar a identificação de genes fortes por MSV. Ainda na seção 6.1.2, é apresentado um resultado preliminar utilizando o critério de entropia condicional média para estimar a arquitetura de uma rede genética probabilística (PGN) em dados de *microarray* de malária. Resultados da aplicação da função critério proposta para construção de W-operadores com o objetivo de realizar filtragem de imagens e reconhecimento de texturas são apresentados na seção 6.1.3. Finalizamos este capítulo com a seção 6.2 que introduz os conceitos gerais do sistema implementado para identificação de genes fortes que separam dois estados biológicos por MSV, incluindo o conceito de índice de credibilidade e alguns resultados que foram validados posteriormente pela nossa metodologia proposta (seção 6.1.2).

Este texto encerra-se com a conclusão deste trabalho (capítulo 7), bem como as considerações sobre os trabalhos em andamento e futuros. A figura 1.1 mostra um esquema de como os capítulos desta dissertação são relacionados entre si.

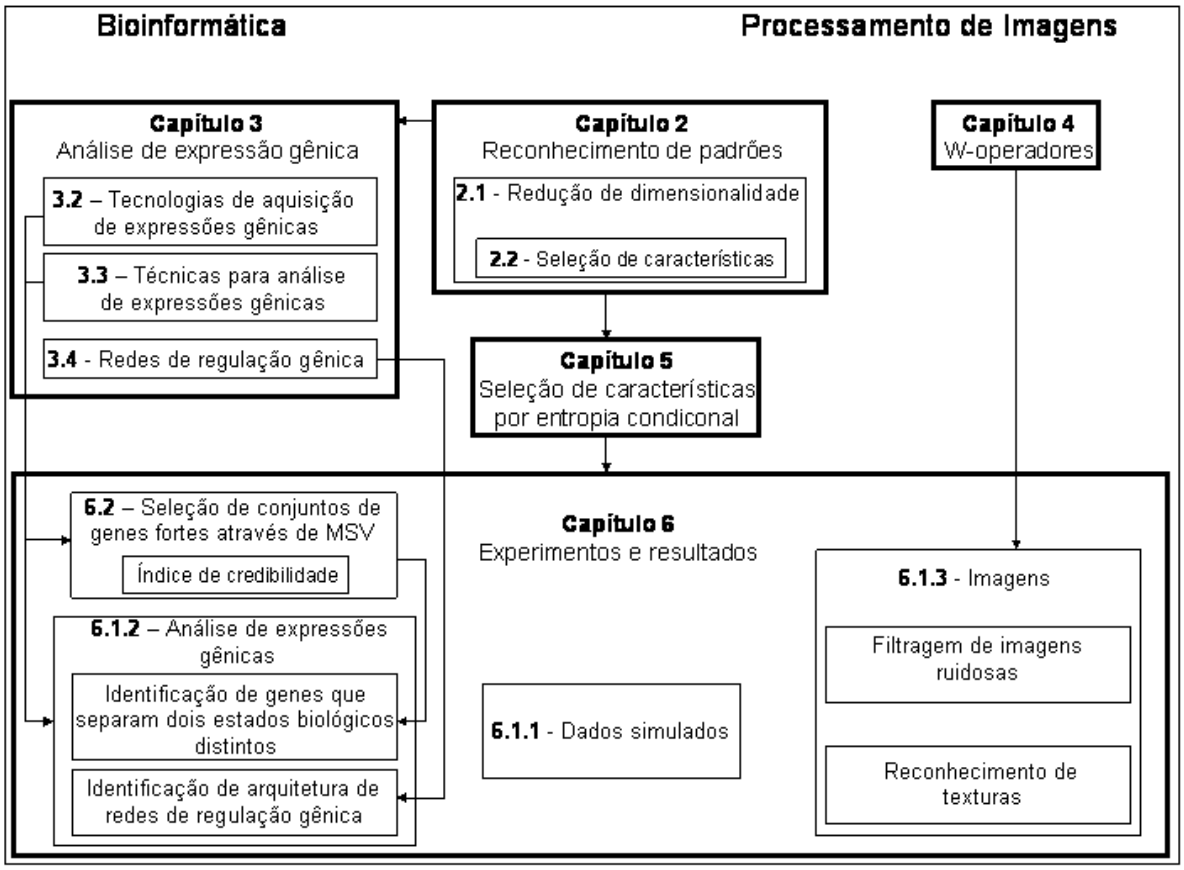


Figura 1.1: Esquema sobre a relação entre os capítulos desta dissertação.

Parte I

Revisão e conceitos básicos

Capítulo 2

Reconhecimento de padrões

2.1 Reconhecimento de padrões e redução de dimensionalidade

Nos últimos 50 anos, a pesquisa em reconhecimento de padrões contribuiu com avanços que possibilitaram aplicações complexas e diversificadas [44]. Dentre essas aplicações, pode-se destacar:

- Bioinformática: análise de seqüências de DNA; análise de dados de expressão gênica (*microarray*, SAGE);
- Mineração de dados (*data mining*): busca por padrões significativos em espaços multi-dimensionais, geralmente obtidos de grandes bancos de dados e "data warehouses";
- Classificação de documentos da Internet;
- Análise de imagens de documentos para reconhecimento de caracteres (*Optical Character Recognition - OCR*);
- Inspeção visual para automação industrial;
- Busca e classificação em base de dados multimídia;
- Reconhecimento biométrico de faces, íris ou impressões digitais;

- Sensoriamento remoto por imagens multispectrais;
- Reconhecimento de fala.

Atribuir um *rótulo* a um determinado *objeto* ou *padrão* é o cerne de reconhecimento de padrões. Inicialmente, temos os objetos do mundo real, sendo desejado particioná-los em *classes* com base em suas respectivas características. Objetos que partilham alguma relação particular entre si são pertencentes à mesma classe, ou seja, possuem um mesmo *rótulo*.

Há diversas abordagens para se realizar reconhecimento de padrões. Dentre elas, este trabalho se encaixa justamente na abordagem estatística, onde cada padrão é representado por um vetor aleatório de n características $\mathbf{X} = (X_1, X_2, ..., X_n)$ [70]. Cada padrão observado $\mathbf{x_i} = (x_1, x_2, ..., x_n)$ é uma amostra de $\mathbf{X}$.

Um sistema de reconhecimento estatístico de padrões é composto principalmente pelos seguintes subsistemas principais [31, 44]:

- sistema de aquisição dos dados, através de sensores ou câmeras, por exemplo;
- sistema de pré-processamento, para eliminar ruídos e normalizar os dados;
- extrator de características, que cria um vetor de características à partir dos dados obtidos;
- sistema de redução de dimensionalidade, onde se analisa o conjunto de características e devolve um outro conjunto contendo apenas algumas das características mais importantes, ou uma combinação de algumas delas;
- classificador, que toma uma certa decisão após a análise de um determinado padrão.

Dado um conjunto de amostras de treinamento, o objetivo principal em reconhecimento de padrões é o de projetar um classificador que infira um determindado rótulo a um novo padrão a partir desse conjunto com a menor margem de erro possível. Se cada uma dessas amostras do conjunto de treinamento já possuír um rótulo associado conhecido, trata-se de *classificação supervisionada*. Existe também a *classificação não-supervisionada* na qual as amostras não possuem rótulo conhecido a priori [74]. Nossa pesquisa tem se concentrado no primeiro tipo de classificação.

Dimensionalidade é o termo atribuído ao número de características utilizadas na representação de *padrões* de objetos, ou seja, à dimensão do vetor $\mathbf{X}$. Reduzir a dimensionalidade significa selecionar um subespaço do espaço de características para representar os padrões. A redução de dimensionalidade faz-se necessária para evitar o *problema da dimensionalidade* [44].

O problema da dimensionalidade ou comportamento da "curva em U" [44] é um fenômeno onde o número de amostras de treinamento exigido para que um classificador tenha um desempenho satisfatório é dado por uma função exponencial da dimensão do espaço de características. Este é o principal motivo pelo qual a realização de redução de dimensionalidade se faz importante em problemas de classificação nos quais os padrões medidos possuem um número elevado de atributos e apenas um número limitado de amostras de treinamento..

A figura 2.1 ilustra o problema da "curva em U". Considere um número de amostras de treinamento fixo. Para dimensões entre zero e m_1, adicionar características implica melhores resultados de classificação, pois o número de características nessa região é insuficiente para separar as classes. Entre m_1 e m_2, a adição de características não diminui significativamente a taxa de erro do classificador, implicando que as características mais importantes já foram inseridas até o ponto m_1. O problema da dimensionalidade ocorre de fato na região posterior a m_2 onde a adição de características piora o desempenho do classificador devido ao número insuficiente de amostras em relação ao número de características.

Existem basicamente duas abordagens para se efetuar redução de dimensionalidade: fusão de características e seleção de características [17]. Os algoritmos de fusão de características criam novas características a partir de transformações ou combinações do conjunto original. Já os algoritmos de seleção selecionam, de acordo com algum critério, o melhor subconjunto de características.

2.2 Seleção de características

Seja $\mathbf{X} = (X_1, X_2, ..., X_n)$ um vetor aleatório denominado *vetor de características*. Seja Y uma variável aleatória denominada *classe* ou *rótulo*. Classificar um *padrão* $\mathbf{x} = (x_1, x_2, ..., x_n)$, isto é, uma amostra de $\mathbf{X}$, é associar a ele um *rótulo* $y \in \{0, 1, ..., c\}$.

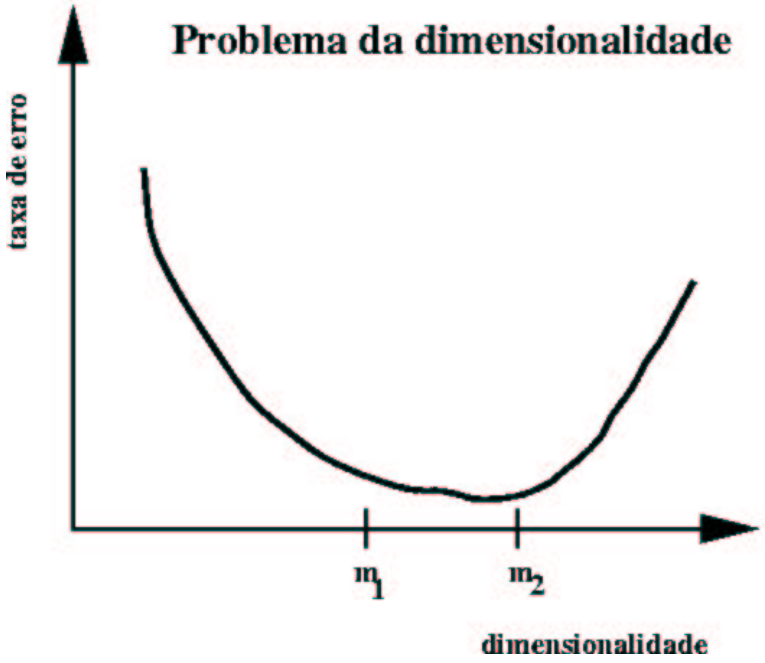


Figura 2.1: Gráfico da taxa de erro em função da dimensionalidade com número fixo de amostras ilustrando o problema da "curva em U".

Em reconhecimento de padrões por classificação supervisionada, dado um conjunto de amostras de treinamento T onde cada amostra é representada pelo par $(\mathbf{x}, y)$, deseja-se obter um bom classificador representado por uma função ψ tal que

$$\psi(\mathbf{x}) = y$$

Selecionar características significa tentar descobrir um subconjunto $\mathcal{Z}$ do conjunto potência $\mathcal{P}(\mathcal{I})$ (conjunto de todos os possíveis subconjuntos de $\mathcal{I}$), onde $\mathcal{I}$ é o conjunto de índices $\mathcal{I} = \{1, 2, ..., n\}$ do espaço total de características, tal que $\mathbf{X}_\mathcal{Z}$ seja um bom subespaço representante de $\mathbf{X}$. Por exemplo, se $\mathcal{Z} = \{1, 3, 5\}$, então $\mathbf{X}_\mathcal{Z} = \mathbf{X}_{\{\mathbf{1},\mathbf{3},\mathbf{5}\}} = \{X_1, X_3, X_5\}$.

Após a seleção de características, projeta-se um classificador ψ baseado em $\mathbf{X}_\mathcal{Z}$ tal que

$$\psi(\mathbf{x}_\mathcal{Z}) = y$$

Como se pode ver, seleção de características é um problema de otimização que, dado

um conjunto de n características, objetiva selecionar um subconjunto de tamanho d ($d \leq n$) que minimiza uma determinada função critério. Ou seja, este problema é resolvido selecionando-se $\mathcal{Z}^* \subseteq \mathcal{I}$ de acordo com a seguinte equação (2.1).

$$\mathcal{Z}^* : \mathcal{F}(\mathbf{X}_{\mathcal{Z}^*}) = min_{\mathcal{Z} \subseteq \mathcal{I}}\{\mathcal{F}(\mathbf{X}_{\mathcal{Z}})\} \tag{2.1}$$

onde $\mathcal{F}(\cdot)$ denota a função critério. Dependendo da função critério, pode ser conveniente maximizá-la ao invés de minimizá-la.

É importante notar que a exploração de todos os elementos de $\mathcal{P}(\mathcal{I})$ solucionaria o problema, mas isto é impraticável em geral. Há algumas heurísticas de busca que tentam obter um conjunto sub-ótimo explorando um espaço de busca muito menor do que o espaço inteiro das combinações.

2.2.1 Algoritmos

Existem diversos algoritmos que realizam seleção de características. Obviamente, o algoritmo de busca exaustiva que testa todos os subespaços possíveis de características é o que sempre devolve a solução ótima, porém sua complexidade de tempo é proibitiva em casos práticos. O algoritmo chamado *branch-and-bound* proposto em [54] devolve a solução ótima em situações nas quais a função critério é monotônica, ou seja, se $\mathcal{F}(\mathbf{Z_i} \cup \mathbf{Z_j}) \leq \mathcal{F}(\mathbf{Z_i})$ (ou $\mathcal{F}(\mathbf{Z_i} \cup \mathbf{Z_j}) \geq \mathcal{F}(\mathbf{Z_i})$) para todo $\mathbf{Z_i}, \mathbf{Z_2} \subseteq \mathbf{X}$. Mas no pior caso, o algoritmo explora todas as configurações, tendo portanto complexidade exponencial.

Os algoritmos sub-ótimos não garantem que a solução ideal seja devolvida, mas são eficientes. Existem 3 tipos principais de algoritmos sub-ótimos: determinísticos com solução única, determinísticos de múltiplas soluções e estocásticos de múltiplas soluções.

Os métodos determinísticos de múltiplas soluções devolvem inúmeros conjuntos solução, porém ao contrário dos estocásticos, devolvem sempre os mesmos conjuntos solução para os mesmos dados de entrada. Já os métodos estocásticos de múltiplas soluções, além de devolver diversos conjuntos de características, duas execuções desses métodos aplicados aos mesmos dados de entrada podem devolver diferentes conjuntos solução.

Neste trabalho, focalizamos os métodos determinísticos com solução única, isto é, que devolvem uma única solução que é sempre a mesma para toda execução sobre os mesmos

dados de entrada. Alguns dos algoritmos mais importantes dessa classe são descritos abaixo.

Melhores Características Individuais

Este algoritmo analisa cada característica individualmente e seleciona as d melhores [45, 70].

MCI($\mathbf{X}$, d)
$\mathbf{Z} \leftarrow \phi$
ENQUANTO $|\mathbf{Z}| < d$ FAÇA
  $\mathbf{Z} \leftarrow \mathbf{Z} \cup \{X_i : \mathcal{F}(X_i) = min_{1 \leq j \leq n} \mathcal{F}(X_j), \forall X_j \notin \mathbf{Z}\}$
DEVOLVA $\mathbf{Z}$

Tal método é simples computacionalmente mas não garante que o melhor subconjunto seja encontrado, pois duas características boas individualmente podem formar um subconjunto ruim quando associadas entre si.

Busca Seqüencial para Frente (SFS)

O princípio da Busca Seqüencial para Frente (do inglês: SFS - Sequential Forward Search) é relativamente simples: dado um espaço $\mathbf{Z}$ inicialmente nulo, procure uma característica X_j tal que $\mathbf{Z} \cup \{X_j\}$ seja o melhor espaço dentre todos os espaços $\mathbf{Z} \cup \{X_i\}, 1 \leq i \leq n$, onde n é o número de características do espaço inteiro $\mathbf{X}$. Adicione essa característica a $\mathbf{Z}$ e repita o mesmo procedimento para esse novo conjunto. Esse procedimento é executado d vezes [45, 70].

SFS($\mathbf{X}$, $\mathbf{Z}$, d)
ENQUANTO $|\mathbf{Z}| < d$ FAÇA
  $\mathbf{Z} \leftarrow \mathbf{Z} \cup \{X_i : \mathcal{F}(\mathbf{Z} \cup X_i) = min_{1 \leq j \leq n} \mathcal{F}(\mathbf{Z} \cup X_j), \forall X_j \notin \mathbf{Z}\}$
DEVOLVA $\mathbf{Z}$

Em geral, este método obtém resultados melhores que o método anterior, mas não garante que a solução seja ótima, devido ao efeito *nesting*. Tal efeito ocorre quando o subconjunto ótimo não contém os elementos do conjunto selecionado, devido ao fato de não ser possível retirar características por este método. Sua vantagem é ser eficiente com-

putacionalmente quando se deseja obter um pequeno subconjunto em relação ao espaço total de características.

Busca Seqüencial para Trás (SBS)

O método de Busca Seqüencial para Trás (do inglês: SBS - Sequential Backward Search) é análogo ao SFS, mas ao invés de colocar, retira características do espaço $\mathbf{Z}$ inicialmente igual ao espaço total das características $\mathbf{X}$ [45, 70].

SBS($\mathbf{X}$, $\mathbf{Z}$, d)
ENQUANTO $|\mathbf{Z}| > d$ FAÇA
  $\mathbf{Z} \leftarrow \mathbf{Z} - \{X_i : \mathcal{F}(\mathbf{Z} - X_i) = min_{1 \leq j \leq n} \mathcal{F}(\mathbf{Z} - X_j), \forall X_j \in \mathbf{Z}\}$
DEVOLVA Z

Da mesma forma que no SFS, este método não evita o efeito *nesting*, já que é possível que sejam eliminadas do subconjunto solução características que pertençam ao subconjunto ótimo. O SBS é eficiente computacionalmente quando se deseja obter um subconjunto de tamanho próximo ao tamanho do espaço total de características.

Algoritmos de busca seqüencial generalizada (GSFS e GSBS)

São algoritmos que inserem (GSFS) ou removem (GSBS) subconjuntos de características, ao invés de fazerem com apenas uma por vez [67, 70].

Mais l - menos r (PTA)

Este método (do inglês: PTA - Plus l - Take Away r) foi criado com o objetivo de amenizar o efeito *nesting* [67, 70]. O algoritmo aumenta o conjunto de características em l elementos usando SFS, e depois elimina r características usando SBS. Os valores l e r são parâmetros a serem definidos pelo usuário.

Algoritmos de busca seqüencial flutuante

Tanto o algoritmo de busca para frente (SFFS) quanto o de busca para trás (SFBS) são generalizações do método *mais l - menos r*, em que os valores de l e r são determina-

dos e atualizados dinamicamente [57, 70]. O fluxograma da figura 2.2 esquematiza o funcionamento do SFFS.

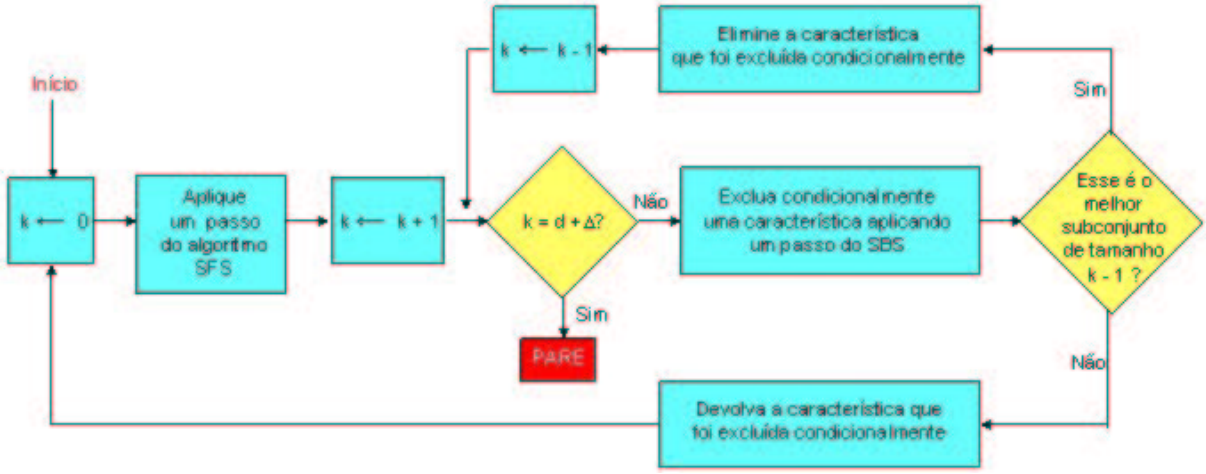


Figura 2.2: Fluxograma simplificado do algoritmo SFFS. Adaptado de [45]. Normalmente utiliza-se $\Delta = 3$ ($k = d + 3$ como critério de parada).

Esses métodos são computacionalmente eficientes e produzem soluções muito próximas da solução ótima. São métodos que melhor combinam tempo de execução com qualidade dos resultados [44, 45].

Existem também os métodos de busca seqüencial flutuante adaptativos (ASFFS e ASFBS) que são generalizações onde conjuntos de características podem ser inseridos por vez [67]. Esses conjuntos têm seu tamanho e seu conteúdo determinados dinamicamente. Tais métodos produzem resultados um pouco melhores, porém são muito mais lentos e complexos que os não adaptativos [19].

2.2.2 Funções critério

Em algoritmos de seleção de características, a *função critério* é uma parte fundamental. O objetivo de uma função critério é o de selecionar subconjuntos de características que separem bem as classes, de maneira a facilitar o trabalho do classificador. Discutimos, a seguir, algumas funções popularmente usadas.

Desempenho do classificador

Um critério bastante utilizado é o do erro de classificação. Quando não se sabe a distribuição dos dados, utiliza-se os padrões de treinamento e de teste no espaço determinado pelo conjunto de características para avaliar o desempenho de um classificador [17]. Quanto menor o erro, melhor é o conjunto de características.

É importante tomar o cuidado de não estimar a probabilidade do erro do classificador após a seleção de características com base no conjunto de treinamento e de testes utilizado no processo de seleção. Caso contrário, o classificador será ajustado especificamente para o conjunto de padrões utilizado em seu projeto, e a estimativa da probabilidade de erro será muito otimista [17].

Distâncias entre classes

Há vários critérios que utilizam distâncias entre padrões de classes diferentes no espaço de características a partir de um conjunto de treinamento. Dentre as principais, temos [17, 70]:

- *Distância entre os centróides das classes:* para calcular essa medida, basta determinar os centróides das classes e medir a distância entre eles.
- *Distância entre vizinhos mais próximos, mais distantes e média:* no cálculo dessas distâncias, devemos considerar, respectivamente, o mínimo, o máximo ou a média das distâncias entre os padrões de treinamento de duas classes diferentes.
- *Distância baseada em matrizes de espalhamento:* utilizam medidas de separabilidade baseadas em análise de discriminantes.
- *Distância de Mahalanobis:* utilizada para medir a distância entre classes de padrões.
- *Distância de Bhattacharyya e divergência:* baseia-se nas funções densidade de probabilidade das classes, de forma que a distância espacial entre os conjuntos não seja considerada, mas sim a diferença entre a forma deles.
- *Distâncias nebulosas:* medidas que utilizam informações obtidas a partir da *fuzzyfucação* entre conjuntos (transformação dos conjuntos de treinamento em conjuntos

nebulosos), como os suportes dos conjuntos e os coeficientes de pertinência dos padrões [12, 18].

A maior parte das funções critério baseadas em distância tendem a privilegiar características que deixem as classes linearmente separáveis (fig. 2.3). Existem casos onde um subespaço de características é considerado um bom separador, mesmo que ele não deixe as classes linearmente separáveis. Exemplos disso estão ilustrados na figura 2.4. Um outro problema é que tais critérios ficam restritos apenas a encontrar subespaços de características que separam duas classes, embora na maior parte dos problemas de reconhecimento de padrões, existam mais de duas classes possíveis.

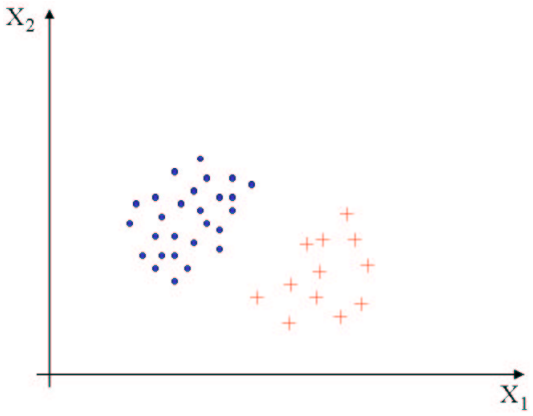


Figura 2.3: Classes linearmente separáveis.

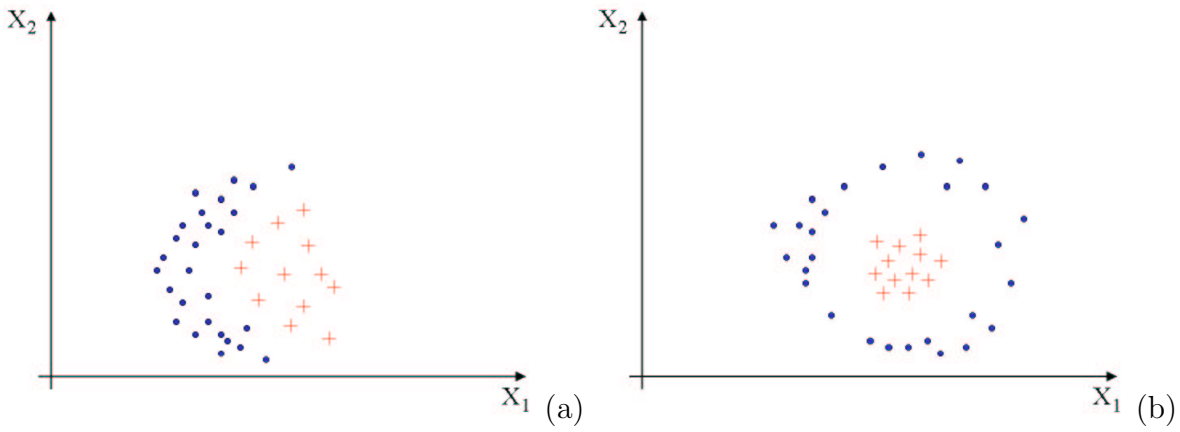


Figura 2.4: (a) classes côncavas entre si; (b) classe interna à outra.

Para contornar esses problemas, propomos um critério para seleção de características que não se baseia na geometria dos pontos formados pelos padrões no espaço, mas sim no grau de informação que um determinado subespaço de características fornece com relação ao comportamento da variável de classe, independente do número de valores distintos que esta variável possa assumir. Esse critério é baseado em princípios de teoria estatística como *entropia* e *informação mútua* (ver capítulo 5). Para uma breve revisão bibliográfica sobre seleção de características através de teoria da informação, ver a próxima seção (seção 2.3). Tal critério escolhe o melhor subespaço de características de acordo com a distribuição de probabilidades condicionais entre suas instâncias e as classes, sendo independente do erro do classificador.

2.3 Seleção de características através de teoria da informação

A teoria da informação, também conhecida como teoria matemática da comunicação, foi formulada por Claude Shannon em [62]. Desde então, vem sendo utilizada não só em pesquisas de telecomunicações, como também em diversos outros contextos. Particularmente com relação ao reconhecimento de padrões, nos últimos anos os pesquisadores da área tem dado cada vez mais importância aos conceitos dessa teoria como informação mútua e entropia.

A informação mútua (ver definição no capítulo 5) é uma medida bastante usada para análise de dependência estocástica de variáveis aleatórias discretas [22, 49, 68]. Ela é aplicada em seleção de características para identificar subespaços que predizem o valor da classe [31, 39].

Em seu trabalho de categorização de textos, Lewis examinou o impacto do tamanho do espaço de características na eficiência da categorização, ordenando cada característica (palavra) pela sua informação mútua com as possíveis categorias. As primeiras d características foram escolhidas para compor o espaço de características, e diferentes valores de d foram examinados [50].

Bonnlander e Weigend propuseram um método para identificar e eliminar as variáveis de entrada de uma rede neural irrelevantes para predizer os valores das variáveis de saída.

Este método aplica informação mútua como medida de relevância das variáveis de entrada [14].

Viola propôs um novo procedimento para avaliar e manipular a entropia empírica de uma distribuição (EMMA - "EMpirical entropy Manipulation and Analysis") [72]. EMMA pode ser usado para encontrar projeções de pequena dimensionalidade altamente informativas em um espaço de grande dimensionalidade.

Zaffalon e Hutter utilizaram informação mútua a fim de obter uma robusta seleção de características [77]. Definiram dois tipos de filtro: o *forward filter* (FF) e o *backward filter* (BF). De acordo com um determinado limiar ϵ, uma característica é incluída no espaço de características se sua informação mútua com a variável de classe for maior ou igual a ϵ no caso do filtro FF, ou excluída se sua informação mútua for menor ou igual a ϵ no caso do filtro BF.

Fleuret criou um método que seleciona características que maximizam sua informação mútua com a classe, condicionalmente às características já escolhidas [35]. Este critério, chamado de Maximização da Informação Mútua Condicional (CMIM), não seleciona características que sejam similares às já selecionadas, mesmo que elas sejam muito boas individualmente, porque elas não carregam qualquer informação adicional sobre a classe.

Capítulo 3

Análise de expressão gênica

3.1 Introdução

A taxa de produção de dados de expressões gênicas tem crescido de maneira explosiva nos últimos tempos. Devido a isso, releva-se a necessidade de métodos automáticos que auxiliem os cientistas a organizar, destrinchar e entender essa verdadeira "montanha"de dados para a geração de conhecimento biológico. Alguns dos principais problemas dessa área são a análise de dados de níveis de expressão gênica, seqüenciamento de genes, proteínas e aminoácidos [51].

Genes são elementos importantes dos sistemas de controle de organismos, constituindo um tipo de rede de comunicação que processa informação biológica e regula vias metabólicas de células (fig. 3.1). Eles apresentam a propriedade de se expressar, criando cópias de segmentos de DNA na forma de RNA mensageiro (mRNA). Esses RNA mensageiros atravessam orifícios do núcleo celular para o citoplasma, onde são traduzidos em seqüências de proteínas que constroem enzimas que catalisam reações metabólicas ou voltam ao núcleo para interagir com o DNA e regular a síntese de mRNA [6].

3.2 Tecnologias de aquisição de expressões gênicas

Atualmente existem tecnologias que permitem medir fenômenos moleculares como decodificação de DNA, descrição de estruturas proteicas, estimação de concentração de mRNA

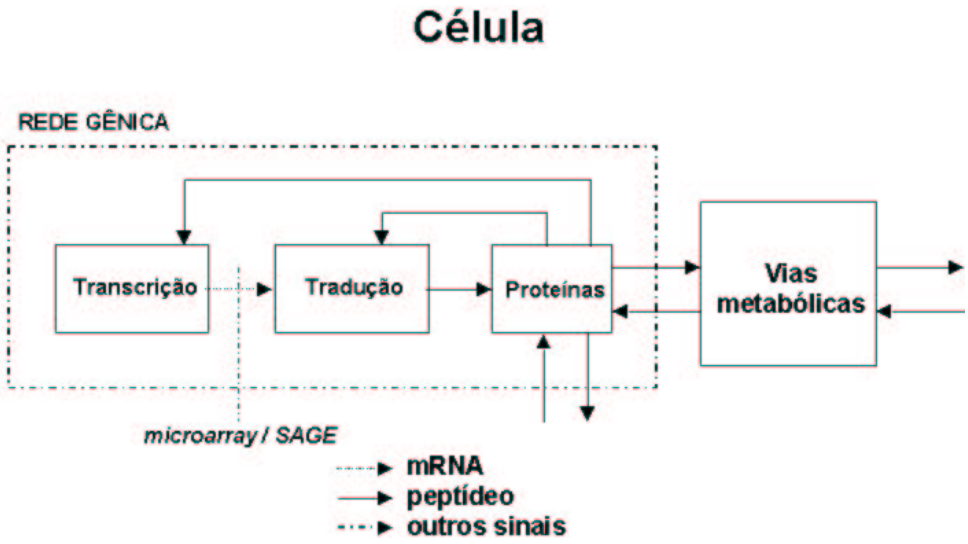


Figura 3.1: Dinâmica da célula (adaptada de [6]).

e proteínas nas células. As tecnologias mais conhecidas e amplamente utilizadas para medir concentrações de mRNA (isto é, expressões gênicas) são: *microarray*, *oligonucleotide chips*, RT-PCR e SAGE (Serial Analysis of Gene Expression). Nossa pesquisa tem lidado com matrizes de expressão produzidas por *microarray* e por SAGE. Ambas produzem dados que podem ser representados na forma de matrizes $m \times n$ de valores reais de expressão, onde cada linha da matriz é uma amostra (chip de *microarray* ou biblioteca de SAGE) e cada coluna representa um gene (figura 3.2). Nesta seção veremos uma breve revisão do processo de obtenção dessas matrizes por essas duas tecnologias.

3.2.1 *Microarray*

A introdução da técnica de cDNA *microarray* revolucionou o campo da biotecnologia, pois aumentou drasticamente a capacidade de medição dos fenômenos relacionados às expressões gênicas, além de permitir modelos quantitativos para esses fenômenos. Uma das primeiras experiências bem sucedidas utilizando essa tecnologia foi o mapeamento do genoma da levedura em 1996 [61]. Depois disso, várias aplicações em diagnósticos vêm sendo desenvolvidas, consolidando essa técnica como promissora ferramenta em biologia molecular para o presente e para o futuro [34].

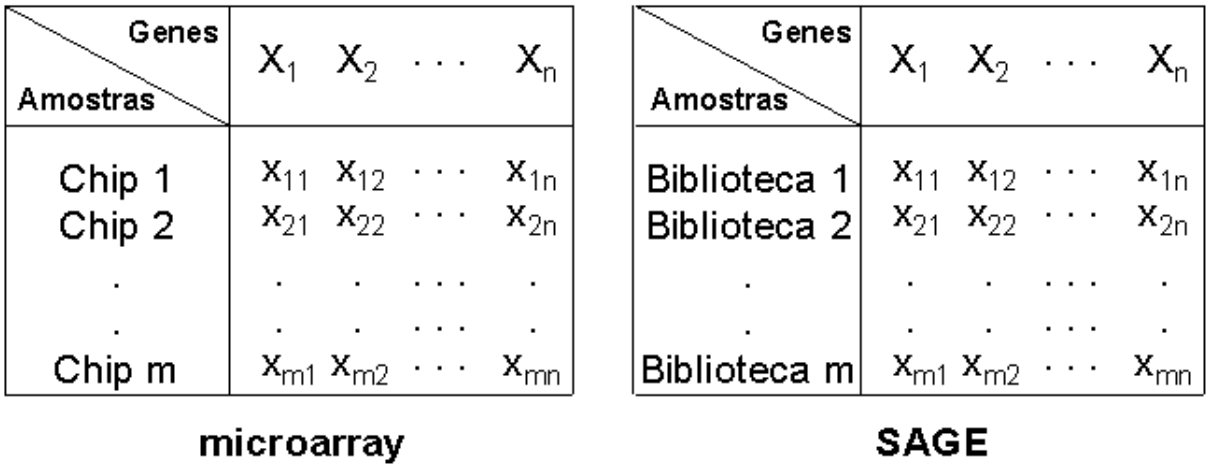


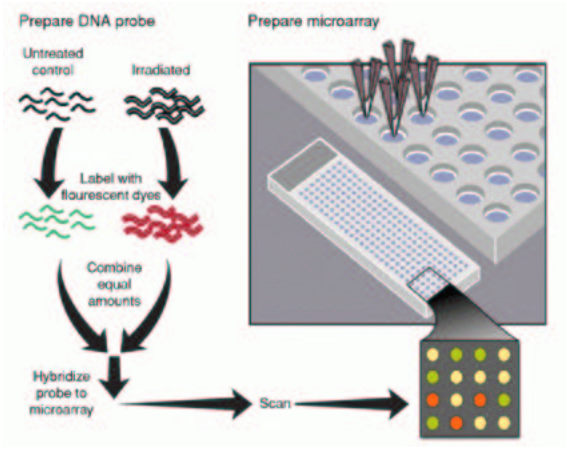
Figura 3.2: Matrizes de expressão obtidas de *microarray* e SAGE.

A figura 3.3.a[1] esquematiza o processo de obtenção da imagem de *microarray*. O processo se inicia com a utilização de um braço mecânico de alta precisão que deposita pequenas quantidades de DNA em uma lâmina de vidro ou de nailon denominadas *chips*. Em seguida utiliza-se como experimento dois mRNAs cultivados em condições distintas. Uma delas será submetida a uma transcrição reversa Cy3 (pigmento fluorescente verde) e a outra sofrerá uma transcrição reversa Cy5 (pigmento fluorescente vermelho). Então une-se os dois mRNAs formando um cDNA que será hibridizado (misturado) com cada uma das seqüências de DNA misturadas na lâmina formando os *spots*. Cada um dos spots se expressará de tal forma a emitir uma fluorescência que é captada por um microscópio (scanner) específico para tal tarefa. Finalmente, a matriz de expressões é obtida através da análise das imagens (fig. 3.3.b[2]) fornecidas pelo microscópio para subseqüente análise dos dados [42].

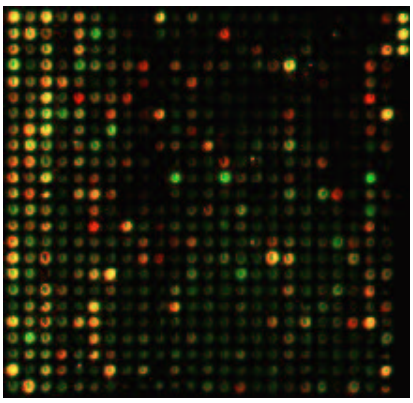
3.2.2 SAGE

Serial Analysis of Gene Expression (SAGE) [71] é uma técnica que permite uma análise rápida e detalhada de milhares de mRNA transcritos em uma célula. Esta técnica possui dois princípios básicos:

[1] retirada do site `http://www.llnl.gov/str/JulAug03/Wyrobek.html`
[2] retirada do site `http://www.gene-chips.com/sample1.html`



(a)



(b)

Figura 3.3: (a) processo de obtenção da imagem de *microarray*; (b) exemplo de imagem de *microarray*.

- uma pequena seqüência de nucleotídeos, denominada "tag", pode identificar o transcrito original de onde foi retirado;

- a ligação dessas *tags* permite uma rápida análise de seqüenciamento de multiplos transcritos.

A figura 3.4[3] esquematiza com maior grau de detalhe o processo envolvido no SAGE. Em síntese, o processo dá início através da criação do cDNA a partir de mRNA. Uma seqüência de pares de bases (10-17) é retirada de um local específico de cada cDNA, formando uma *tag* (etiqueta), servindo como um identificador. Então as *tags* são unidas em uma longa dupla fita de DNA que pode ser amplificada e seqüenciada.

Uma vantagem deste método é que a seqüência de mRNA não precisa ser conhecida, podendo portanto detectar genes desconhecidos anteriormente. Porém, ele é um pouco mais complexo, exigindo uma maior quantidade de seqüenciamento [23].

SAGE foi utilizado para analisar o conjunto de genes expressos durante três diferentes fases do ciclo celular da levedura, além de ter sido usado para monitorar a expressão de cerca de 45 mil genes humanos em células normais de cólon, tumores de cólon e tumores pancreáticos [23].

[3]retirada do site http://www.bioteach.ubc.ca/MolecularBiology/PainlessGeneExpressionProfiling

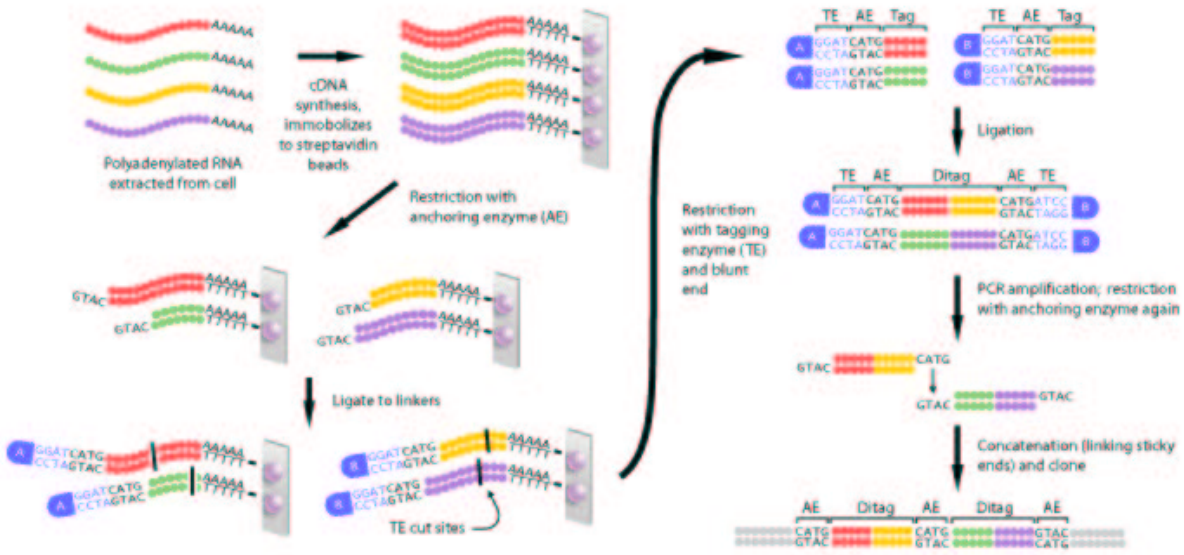


Figura 3.4: Etapas envolvidas no SAGE.

3.3 Técnicas para análise de expressões gênicas

Esta seção apresenta uma síntese das principais metodologias de reconhecimento de padrões propostas para analisar expressões gênicas. Grande parte dessa revisão também pode ser encontrada no artigo de Li *et al* [51].

3.3.1 Fold ("Dobra")

Este é um dos métodos mais simples. Neste método, se o nível de expressão de um gene foi alterado por um número predeterminado de "dobras" (2, 3 ou 4), dizemos que ele mudou seu próprio estado (ligado para desligado ou vice-versa) devido ao tratamento. Um problema com este método é que dificilmente ele revela a correlação desejada entre os dados de expressão e a função biológica, já que o fator predeterminado de dobras tem diferentes significados dependendo dos níveis de expressão de vários genes. Um outro problema é que esta técnica analisa os genes individualmente, não levando em conta a rede de regulação que existe entre eles.

3.3.2 Teste-T

Este é um outro método simples para análise de expressão gênica. Ele manipula o logaritmo dos níveis de expressão, e requer comparação com a média e a variância dos grupos de tratamento e de controle. Uma desvantagem desta técnica é que ela requer inúmeros experimentos de tratamento e de controle para produzir resultados confiáveis [5, 55].

3.3.3 Análise de Componentes Principais (PCA)

O uso desta técnica na análise de expressões gênicas foi sugerida por diversos pesquisadores. PCA (do inglês: Principal Component Analysis) é uma técnica matemática linear que encontra bases vetoriais que expandem o espaço de dados (espaço das expressões gênicas). Uma componente principal pode ser encarada como um padrão principal no conjunto de dados. Quanto mais componentes principais forem utilizadas para modelar o espaço, melhor será a representação. Na maior parte dos casos, PCA reduz a dimensionalidade do espaço e a necessidade de eliminação de ruído sem muita perda de generalidade ou informação.

Uma vantagem do PCA á a facilidade de utilizar e entender seu algoritmo. Sua desvantagem é que ele só funciona bem para problemas onde os dados de expressão sejam linearmente separáveis devido a sua natureza linear. PCA é uma poderosa técnica quando combinada com uma outra técnica de classificação como agrupamento k-médias (seção 3.3.4) ou mapas auto-organizáveis.

3.3.4 Agrupamento k-médias

Esta técnica é uma abordagem de agrupamento (do inglês: clustering) divisível. Particiona-se os dados (genes ou experimentos) em grupos que tem padrões de expressão similares. k é o número de grupos (clusters) definidos pelo usuário. Este algoritmo é fácil de implementar, mas o parâmetro k dificilmente é conhecido de antemão, freqüentemente envolvendo um processo de tentativa e erro.

3.3.5 Agrupamento hierárquico

Há inúmeras variantes de algoritmos de agrupamento hierárquico que podem ser aplicadas para análise de expressões. Dentre elas estão: *single-linkage, complete-linkage, average-linkage, weighted pair-group averaging* e *within pair-group averaging* [2, 33, 58, 75]. Esses algoritmos diferem apenas na maneira com que as distâncias são calculadas entre os grupos crescentes e os elementos restantes no conjunto de dados. Eles costumam gerar uma pontuação de similaridade (*score*) para todas as combinações dos genes, armazenando as pontuações numa matriz para finalmente unir aqueles genes que possuem a maior pontuação, continuando a unir progressivamente menos pares de similaridade.

No processo de agrupamento, após o cômputo da matriz de similaridade, os pares mais relacionados são identificados na parte de cima da diagonal da matriz. Um nó na hierarquia é criado para o par de maior pontuação, tira-se a média dos perfis de expressão dos dois genes, e os elementos unidos são ponderados pelo número de elementos que eles contêm. A matriz então é atualizada, substituindo os elementos unidos pelo nó. Para n genes, o processo é repetido $n - 1$ vezes até restar um único elemento que contém todos os genes.

Wen *et al* [75] usaram agrupamento e técnicas de mineração de dados (data mining) para analisar dados de expressão. Eles mostraram como a integração dos resultados que foram obtidos de várias métricas de distância pode revelar diferentes, porém significativos padrões nos dados. Eisen *et al* [33] também demonstraram o poder do agrupamento hierárquico na análise de expressões.

As vantagens do agrupamento hierárquico são sua simplicidade e o fato de produzir resultados mais fáceis de serem visualizados e interpretados do que os gerados pelo algoritmo k-médias. Porém existem duas desvantagens. A primeira é que devido a sua natureza gulosa, pequenos erros cometidos nos estágios iniciais do algoritmo tendem a se propagar problematicamente durante o processo [60], pois não tem a propriedade de voltar e reparar certos erros (backtracking). Uma segunda desvantagem é que a medida que os grupos crescem, o vetor que representa o grupo pode não representar mais nenhum dos genes do grupo, tornando os padrões de expressão dos genes menos relevantes [58]. Algumas técnicas híbridas vêm sendo criadas para tentar descobrir o momento certo para o algoritmo parar de juntar elementos.

3.3.6 Modelos de mistura e maximização da esperança (EM)

Modelos de mistura são modelos probabilísticos construídos pelo uso de combinações convexas positivas de distribuições tomadas de uma família de distribuições [4, 37]. O algoritmo EM é um algoritmo iterarivo que procede em dois passos alternativos, o passo E (esperança) e o passo M (maximização). A aplicação do algoritmo EM ao correspondente modelo de mistura pode servir como uma análise complementar à do agrupamento hierárquico. Isto porque o modelo de mistura dá uma pista de qual é o número verdadeiro de grupos distintos, uma importante questão para os biólogos no estudo das expressões. O problema é que geralmente o número de amostras para alguns grupos é muito pequeno para estimar os parâmetros do modelo de mistura.

3.3.7 Gene Shaving

Gene shaving é um método estatístico para descoberta de padrões em dados de expressão gênica. O algoritmo original utiliza PCA [41]. Alguns métodos substituem o PCA por uma variedade não-linear. *Gene shaving* busca identificar grupos de genes com padrões de expressão coerentes e grande variação nas amostras. Tal propriedade é importante, sendo uma vantagem que não existe em simples métodos de agrupamento. Os autores que propuseram este método analisaram expressões de pacientes com uma grande e difusa B-célula lymphoma e identificaram um pequeno grupo de genes cuja expressão é altamente preditiva de sobrevivência.

Existem duas variedades do algoritmo original: supervisionado (ou parcialmente supervisionado) e não supervisionado. No supervisionado e no parcialmente supervisionado, informações disponíveis sobre os genes e as amostras podem ser utilizadas para rotular seus dados como um meio de influenciar o processo de agrupamento e garantir grupos significativos. *Gene shaving* permite que genes possam pertencer a mais de um grupo. Essas duas propriedades fazem com que o *gene shaving* seja uma técnica bastante diferente da maior parte dos agrupamentos hierárquicos e de outros métodos para análise de dados de expressão.

A desvantagem deste método é a necessidade de um esforço computacional muito grande devido a repetitiva computação da maior componente principal para um grande conjunto de variáveis.

3.3.8 Máquinas de Suporte Vetorial (MSV)

Máquinas de Suporte Vetorial (MSV) (do inglês: Support Vector Machines) consistem em uma técnica de classificação supervisionada, pois os vetores são classificados com base em vetores de referência conhecidos. MSV resolve o problema de mapear os vetores de expressão gênica do espaço de expressões em um espaço de características de maior dimensão, em que a distância é medida usando uma função matemática conhecida como uma função núcleo (*kernel*), e então os dados podem ser separados em duas classes [58]. Vetores de expressão podem ser vistos como pontos em um espaço n-dimensional. Em análise de expressões, conjuntos de genes são identificados para representar um padrão alvo de expressão. A MSV é então treinada para distinguir os pontos que representam o padrão alvo (pontos positivos no espaço de características) dos pontos que não o representam (pontos negativos).

Com um espaço de características apropriadamente escolhido de dimensionalidade suficiente, qualquer conjunto de treinamento consistente pode ser separável [16]. MSV é uma técnica linear que usa hiperplanos que separam superfícies entre pontos positivos e negativos no espaço de características. Especificamente, MSV seleciona o hiperplano que provê a maior margem entre a superfície do plano e os pontos positivos e negativos, buscando evitar ao máximo a mistura entre os dois conjuntos.

A técnica de MSV pode ser considerada como não linear por causa da possibilidade de mapear fronteira linear no espaço de características para fronteira não-linear no espaço das expressões gênicas. A grande vantagem do MSV é que ele oferece a possibilidade de treinar classificadores não-lineares generalizáveis em um espaço de alta dimensão utilizando um pequeno conjunto de treinamento. Para conjuntos de treinamento maiores, MSV tipicamente seleciona um pequeno conjunto que é suficiente para construir o classificador, minimizando assim um esforço computacional durante os testes.

Uma das desvantagens desta técnica é que se a função núcleo, os parâmetros e as penalidades não forem escolhidas adequadamente, MSV pode não ser capaz de encontrar um hiperplano separador [16].

Na seção 6.2, ilustramos uma técnica recentemente implementada pelo Prof. Dr. Paulo J. S. Silva do IME-USP que utiliza MSV para selecionar conjuntos de *genes fortes*, isto é, genes que resistem ao máximo às margens de erro nas medidas de expressão gênica

[47]. Além disso, na seção 6.1.2 será mostrado como nossa técnica baseada em entropia condicional foi útil para corroborar os resultados produzidos por MSV.

3.4 Redes de regulação gênica

Serão expostas aqui teorias e aplicações sobre modelos de redes de regulação gênica que já foram desenvolvidas, como uma espécie de revisão bibliográfica referente a esse assunto. Essas informações podem ser encontradas também no sétimo tópico de [6].

A arquiterura de uma rede consiste de ligações representando conexões biomoleculares e regras ou funções que representam as interações moleculares na célula [24]. Em síntese, a arquitetura define como cada variável depende das outras. Essas variáveis podem representar valores de atividade molecular como, por exemplo, níveis de expressão de genes, em um dado modelo.

A dinâmica de uma rede reflete a evolução das variáveis no tempo. O estado da rede é o valor dessas variáveis de estado em um dado instante de tempo. Uma trajetória é uma seqüência de transições de estado. Quando o sistema volta para um estado ocorrido anteriormente, em um modelo determinístico, ele seguirá um mesmo ciclo de estados para sempre. Esse fenômeno é denominado atrator, sendo o resultado final do sistema [76].

A atividade da célula é determinado pela expressão dos genes (níveis de mRNA transcritos). Esses níveis também indicam a quantidade de proteína disponível na célula em um determinado instante de tempo. Com múltiplas amostras, podemos observar a disponibilidade de mRNA ao longo do tempo ou sob diferentes condições (como em resposta a estímulos, falta de recursos, evolução de câncer, tecidos normais, tecidos doentes, etc). Grande parte dos estudos tomam níveis de mRNA como variáveis de estado por eles serem fáceis de medir.

Uma rede de regulação gênica pode ser complexa o suficiente a ponto de tornar a sua reconstrução um problema difícil. Dados N genes, há um número exponencial de possíveis interações entre os genes. Este é o problema da dimensionalidade como discutido na seção 2.1. O problema da dimensionalidade torna as tarefas de modelagem, identificação e simulação de redes mais difíceis. A seleção cuidadosa das variáveis de entrada, sem perder nenhuma das importantes, e o uso de informação a priori sobre o que é conhecido

biologicamente podem ser cruciais para lidar com este problema.

3.4.1 Modelos de redes gênicas

Dentre os diversos modelos propostos de redes de regulação gênica, pode-se destacar [23]:

- Determinístico ou Estocástico
- Discreto ou Contínuo

Uma rede determinística é um sistema rígido em que os níveis de expressão de todos os genes em um dado instante de tempo e as interações regulatórias entre eles determinam univocamente o estado das expressões gênicas no próximo estado [69]. Já em um sistema estocástico, um estado de expressões gênicas pode gerar mais de um estado de expressões.

A natureza de uma rede de regulação gênica é, em geral, estocástica. Alguns mecanismos que explicam tal natureza são, por exemplo, degradação de produtos de um gene, a colisão espacial necessária antes de um reagente poder exercer sua influência, equações de reações reversíveis, além de outras [32]. Conseqüentemente, modelos estocásticos descrevem melhor a dinâmica de uma rede de regulação gênica do que modelos determinísticos, apesar dos primeiros serem mais complexos.

Rede Booleana é a rede discreta mais simples de todas, consistindo de n nós representando os genes, que podem ser expressos ou não expressos (estados 0 ou 1). A dinâmica da rede é determinada por n funções, uma para cada nó, que recebe entrada de k nós e determina o próximo estado para aquele nó dos estados de todos os nós de entrada [52, 1].

Obviamente, redes Booleanas são simplificações de redes genéticas, já que os valores das expressões gênicas são contínuos ao invés de discretos. Contudo, genes podem ter complexas interações em redes Booleanas, apresentando um comportamento comparável às redes biológicas. Elas podem ser um bom ponto de partida para modelagem realista de redes gênicas [3].

Um possível refinamento de redes Booleana é a aleatória, onde cada nó pode ter um esquema de entrada e saída diferente e um número diferente de entradas. Cada nó também pode ter diferentes funções Booleanas escolhidas ao acaso. Com isso, uma rede Booleana aleatória é mais realista do que uma Booleana.

Redes gênicas também podem ser modeladas por um conjunto de equações diferenciais não lineares [73]. Parâmetros que indicam a taxa de alteração de cada gene devem ser encontrados. Eles são: $n \times n$ pesos de interação gênica, n termos "bias"e constantes de tempo para cada nó do sistema. A hipótese de instantes de tempo discreto para o próximo estado da rede não é necessário neste caso. Chen *et al* (1999) [20] tentou reduzir o espaço de busca para os parâmetros, fazendo algumas hipóteses adicionais. Desse modo, é possível examinar mais características do que em modelos mais simples.

3.4.2 Identificação de redes

O objetivo de identificação de redes é construir um modelo de larga escala da rede de interações regulatórias entre os genes. Isto requer encontrar a arquitetura de rede dos padrões de expressão.

Para um dado conjunto de medidas, métodos de engenharia reversa tentam inferir as redes regulatórias desconhecidas pelo uso de um modelo paramétrico, e adaptar os parâmetros aos dados reais.

Se a arquitetura do modelo é desconhecida, o modelo paramétrico terá que ser muito geral e simplificado. Por outro lado, se a arquitetura é bem conhecida, podemos usar um modelo mais detalhado para estimar parâmetros relacionados a mecanismos individuais [23]. Identificação prévia da arquitetura [66] serve para restringir a conectividade da rede, podendo reduzir bastante a quantidade necessária de amostras para uma boa estimação dos parâmetros, contornando o problema da dimensionalidade.

Diversos métodos para a identificação de redes gênicas - arquitetura e dinâmica - foram propostos. Alguns deles serão listados a seguir.

Somogyi et al [66] propôs duas premissas para a identificação de redes gênicas. A primeira (mecânico-reducionista) é a seguinte:

"Um gene para toda função e uma função para todo gene"

Essa premissa acarreta uma série de implicações:

- Redução completa de organismo em genes
- Determinação de atividades e estrutura de proteínas

- Mapeamento de interações moleculares entre produtos de genes
- Montagem de banco de dados de mecanismos moleculares
- Síntese da soma de suas partes

A segunda premissa é:

"Função gênica é distribuída através de uma rede de processamento paralelo"

Essa premissa implica:

- Identificar genes e elementos da rede gênica
- Determinar estados da rede (padrões de expressão)
- Mapeamento de trajetórias e atratores alternativos
- Trajetórias paralelas sugerem entradas compartilhadas
- Ligações temporais determinadas por formas ondulatórias
- Engenharia reversa computacional da rede

O algoritmo de Akutsu et al [1] pode identificar uma rede Booleana de n nós de $O(\log n)$ pares de transição de estados, por busca exaustiva de todas as funções Booleanas possíveis até encontrar um conjunto que se adeque a todos os dados.

O algoritmo REVEAL (Reverse Engineering Algoritm) [52] é baseado na comparação das entropias de Shannon dos dados de entrada e saída que resulta no mínimo em grau k para cada nó, e assim a rede mínima pode ser inferida [65]. A partir daí é realizada uma busca exaustiva pela regra que adequa-se aos dados, como no algoritmo de Akutsu *et al.*

O algoritmo Predictor [43] determina o conjunto de redes Booleanas consistente com um conjunto de estados constantes de padrões de expressão gênica, cada um gerado por uma perturbação na rede gênica. Este método é aplicado interativamente com o método Chooser, que usa uma metodologia baseada em entropia para propor um experimento de perturbação adicional para discriminar entre o conjunto de redes determinados pelo

Predictor. Deste modo, a rede gênica é sucessivamente refinada usando os dados de expressão cumulativa.

Redes Bayesianas tem sido sugeridas para analisar dados de expressão [5, 30, 36]. Utilza-se conhecimento a priori da arquitetura da rede regulatória para construção de alguns modelos. Assim, uma rede Bayesiana é utilizada para selecionar o modelo que melhor se adapata aos dados de expressão [38]. Gifford *et al* utilizou esta abordagem para distinguir entre dois modelos de regulação da galactose [40]. Friedman *et al* analisaram dados de expressão para identificar interações entre genes de várias vias metabólicas e regulatórias [36, 56].

Wahde e Hertz [73] modelaram redes de regulação gênica baseados na formulação de redes neurais recorrentes de tempo contínuo e propuseram um método para determinar os parâmetros de tais redes.

Barrera et al [9] propôs uma técnica para estimar a dinâmica de redes discretas multivaloradas de dados de trajetória e sistemas de envelope (ou seja, um par de sistemas com trajetórias acima e abaixo do sistema alvo) obtido por simulações exaustivas, baseadas no conhecimento biológico sobre a rede. O envelope restringe o espaço de sistemas candidatos, simplificando o problema de estimação.

Capítulo 4

W-operadores

4.1 Introdução

Um W-operador é um operador de transformação de imagem binária que é localmente definido dentro de uma janela W e invariante à translação [10]. Isto significa que ele depende apenas das formas da imagem de entrada vistas através da janela W e que a regra de transformação aplicada é a mesma para todos os pixels da imagem. Um W-operador é caracterizado por uma função Booleana que depende do número de variáveis de W. Erosão, dilatação, abertura, fechamento, detecção de contorno, *hit-miss*, filtro da mediana e esqueletonização são alguns exemplos de W-operadores.

Estimar um W-operador a partir dos dados de treinamento é um problema de otimização. Os dados de treinamento fornecem uma amostra de uma distribuição conjunta das formas observadas e sua classificação (valor Booleano associado às formas observadas na imagem de saída). Uma função perda mede o custo de um erro de classificação de uma forma. Um erro de operador é a esperança da função perda sob a distribuição conjunta. Dado um conjunto de W-operadores, o operador ideal é aquele que tem erro mínimo. Como, na prática, a distribuição conjunta é conhecida apenas pelas suas amostras, ela deve ser estimada. Isto implica que o erro dos operadores também deve ser estimado e, conseqüentemente, o próprio operador ideal deve ser estimado. Estimar um W-operador é uma tarefa fácil quando a amostragem da distribuição conjunta é grande. Entretanto, este dificilmente é o caso. Usualmente, o problema envolve grandes janelas com massa

não concentrada de distribuições de probabilidade conjunta, o que requer uma quantidade proibitiva de dados de treinamento.

Uma abordagem para lidar com a falta de dados de treinamento é restringir o espaço de operadores considerado. De fato, quando o número de operadores candidatos diminui, é necessário menos dados de treinamento para obter boas estimações do melhor operador candidato [26]. Usualmente, o espaço do operador é restrito com base em algum conhecimento *a priori* sobre as características desejadas do operador ideal.

4.2 Definição e propriedades

Seja E o plano dos inteiros e "-" a translação em E. Uma *imagem binária* ou simplesmente *imagem* é uma função f de E em $\{0,1\}$. Uma imagem f pode ser representada, equivalentemente, por um subconjunto X de E pela seguinte relação: $\forall x \in E, x \in X \Leftrightarrow f(x) = 1$.

A translação de uma imagem $X \subseteq E$ por um vetor $h \in E$ é a imagem $X_h = \{x \in E : x - h \in X\}$.

Seja $\mathcal{P}(E)$ o conjunto potência de E, isto é, o conjunto de todos os subconjuntos possíveis de E. Uma *transformação de imagem* ou *operador* é um mapeamento Ψ de $\mathcal{P}(E)$ em $\mathcal{P}(E)$.

Um operador Ψ é chamado de *invariante à translação* se e somente se, para todo $h \in E$,

$$\Psi(X_h) = \Psi(X)_h.$$

Seja W um subconjunto finito de E. Um operador Ψ é *localmente definido* na janela W se e somente se, para todo $x \in E$,

$$x \in \Psi(X) \Leftrightarrow x \in \Psi(X \cap W_x).$$

Um operador é chamado de *W-operador* se ele é invariante à translação e localmente definido em uma janela finita W. Qualquer W-operador Ψ pode ser caracterizado por uma função Booleana ψ de $\mathcal{P}(W)$ em $\{0,1\}$ através da relação, $\forall x \in E$,

$$x \in \Psi(X) \Leftrightarrow \psi(X_{-x} \cap W) = 1.$$

Portanto, escolher um W-operador Ψ é equivalente a escolher sua função Booleana ψ correspondente.

4.3 Construção de W-operadores

Construir um W-operador significa escolher um operador de uma família de candidatos para realizar uma determinada tarefa. É um problema de otimização onde o espaço de busca é a família de operadores candidatos e o critério de otimização é uma medida da qualidade do operador. Na formulação usualmente adotada, o critério é baseado em um modelo estatístico para as imagens associadas a uma medida de similaridade de imagens: a função perda.

Sejam **S** e **I** dois conjuntos aleatórios discretos definidos em E, isto é, realizações de **S** e **I** são imagens obtidas de acordo com alguma distribuição de probabilidade em $\mathcal{P}(E)$. Transformações de imagens são modeladas em um dado contexto pelo processo aleatório (**S**, **I**), onde o processo **S** representa as imagens de entrada e **I** as imagens de saída. O processo **I** depende do processo **S** de acordo com uma distribuição condicional.

Dado um espaço de operadores $\mathcal{O}$ e uma função perda ℓ de $E \times E$ em $\mathbb{R}^+$, o erro $Er[\Psi]$ de um operador $\Psi \in \mathcal{O}$ é a esperança de $\ell(\Psi(\mathbf{S}), \mathbf{I})$, isto é, $Er[\Psi] = E[\ell(\Psi(\mathbf{S}), \mathbf{I})]$. O operador *ideal* Ψ_{opt} é aquele que tem mínimo erro, isto é, $Er[\Psi_{opt}] \leq Er[\Psi], \forall \Psi \in \mathcal{O}$.

Um processo aleatório conjunto (**S**, **I**) é *conjuntamente estacionário* em relação a uma janela W finita, se a probabilidade de ver, através de W, uma forma na imagem de entrada associada a um valor Booleano na imagem de saída é a mesma para qualquer translação de W, isto é, para todo $x \in E$,

$$P(S \cap W_x, I(x)) = P(S \cap W, I(o)),$$

em que S é uma realização de **S**, I é a função Booleana equivalente a uma realização de **I**, e o é a origem de E.

Para tornar o modelo utilizável na prática, suponha que (**S**, **I**) seja conjuntamente estacionário em relação à janela finita W. Sob esta hipótese, o erro de predizer uma imagem da observação de uma outra imagem pode ser substituído pelo erro de predizer um pixel da observação de uma forma através de W e, conseqüentemente, o operador ótimo Ψ_{opt} é sempre um W-operador. Então, o problema de otimização pode ser formulado no espaço das funções Booleanas definidas em $\mathcal{P}(W)$, com processos aleatórios conjuntos em $(\mathcal{P}(W), \{0,1\})$ e funções perdas ℓ de $\{0,1\} \times \{0,1\}$ em $\mathbb{R}^+$.

Na prática, as distribuições $(\mathcal{P}(W), \{0,1\})$ são desconhecidas e devem ser estimadas, o que implica em estimar $Er[\psi]$ e ψ_{opt}. Quando a janela é pequena ou a distribuição tem uma massa de probabilidades concentrada em algum lugar, a estimação é fácil. Entretanto, isto raramente acontece. Usualmente, temos grandes janelas com distribuições de massa não concentradas, o que requer uma quantidade enorme de dados de treinamento.

Uma abordagem para lidar com a falta de dados é restringir o espaço de busca. O erro estimado de um operador em um espaço restrito pode ser decomposto como a adição do incremento do erro do operador ótimo (isto é, aumento no erro do operador ótimo pela redução do espaço de busca) e o erro de estimação do espaço restrito. Uma restrição é benéfica quando o erro de estimação da restrição diminui (isto é, em relação ao erro de estimação do espaço inteiro) mais que o incremento do erro do operador ótimo. As restrições conhecidas são heurísticas propostas por especialistas.

4.3.1 W-operadores ótimos

Formular o problema de otimização requer dar a noção do que é ser o melhor. Se uma imagem ruidosa é observada, o problema de restauração é encontrar um operador que filtra a imagem de tal modo a estimar a imagem original não ruidosa. Se uma imagem é observada e um contorno é desejado, o problema é encontrar um detector de contorno que opera na imagem para estimar o contorno verdadeiro. Se quisermos encontrar um padrão em uma imagem, o problema é encontrar um algoritmo de reconhecimento de padrões para marcar os locais onde cópias do padrão estão localizadas. Essas diversas tarefas podem ser realizadas levando-se em conta uma imagem de entrada observada, processando-a por um operador, e então comparando a imagem de saída com uma imagem desejada. O problema é probabilístico porque o operador deve ser aplicado a um conjunto aleatório

de imagens observadas, sendo que essas imagens devem ser processadas para estimar um conjunto aleatório de imagens desejadas [29].

Suponha que uma dada janela seja transladada para um pixel z e que os n valores da imagem observada na janela transladada W_z formam o vetor aleatório $\mathbf{X}$. Além disso, suponha que o valor no pixel z da imagem desejada seja Y. Se Ψ é um operador arbitrário com função característica ψ, então $\psi(\mathbf{X})$ serve como um estimador de Y. Existem duas possibilidades:

- $\psi(\mathbf{X}) = Y$: neste caso não há erro;
- $\psi(\mathbf{X}) \neq Y$: há um erro.

O *erro médio absoluto* (MAE - Mean Absolute Error) é o valor esperado de $|\psi(\mathbf{X}) - Y|$, ou simplesmente o número de erros de classificação cometidos pela função ψ dividido pelo total de pixels da imagem.

O objetivo é encontrar o operador que tem o mínimo MAE para um dado problema de estimação. Pode existir mais de um operador com MAE mínimo. Qualquer operador que tenha MAE mínimo é chamado de *operador ótimo* ou *filtro ótimo*. Um operador ótimo é definido em termos das probabilidades condicionais de ψ:

$$\psi_{opt}(\mathbf{X}) = \begin{cases} 1 & \text{se } P(Y=1|\mathbf{X}) > 0.5 \\ 0 & \text{se } P(Y=1|\mathbf{X}) \leq 0.5 \end{cases} \tag{4.1}$$

Um filtro ótimo é determinado apenas pelas probabilidades condicionais dadas na equação anterior. Denotando o erro de um operador Ψ por $Er[\Psi]$, temos que o erro do operador ótimo é dado pela seguinte equação:

$$Er[\Psi_{opt}] = \sum_{\{\mathbf{x}:P(Y=1|\mathbf{x})>0.5\}} P(\mathbf{x})P(Y=0|\mathbf{x}) + \sum_{\{\mathbf{x}:P(Y=1|\mathbf{x})\leq 0.5\}} P(\mathbf{x})P(Y=1|\mathbf{x}) \tag{4.2}$$

Se forem listadas as possíveis observações $\mathbf{x}$ em uma tabela com as probabilidades $P(\mathbf{x})$ e probabilidades condicionais $P(Y=0|\mathbf{x})$, então o erro é obtido pela soma dos produtos das probabilidades e das probabilidades condicionais correspondendo aos valores (0 ou 1) não escolhidos por ψ_{opt}.

4.3.2 Construção de W-operadores ótimos

Na prática, um filtro ótimo é estatisticamente estimado dos dados de treinamento. Ele é obtido através da estimação das probabilidades condicionais de um conjunto de pares de imagens (ideal e observada). Para cada observação possível $\mathbf{x}$ na janela, a estimação da probabilidade condicional $P(Y = 1|\mathbf{x})$ é dada por:

$$\hat{P}(Y = 1|\mathbf{x}) = \frac{\text{número de vezes em que } y = 1 \text{ quando } \mathbf{x} \text{ é observado}}{\text{número de vezes em que } \mathbf{x} \text{ é observado entre as amostras}} \tag{4.3}$$

O estimador construído, Ψ_{est} do filtro ótimo pode ou não ser ótimo. Ele é ótimo se e somente se $\hat{P}(Y = 1|\mathbf{x}) > 0.5$ quando $P(Y = 1|\mathbf{x}) > 0.5$. Se o número de amostras for suficientemente grande, então $\hat{P}(Y = 1|\mathbf{x})$ é uma boa aproximação de $P(Y = 1|\mathbf{X})$ e é esperado que o filtro estimado seja próximo ao ótimo. Entretanto, para um número de amostras insuficiente, $\hat{P}(Y = 1|\mathbf{x})$ pode diferir consideravelmente de $P(Y = 1|\mathbf{X})$, sendo que o filtro estimado pode divergir do ótimo. Isso significa que o erro do filtro estimado depende do número m de amostras. Além disso, esse erro também depende dos dados das amostras coletadas.

Um caso extremo é quando uma determinada observação $\mathbf{x}$ não aparece no conjunto de treinamento. Neste caso, não há como estimar $P(Y = 1|\mathbf{x})$. Algum critério extra deve ser empregado neste caso para decidir se $\psi_{est}(\mathbf{x}) = 1$ ou $\psi_{est}(\mathbf{x}) = 0$. Isto é conhecido como o *problema de generalização*.

Então, caso esteja sendo usada uma janela muito grande para a quantidade de amostras disponíveis, o melhor W-operador pode ser obtido em uma subjanela contida na janela considerada. Assim, o problema de descobrir a subjanela ideal a partir da qual será construído o W-operador ótimo, pode ser formulado como um problema de seleção de características, onde cada pixel da janela corresponde à uma característica.

As figuras 4.1, 4.2, 4.3 e 4.4 mostram um exemplo que ilustra o processo de obtenção das amostras de treinamento para selecionar a subjanela na qual o W-operador será construído. Dada uma janela $M \times N$, para cada pixel (i, j) da imagem observada, a janela é transladada sobre a imagem observada de tal forma que o centro da janela esteja em (i, j). Através desta janela, observa-se formas (padrões) da imagem, ou seja, toda translação da janela fornece uma observação (instância) da amostra de treinamento. Os

rótulos dessas instâncias são emitidos observando-se o valor da posição (i, j) na imagem ideal. Portanto, uma amostra do conjunto de treinamento é composta pela forma vista na imagem observada através da janela e pelo rótulo dado pela posição (i, j) da imagem ideal. A figura 4.5 mostra as linhas da tabela do conjunto de treinamento correspondentes às translações da janela, em ordem de varredura, de $(3, 3)$ até $(7, 3)$ e de $(20, 10)$ até $(24, 10)$ sobre as imagens mostradas nas figuras 4.1, 4.2, 4.3 e 4.4 (cada linha dessa tabela corresponde a uma amostra).

Mostramos na seção 6.1.3 como nossa metodologia proposta baseada em seleção de características por análise de entropia condicional (capítulo 5) foi empregada para obter W-operadores próximos aos ideais aplicados em filtragem de imagens e reconhecimento de texturas.

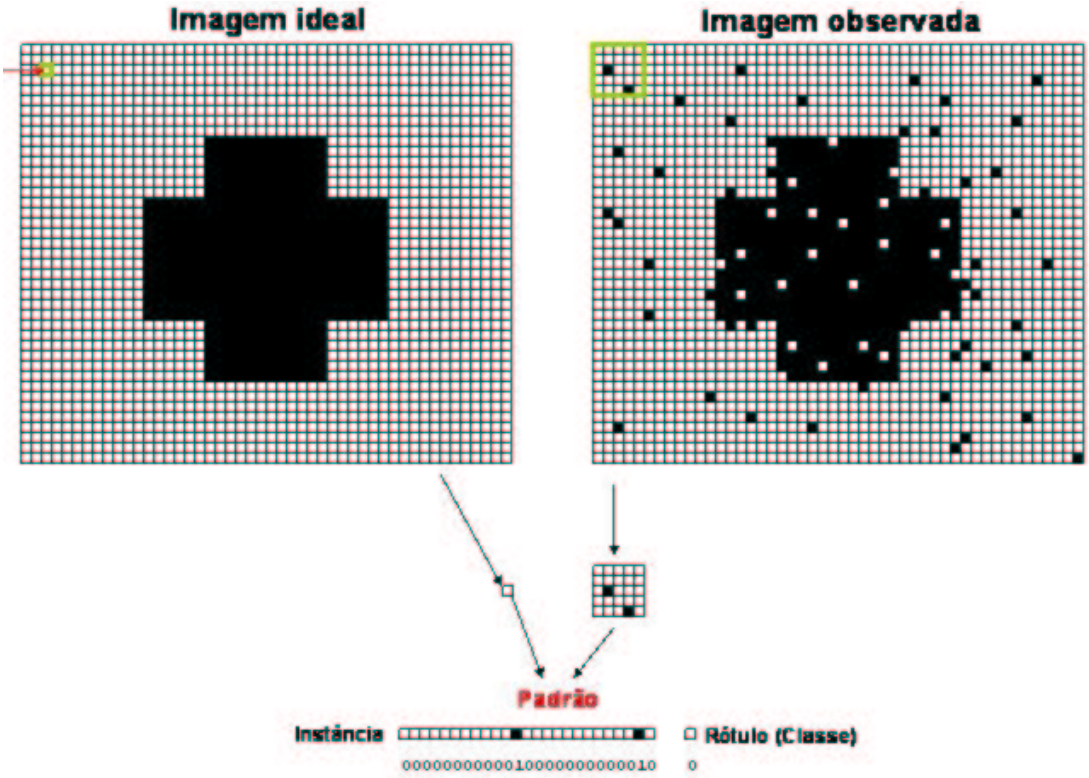


Figura 4.1: Obtenção de uma amostra que compõe o conjunto de treinamento para construção de um W-operador (posição (3,3)).

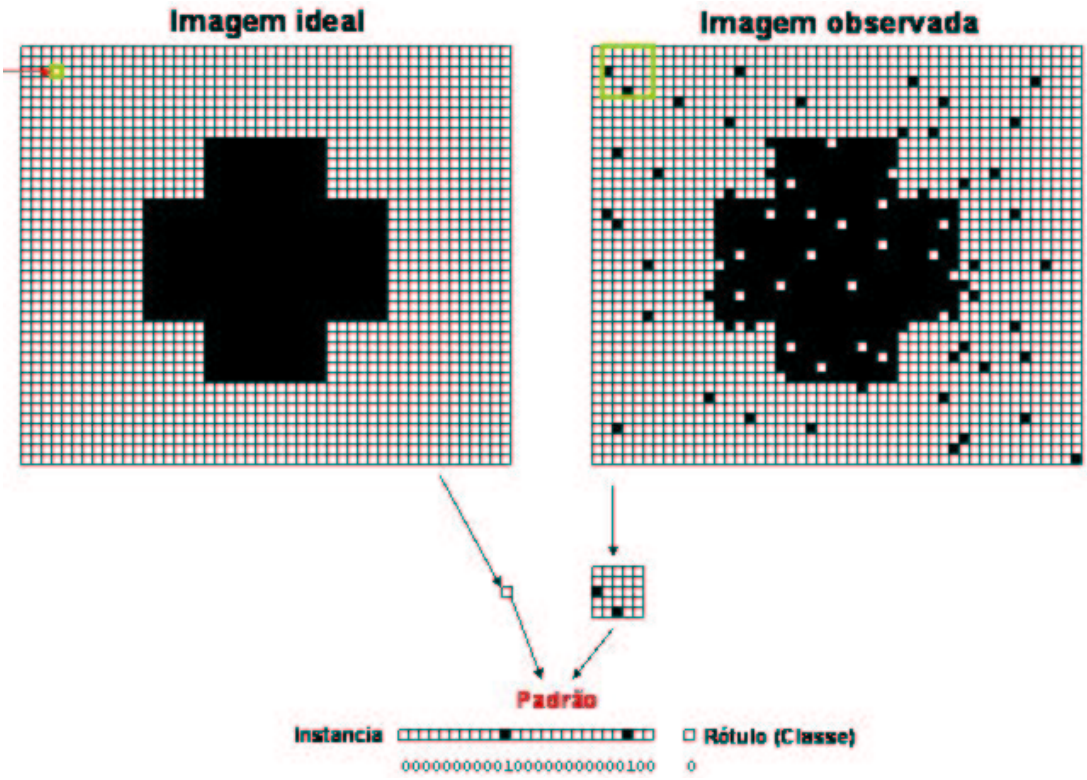


Figura 4.2: Obtenção de uma amostra que compõe o conjunto de treinamento para construção de um W-operador (posição (4,3)).

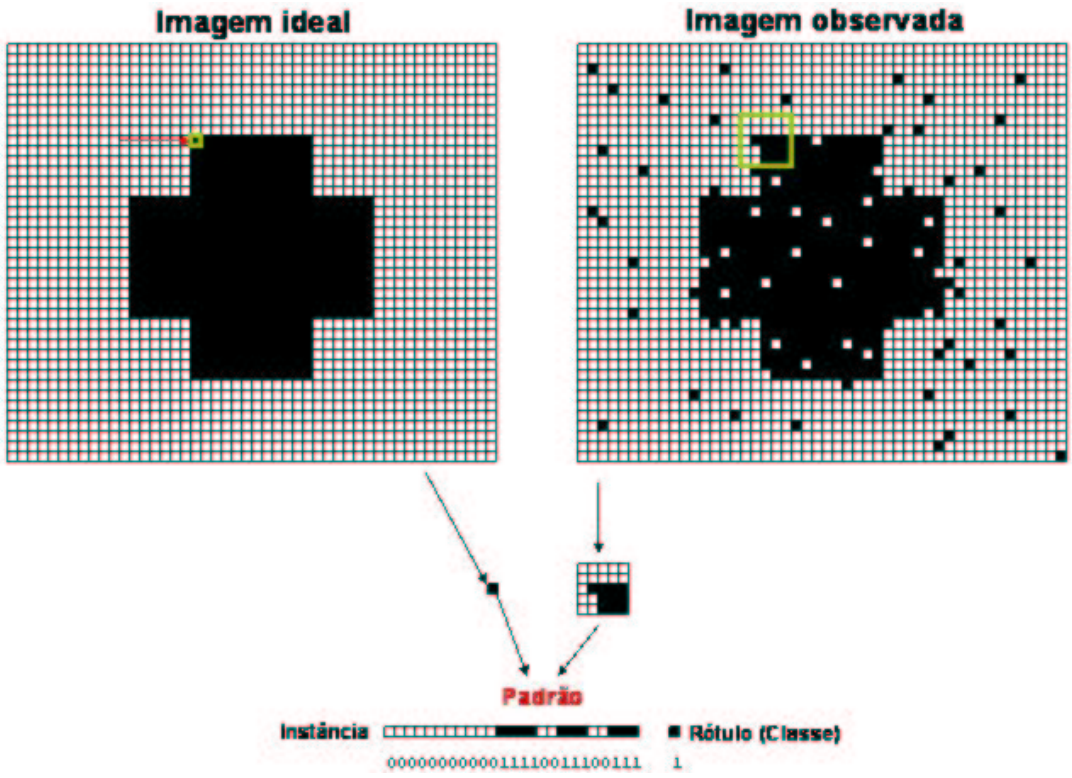


Figura 4.3: Obtenção de uma amostra que compõe o conjunto de treinamento para construção de um W-operador (posição (20,10)).

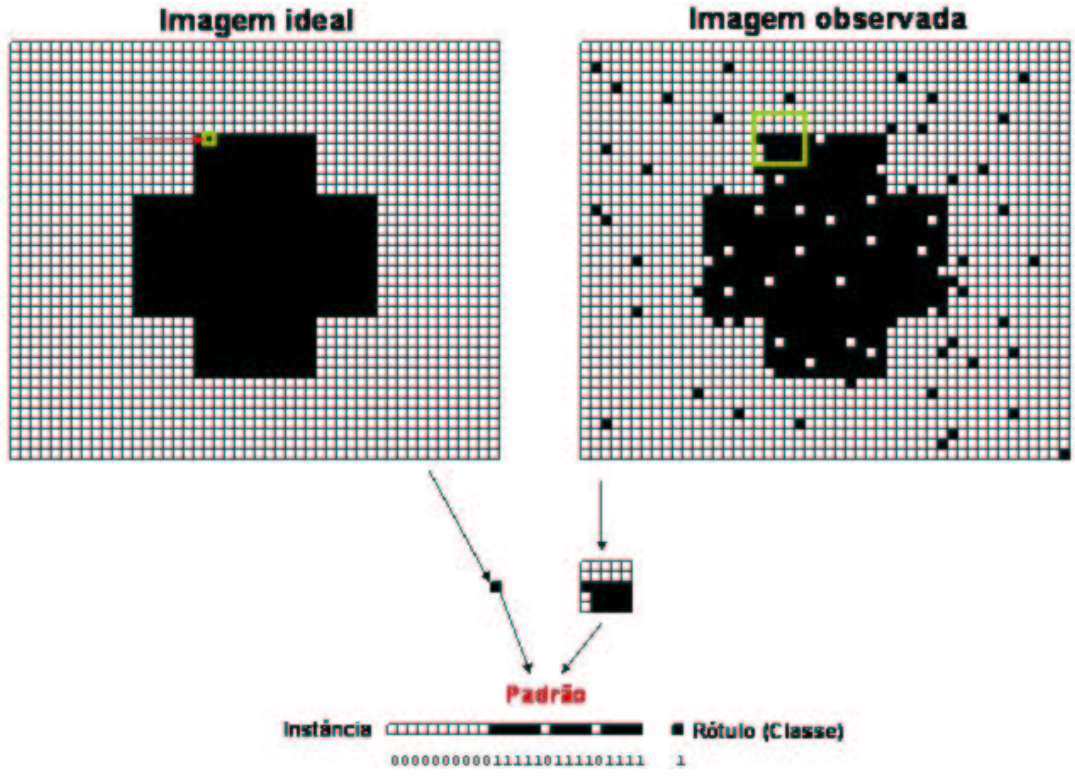


Figura 4.4: Obtenção de uma amostra que compõe o conjunto de treinamento para construção de um W-operador (posição (21,10)).

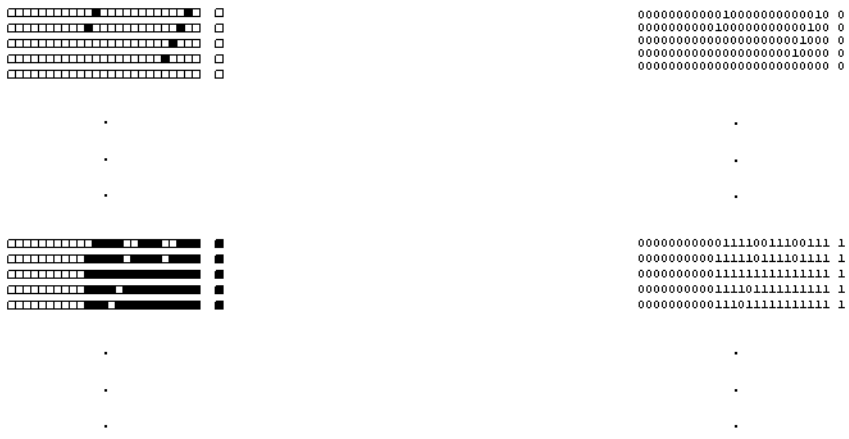


Figura 4.5: Linhas do arquivo do conjunto de treinamento para construção de um W-operador (translações em ordem de varredura de (3,3) até (7,3) e de (20,10) até (24,10) sobre as imagens mostradas nas figuras 4.1, 4.2, 4.3 e 4.4.

Parte II

Metodologia proposta para seleção de características

Capítulo 5

Seleção de características por análise de entropia condicional

5.1 Introdução

O núcleo da contribuição desta dissertação encontra-se descrito neste capítulo. Redução de dimensionalidade e seleção de características exercem uma função primordial em reconhecimento de padrões. O principal objetivo da redução de dimensionalidade é obter um subespaço bastante reduzido de características que seja adequado para representar os padrões em questão. Em seleção de características, a *função critério* é a responsável por definir a qualidade de um determinado subespaço de características. Propomos um critério baseado em entropia condicional para medir a qualidade de um determinado subespaço de características, dependendo de sua dimensão e do número de amostras de treinamento. A seguir definimos uma função critério com base em princípios estatísticos da teoria da informação.

5.2 Critério para seleção de características: entropia condicional média

Este capítulo segue a notação introduzida na seção 2.2.

Seja $\mathbf{x}_{\mathcal{Z}}$ uma amostra de $\mathbf{X}_{\mathcal{Z}}$ e Y uma variável aleatória representando o conjunto de rótulos. O interesse está em descobrir alguma maneira de medir quantitativamente a predição do comportamento de Y com base em $\mathbf{x}_{\mathcal{Z}}$. Se Y for fortemente predito por $\mathbf{x}_{\mathcal{Z}}$, significa que dado $\mathbf{x}_{\mathcal{Z}}$, pode-se inferir o valor de Y com alta probabilidade de acerto. A resposta a esta questão é encontrada na teoria da informação formulada por Claude Shannon [62].

O conceito de *entropia* (entropia de Shannon) é o de uma medida de informação calculada pelas probabilidades de ocorrência de eventos individuais ou combinados [63]. Sejam X e Y variáveis aleatórias e P a função probabilidade. Formalmente, a entropia de X é definida como:

$$H(X) = -\sum_{x \in X} P(x) log P(x) \tag{5.1}$$

A *entropia conjunta* de X e Y é definida como:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) log P(x,y) \tag{5.2}$$

E a *entropia condicional* de Y dado X:

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(y|x) log P(y|x) \tag{5.3}$$

Observação: caso a probabilidade $P(\cdot)$ seja zero, por convenção adota-se $log 0 = 0$ para o cálculo da entropia H.

A fórmula da entropia condicional encontrada na literatura é ligeiramente diferente da equação 5.3, pois a primeira é ponderada pelas probabilidades de x. Neste texto, tal fórmula é denominada *entropia condicional média* $E[H(Y|X)]$, como definida na equação 5.4. A definição desta mesma fórmula levando em conta a notação introduzida na seção 2.2 é dada pela equação 5.6. Uma modificação da fórmula da entropia condicional média, que atribui uma ponderação positiva às instâncias não observadas ($P(x) = 0$), encontra-se na equação 5.7.

$$E[H(Y|X)] = -\sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) log P(y|x) \tag{5.4}$$

Informação mútua, M, (também conhecida como *ganho de informação* [39]) é definida como uma soma das entropias individuais menos a entropia conjunta, sendo uma medida de correlação entre duas variáveis X e Y [63]. A entropia condicional média $E[H(Y|X)]$ é a diferença da entropia conjunta $H(X,Y)$ com relação a entropia individual $H(X)$. Então:

$$M(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - E[H(Y|X)] \tag{5.5}$$

pois $E[H(Y|X)] = H(X,Y) - H(X)$ [39].

A idéia central é encontrar o subespaço $\mathbf{X}_{\mathcal{Z}}$ de características que maximiza a informação mútua de todas as possíveis instâncias $\mathbf{x}_{\mathcal{Z}_i}, 1 \leq i \leq m$ com relação a Y. Em outras palavras, maximizar a informação mútua neste caso é equivalente a encontrar o subespaço de características que realiza a melhor predição do rótulo ou classe de um determinado padrão pertencente às amostras de treinamento. Isto porque a equação 5.5 pode ser interpretada do seguinte modo. Caso X consiga organizar adequadamente a informação sobre Y ($E[H(Y|X)]$ baixo), mesmo que Y tenha um comportamento muito caótico ($H(Y)$ alto), então a informação obtida de Y através de X será bastante valiosa ($M(X,Y)$ alto).

Como os valores de Y são fixos para um determinado conjunto de treinamento, $H(Y)$ terá sempre o mesmo valor para qualquer conjunto $\mathbf{X}_{\mathcal{Z}}$. Portanto, dada esta constatação e a equação 5.5, quanto menor a informação mútua, maior a entropia condicional. Isto implica que a entropia condicional $H(Y|\mathbf{X}_{\mathcal{Z}} = \mathbf{x}_{\mathcal{Z}_i})$ é suficiente para avaliar quantitativamente a informação de Y condicionada a uma possível instância $\mathbf{x}_{\mathcal{Z}_i}$ de $\mathbf{X}_{\mathcal{Z}}$. Com base na equação 5.3, temos que a fórmula de $H(Y|\mathbf{X}_{\mathcal{Z}} = \mathbf{x}_{\mathcal{Z}_i})$ é dada pela seguinte equação:

$$H(Y|\mathbf{X}_{\mathcal{Z}} = \mathbf{x}_{\mathcal{Z}_i}) = -\sum_{y=1}^{c} P(y|\mathbf{x}_{\mathcal{Z}_i}) log P(y|\mathbf{x}_{\mathcal{Z}_i}) \tag{5.6}$$

onde c é o número de classes de Y.

A motivação para o estudo da entropia como função critério para seleção de características surge da capacidade que esse conceito estatístico possui de medir o grau de aleatoriedade (ou de incerteza) de variáveis individuais ou combinadas. Dada a distribuição de $\mathbf{X}_{\mathcal{Z}}$, quanto menor o grau de aleatoriedade de Y condicionado aos valores de

$\mathbf{X}_{\mathcal{Z}}$, mais informação teremos sobre o comportamento de Y quando tomamos como referência os valores de $\mathbf{X}_{\mathcal{Z}}$. Um caso extremo é quando Y for totalmente determinado por $\mathbf{X}_{\mathcal{Z}}$, tendo grau de aleatoriedade nula, ou seja, a entropia condicional $H(Y|\mathbf{X}_{\mathcal{Z}})$ neste caso é nula.

Para dar uma idéia sobre o potencial da entropia como critério para seleção de características, considere o gráfico onde os rótulos Y são representados pela abscissa e a probabilidade de um padrão ser rotulado como $Y = y$ dada a ocorrência da instância $\mathbf{x}_{\mathcal{Z}_\mathbf{i}}$ representada pela ordenada (fig. 5.1). Se ele apresentar um pico saliente (massa de probabilidades bem concentrada), significa que a entropia condicional $H(Y|\mathbf{x}_{\mathcal{Z}_\mathbf{i}})$ é pequena, isto é, $\mathbf{x}_{\mathcal{Z}_\mathbf{i}}$ prediz os rótulos de Y com boa confiança. Por outro lado, se o gráfico apresenta-se achatado (massa de probabilidades bem distribuída), a entropia $H(Y|\mathbf{x}_{\mathcal{Z}_\mathbf{i}})$ é alta, significando que $\mathbf{x}_{\mathcal{Z}_\mathbf{i}}$ não prediz Y. Portanto, a entropia condicional pode ser usada como um critério bastante apropriado para realizar seleção de características.

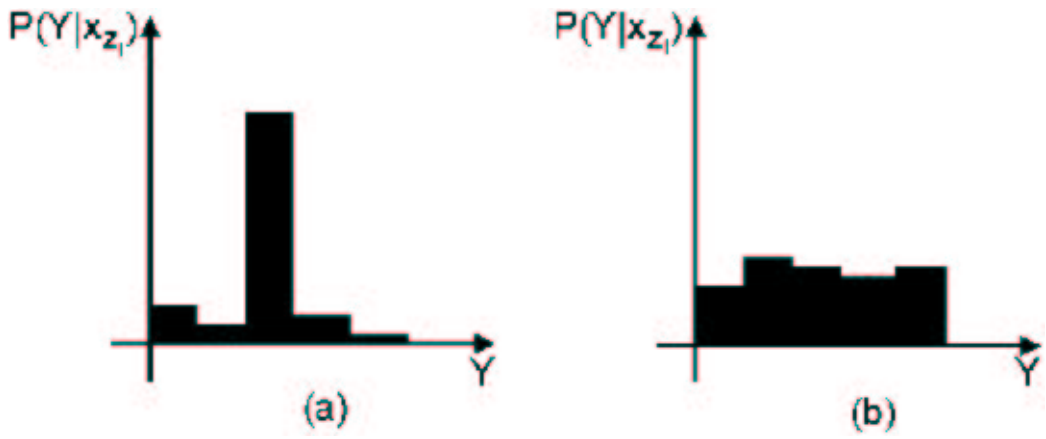


Figura 5.1: (a) Baixa entropia; (b) Alta entropia.

É importante observar que caso Y seja constante, isto é, seu valor seja sempre o mesmo para qualquer amostra de treinamento, $H(Y|\mathbf{X}_{\mathcal{Z}}) = 0$ para todo $\mathcal{Z} \subseteq \mathcal{I} = \{1, 2, ..., n\}$. Ou seja, qualquer subespaço de características prediz Y com 100% de certeza, afinal Y assume um único valor sempre. Porém a informação mútua de Y com relação a qualquer desses subespaços é nula, pois $H(Y) = 0$ (ver equação 5.5). Isto significa que Y é uma variável independente pois nenhum subespaço carrega informação adicional sobre seu valor. A conclusão disso é que deve-se verificar se $H(Y) > 0$ antes de realizar a seleção de características propriamente dita. Caso $H(Y) = 0$, basta devolver o conjunto vazio como

solução.

5.2.1 Entropia condicional média

Agora, assumindo $H(Y) > 0$ como hipótese, para decidir se $\mathbf{X}_{\mathcal{Z}}$ prediz Y, basta calcular a entropia condicional média de todas as possíveis instâncias $\mathbf{x}_{\mathcal{Z}_1}, \mathbf{x}_{\mathcal{Z}_2}, ..., \mathbf{x}_{\mathcal{Z}_m}$ ponderada pelo número de ocorrências de cada uma das instâncias no conjunto de treinamento. A isto, denominamos *entropia condicional média de Y dado $\mathbf{X}_{\mathcal{Z}}$* (denotado $E[H(Y|\mathbf{X}_{\mathcal{Z}})]$) dada pela equação 5.7:

$$E[H(Y|\mathbf{X}_{\mathcal{Z}})] = \sum_{i=1}^{m} \frac{H(Y|\mathbf{x}_{\mathcal{Z}_i}) \cdot o_i}{t} \tag{5.7}$$

onde o_i é o número de ocorrências da instância $\mathbf{x}_{\mathcal{Z}_i}$ no conjunto de treinamento, t é o número de amostras do conjunto de treinamento e m é o número de instâncias possíveis de $\mathbf{X}_{\mathcal{Z}}$. O valor de m é dado por p^d, onde p é o número de valores discretos que cada característica pode assumir, e d é a dimensão de $\mathbf{X}_{\mathcal{Z}}$ (número de características).

A equação anterior funciona bem para os casos onde todas as possíveis instâncias de $\mathbf{X}_{\mathcal{Z}}$ são observadas pelo menos uma vez no conjunto de treinamento. Para os casos onde nem todas as instâncias são observadas, é necessário um refinamento da fórmula para adequá-los. Subespaços de características que possuem muitas instâncias não observadas no conjunto de treinamento são indesejados pois, caso essas instâncias apareçam nas amostras do conjunto de teste, um classificador baseado em tal subespaço acaba sendo forçado a inferir uma classe qualquer a essas instâncias sem nenhum conhecimento *a priori*.

Com a finalidade de amenizar esse problema, suponha que $\mathbf{X}_{\mathcal{Z}_i}$ seja uma instância não observada de $\mathbf{X}_{\mathcal{Z}}$. Como este é um caso indesejado, podemos atribuir entropia máxima a $H(Y|\mathbf{X}_{\mathcal{Z}_i})$. Para isso, basta fazer com que $P(Y|\mathbf{X}_{\mathcal{Z}_i})$ tenha distribuição uniforme, ou seja, $P(y|\mathbf{X}_{\mathcal{Z}_i}) = 1/c$ para todo $y \in Y = \{1, 2, ..., c\}$. Como espera-se que essas instâncias sejam raras nas amostras de teste se o conjunto de treinamento for adequado, parece uma boa idéia fazer com que suas entropias entrem com peso mínimo no cômputo da entropia condicional média. Somando uma constante $\alpha > 0$ à ocorrência de cada uma das possíveis instâncias, garante-se que o menor peso será dado a essas instâncias não

observadas. Assim, a fórmula da entropia condicional média, que também leva em conta as instâncias não observadas, pode ser definida pela equação 5.8. Utilizamos $\alpha = 1$ em todos os experimentos realizados.

$$E[H(Y|\mathbf{X}_{\mathcal{Z}})] = \sum_{i=1}^{m} \frac{H(Y|\mathbf{x}_{\mathcal{Z}_\mathbf{i}}) \cdot (o_i + \alpha)}{\alpha m + t} \tag{5.8}$$

onde $H(Y|\mathbf{x}_{\mathcal{Z}_\mathbf{i}}) = -log(1/c)$ caso $\mathbf{x}_{\mathcal{Z}_\mathbf{i}}$ não tenha sido observado no conjunto de treinamento (entropia máxima).

Então o problema é resolvido selecionando-se $\mathcal{Z}^* \subseteq \mathcal{I}$ de acordo com a seguinte equação (5.9):

$$\mathcal{Z}^* : H(Y|\mathbf{X}_{\mathcal{Z}^*}) = min_{\mathcal{Z} \subseteq \mathcal{I}}\{E[H(Y|\mathbf{X}_{\mathcal{Z}})]\} \tag{5.9}$$

onde $\mathcal{I} = \{1, 2, ..., n\}$ (conjunto de índices do espaço total de n características) e $E[H(Y|\mathbf{X}_{\mathcal{Z}})]$ é dada pela equação 5.8

Portanto, a exploração de todos os possíveis subconjuntos de $\mathcal{I}$ solucionaria o problema, mas isto é impraticável em geral. Há algumas heurísticas de busca que tentam obter um subconjunto sub-ótimo explorando um espaço de busca muito menor do que o espaço inteiro das combinações (ver seção 2.2.1). A seguir, será apresentado o algoritmo que calcula a entropia condicional média de um determinado subespaço de características com base em um conjunto de amostras de treinamento. Antes de mais nada é necessário frisar novamente que qualquer método de seleção de características que utilize o algoritmo a seguir como função critério deverá **minimizá-la** para selecionar os melhores subespaços, conforme definido pela equação 5.9.

Exemplo: observe as matrizes representadas pela tabela 5.1 com 3 características ($d = 3$) onde cada característica pode assumir dois valores (0 ou 1) ($p = 2$) e existem 3 classes possíveis (0, 1 ou 2) ($c = 3$). Cada célula (i, j) indica $P(Y = j|\mathbf{x}_{\mathcal{Z}_\mathbf{i}})$, ou seja, a probabilidade de $Y = j$ dado que $\mathbf{X}_{\mathcal{Z}} = \mathbf{x}_{\mathcal{Z}_\mathbf{i}}$. A coluna nomeada $o + 1$ é o número de ocorrências mais 1 de cada $\mathbf{x}_{\mathcal{Z}_\mathbf{i}}$ no conjunto de treinamento de tamanho 32 ($t = 32$).

F			Y				
f_1	f_2	f_3	0	1	2	$o+1$	H
0	0	0	0.3333	0.5	0.1667	7	0.9206
0	0	1	0.3333	0.3333	0.3333	1	1
0	1	0	0	1	0	3	0
0	1	1	0.75	0	0.25	5	0.5119
1	0	0	0.2857	0.7143	0	8	0.5446
1	0	1	0.25	0.375	0.375	9	0.9851
1	1	0	0.25	0.25	0.5	5	0.9464
1	1	1	1	0	0	2	0
Entropia Condicional Média $E[H(Y\|\mathbf{F})] = 0.6990$							

G			Y				
g_1	g_2	g_3	0	1	2	$o+1$	H
0	0	0	0	1	0	6	0
0	0	1	0.25	0	0.75	5	0.5119
0	1	0	0	0	1	5	0
0	1	1	0.2	0.1	0.7	11	0.7298
1	0	0	1	0	0	2	0
1	0	1	0	0	1	4	0
1	1	0	0	1	0	4	0
1	1	1	0	0	1	3	0
Entropia Condicional Média $E[H(Y\|\mathbf{G})] = 0.2647$							

Tabela 5.1: Exemplo de Entropia Condicional Média calculada para 2 subespaços de características **F** e **G**. Note que **G** tem um poder de influência sobre Y muito maior do que **F** sobre Y. Foi utilizado logaritmo na base 3 para o cálculo das entropias.

5.2.2 Algoritmo

Sejam T o conjunto de amostras de treinamento contendo t pares da forma $((x_1, x_2, ..., x_n), y)$, $P[m][c]$ a matriz de probabilidades condicionais de cada amostra possível $(\mathbf{x}_{\mathcal{Z}_i}, j)$ e $D[m]$ o vetor de contagem de observações de cada instância possível $\mathbf{x}_{\mathcal{Z}_i}$. O algoritmo abaixo

calcula a entropia condicional média de um subespaço de características $\mathbf{X}_{\mathcal{Z}}$ a partir de T.

ECM(X, $\mathcal{Z}$, T, p, c, α)

1. $d \leftarrow$ tamanho de $\mathcal{Z}$; $(O(d))$
2. $m \leftarrow p^d$; número de instâncias possíveis de $\mathbf{X}_{\mathcal{Z}}$ $(O(1))$
3. $P[i][j] \leftarrow 0, \forall 1 \leq i \leq m, \forall 1 \leq j \leq c$; $(O(mc))$
4. $D[i] \leftarrow 0, \forall 1 \leq i \leq m$; $(O(m))$
5. PARA $j \leftarrow 1$ ATÉ t FAÇA $(O(t))$
6. Encontre l, o número da linha em P correspondente à instância $\mathbf{x}^{\mathbf{j}}_{\mathcal{Z}}$ em T; $(O(d))$
7. $P[l][y_j] \leftarrow P[l][y_j] + 1$; $(O(1))$
8. $D[l] \leftarrow D[l] + 1$; $(O(1))$
9. SE existir i tal que $1 \leq i \leq m$ e $D[i] = 0$ ENTÂO $(O(m))$
10. $D[i] \leftarrow D[i] + \alpha, \forall 1 \leq i \leq m$; $(O(1))$
11. $t \leftarrow t + \alpha m$; $(O(1))$
12. PARA $i \leftarrow 1$ ATÉ m FAÇA $(O(m))$
13. SE $D[i] = \alpha$ ENTÃO $(O(1))$
14. $P[i][j] \leftarrow 1/c, \forall 1 \leq j \leq c$; $(O(c))$
15. SENÃO
16. $o_i \leftarrow \sum_{j=0}^{c} P[i][j]$; $(O(c))$
17. $P[i][j] \leftarrow P[i][j]/o_i$; $(O(1))$
18. $H \leftarrow -\sum_{i=1}^{m} \frac{D[i]}{t} \sum_{j=1}^{c} log_c(P[i][j])$; $(O(mc))$
19. DEVOLVA H. $(O(1))$

Discussão

Note que, no passo 18, a base do logaritmo no cálculo da entropia H é c (número de classes possíveis). Na verdade, o valor da base não influencia no resultado da seleção de características, desde que seu valor seja maior que 1. Adotamos a base c como uma forma de normalizar o valor da entropia para que fique no intervalo $0 \leq H \leq 1$. Esta base será utilizada em todos os experimentos descritos no capítulo 6.

A matriz de probabilidades condicionais P pode atuar como um classificador baseado em $\mathbf{X}_{\mathcal{Z}}$. Para classificar um determinado padrão desconhecido, primeiramente verifique o

seu valor $\mathbf{x}_{\mathcal{Z}}$ no padrão. Em seguida, aplique o passo 6 do algoritmo para determinar a linha l em P correspondente ao seu valor. Assim, basta classificá-lo na classe que tiver a maior probabilidade dentre as probabilidades condicionais da linha l em P.

Antes de discutir a complexidade do algoritmo, é preciso mostrar uma possível maneira de realizar o passo 6. Cada uma das características $X'_1, X'_2, ..., X'_d$ possui um valor dentre p valores possíveis, formando um vetor de valores, $(x'_1, x'_2, ..., x'_d)$. Para calcular o índice l correspondente a esse vetor, transforme-o em um inteiro na base decimal pelo seguinte procedimento:

$l \leftarrow 1$; $(O(1))$
PARA $i \leftarrow 0$ ATÉ $d-1$ FAÇA $(O(d))$
  $l \leftarrow l + p^i x'_{d-i}$; $O(1)$
DEVOLVA l; $O(1)$

Exemplo: Seja um subespaço de dimensão $d = 4$ e sua instância observada em uma determinada amostra seja (1, 0, 1, 1). Suponha ainda que cada característica possa ter apenas 2 valores possíveis: 0 ou 1 ($p = 2$). Seu índice correspondente na matriz de probabilidades condicionais é calculado da seguinte forma:

$$1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 + 1 = 8 + 0 + 2 + 1 + 1 = 12$$

Note que caso o índice da tabela comece a partir de 0, o "+ 1" do final é dispensável. Este texto assume que o índice começa a partir de 1.

Outra observação importante é que para subespaços de dimensão grande de características, a tabela de probabilidades condicionais será muito grande, pois seu tamanho é exponencial em funçao de d. Porém, devido a uma propriedade da função critério proposta, não há necessidade de testar por subespaços de dimensão muito grande. A explicação deste fato encontra-se na discussão sobre a complexidade de tempo do algoritmo a seguir e no experimento da seção 6.1.1.

Complexidade de tempo (Ct)

O laço 5 executa t vezes os passos 6, 7 e 8. Como os passos 6, 7 e 8 combinados são $O(d)$, logo o laço 5 possui complexidade $O(td)$.

O laço 12 executa m vezes o passo 13 que por sua vez tem complexidade $O(c)$. Dentro desse laço, ou executa-se o passo 14 ou o passo 16 e 17. Mas tanto o passo 14 quanto os passos 16 e 17 combinados são $O(c)$. Portanto, o laço 12 que engloba também os passos de 13 a 17 tem complexidade $O(mc)$.

Então, somando as complexidades dos laços 5 e 12 com as dos passos de 1 a 4, 9 a 11, 18 e 19, temos:

$$Ct = O(d) + O(1) + O(mc) + O(m) + O(td) + O(m) + O(mc) + O(mc) = O(mc + td)$$
$$Ct = O(mc + td)$$

Como $m = p^d$, pode parecer que este algoritmo é exponencial tanto em complexidade de tempo como de memória, afinal m é o número de linhas da matriz P de probabilidades condicionais. Entretanto, devido a um resultado empírico discutido no experimento da seção 6.1.1 obtido como uma conseqüência do problema da dimensionalidade, o melhor subespaço de dimensão $d > log_p t$ nunca poderá superar o melhor subespaço de dimensão $d \leq log_p t$. Isto quer dizer que é inútil procurar pelo melhor subespaço entre os subespaços com dimensão maior que o logaritmo do número de amostras, sabendo que o melhor de acordo com o critério de entropia condicional média certamente não estará entre eles. Devido a isso, temos que $d = O(log_p t)$ e, portanto:

$$Ct = O(mc + td) = O(p^d c + td) = O(p^{log_p t} c + t log_p t) = O(tc + t log_p t) = O(t(c + log_p t))$$
$$Ct = O(t(c + log_p t))$$

Complexidade de memória (Cm)

Todas as variáveis e estruturas de dados adicionadas pelo algoritmo ocupam memória da ordem de $O(mc) = O(p^d c)$. Como visto na discussão sobre complexidade de tempo e da seção 6.1.1, $d = O(log_p t)$. Portanto a complexidade de memória do algoritmo é:

$$Cm = O(p^{log_p t} c) = O(tc)$$

5.2.3 Normalização e discretização

Em algumas situações, a normalização e a discretização são passos necessários do pré-processamento de um método de seleção de características que utiliza a entropia condicional média. Em particular, a normalização é necessária antes de aplicar qualquer método de seleção de características nos casos em que as características apresentam diferentes escalas de valores. Além disso, tal processo é útil para analisar os sinais produzidos pelos valores das características que possuem pequena variação ao longo das diferentes amostras.

Nos experimentos de análise de expressões gênicas, (vide seções 6.2.3 e 6.1.2), foi aplicada a *transformação normal* [21] como processo de normalização dos dados para analisar genes que apresentam perfis de expressão com pequena variação. A *transformação normal* η é dada por:

$$\eta[X] = \frac{X - E[X]}{\sigma[X]} \tag{5.10}$$

para toda a variável aleatória X, onde $E[X]$ e $\sigma[X]$ são, respectivamente, a média e o desvio padrão de X.

A transtormação normal tem duas propriedades importantes: 1) $E[\eta[X]] = 0$ e $\sigma[\eta[X]] = 1$, para toda variável aleatória X; 2) $\eta[X] = \lambda\eta[X]$, para todo número real λ.

Porém, em certas circunstâncias, deve-se tomar o cuidado de filtrar aqueles genes que possuem sinais de expressão praticamente constantes ao longo das amostras (os chamados *housekeeping genes*) antes da normalização, pois senão há o risco de adição de ruído na análise.

Dependendo da função critério, se os dados forem contínuos, será necessário submetê-los a um processo de discretização (ou quantização). Em particular, esse passo é fundamental com a utilização da entropia condicional média, pois a construção da tabela de probabilidades condicionais só é possível se os dados forem discretos. Essa etapa deve ser realizada com o devido cuidado, pois ela pode afetar severamente os resultados. Normalmente este passo é aplicado após a normalização.

Não há uma regra clara de como os dados devem ser discretizados e de quantos valores

discretos (grau de quantização) as características deverão assumir. Tais decisões deverão ser tomadas de acordo com o tipo de problema, com uma análise prévia dos dados e com base no número de amostras. Se o número de amostras for pequeno, por exemplo, o grau de quantização p deverá ser pequeno (2 ou 3 no máximo) para selecionar subespaços de características com um tamanho razoável, caso contrário, o número de linhas da tabela de probabilidades condicionais será grande demais para ser estimada com precisão (a quantidade de instâncias não observadas será muito alta). Devido a este problema, mesmo se os dados não forem discretos, pode ser que precisem ser submetidos a um grau de quantização menor caso o número de amostras não seja suficiente.

Capítulo 6

Experimentos e resultados

Alguns experimentos foram elaborados de tal forma a ilustrar a eficácia da entropia condicional média como função critério em problemas de redução de dimensionalidade. Na seção 6.1, são descritos experimentos e resultados obtidos em dados sintéticos, na análise de expressões gênicas e na construção de W-operadores aplicados a filtragem de imagens e reconhecimento de texturas. A seção 6.2 descreve uma técnica de MSV implementada pelo prof. Paulo J. S. Silva do IME-USP para identificação de genes fortes, bem como seus resultados. Dois experimentos que utilizam o método de entropia condicional para validar a técnica de MSV são descritos na seção 6.1.2.

6.1 Seleção de características por entropia condicional

6.1.1 Dados simulados

Características não relacionadas

Este experimento consiste em adotar um espaço $\mathbf{X}$ de n características onde nenhuma delas sejam responsáveis pelos valores que a variável Y pode assumir e estudar o comportamento da entropia condicional média com o aumento da dimensionalidade. Uma maneira de obter esse tipo de espaço é estabelecer um conjunto de amostras T contendo t pares $(\mathbf{x}, y)$ onde

cada $\mathbf{x}$ possua algum rótulo $Y = y$ com distribuição uniforme de probabilidades. Ou seja, $\mathbf{x}$ tem rótulo $Y = y$ com probabilidade $1/c$ para todo y tal que $1 \leq y \leq c$. Então, uma amostra é obtida através dos seguintes passos:

1. Escolha aleatoriamente, com distribuição uniforme, uma instância $\mathbf{x}$ de $\mathbf{X}$;
2. Obtenha aleatoriamente, com distribuição uniforme, o rótulo $Y = y$;
3. A amostra é o par $(\mathbf{x}, y)$.

Fixado T, aplicamos os algoritmos SFS e de busca exaustiva para cada uma das dimensões de um espaço total de 8 características, gerando os subespaços $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_8}$, cada um com o seu valor de entropia. Executamos diversas vezes esse tipo de experimento variando a quantidade de amostras, o número p de valores possíveis para cada característica e o número c de classes possíveis. Para cada uma dessas execuções geramos dois gráficos com as dimensões d de $\mathbf{X_d}$ ($\forall d, 1 \leq d \leq n$) representadas pela abscissa e os valores $E[H(Y|\mathbf{X_d})]$ representados pela ordenada, resultantes da aplicação do SFS e da busca exaustiva.

A figura 6.1 mostra o resultado de duas execuções, onde cada uma gerou dois gráficos: um decorrente da aplicação do SFS e o outro da aplicação da busca exaustiva. Em ambas as execuções, o número de valores que cada característica pode assumir é $p = 3$ e o número de classes distintas é $c = 3$. A diferença é que na execução 1 foram utilizadas $t = 3^4 = 81$ amostras, enquanto a execução 2 utilizou $t = 3^6 = 729$ amostras

Discussão

Um resultado empírico interessante evidenciado de diversas execuções desse experimento é que, para os casos onde não há subespaço de características relacionado com Y, o gráfico $d \times E[H(Y|\mathbf{X_d})]$ é dado por uma "curva em U" em que o ponto de mínimo ocorre exatamente na dimensão $D_{min} = log_p t$, fixando-se $\alpha = 1$. Ou seja, mesmo que exista um subespaço $\mathbf{X_d}$ com d maior que $log_p t$ que prediz Y com boa confiança, não será possível identificar esse subespaço como um bom preditor devido a uma quantidade insuficiente de amostras (o número de instâncias não observadas de $\mathbf{X_d}$ é muito alto neste caso). Isto ocorre justamente devido ao problema da dimensionalidade. Porém, é de se esperar que, em muitos casos, a maior parte das características do subespaço selecionado sejam pertencentes também a $\mathbf{X_d}$.

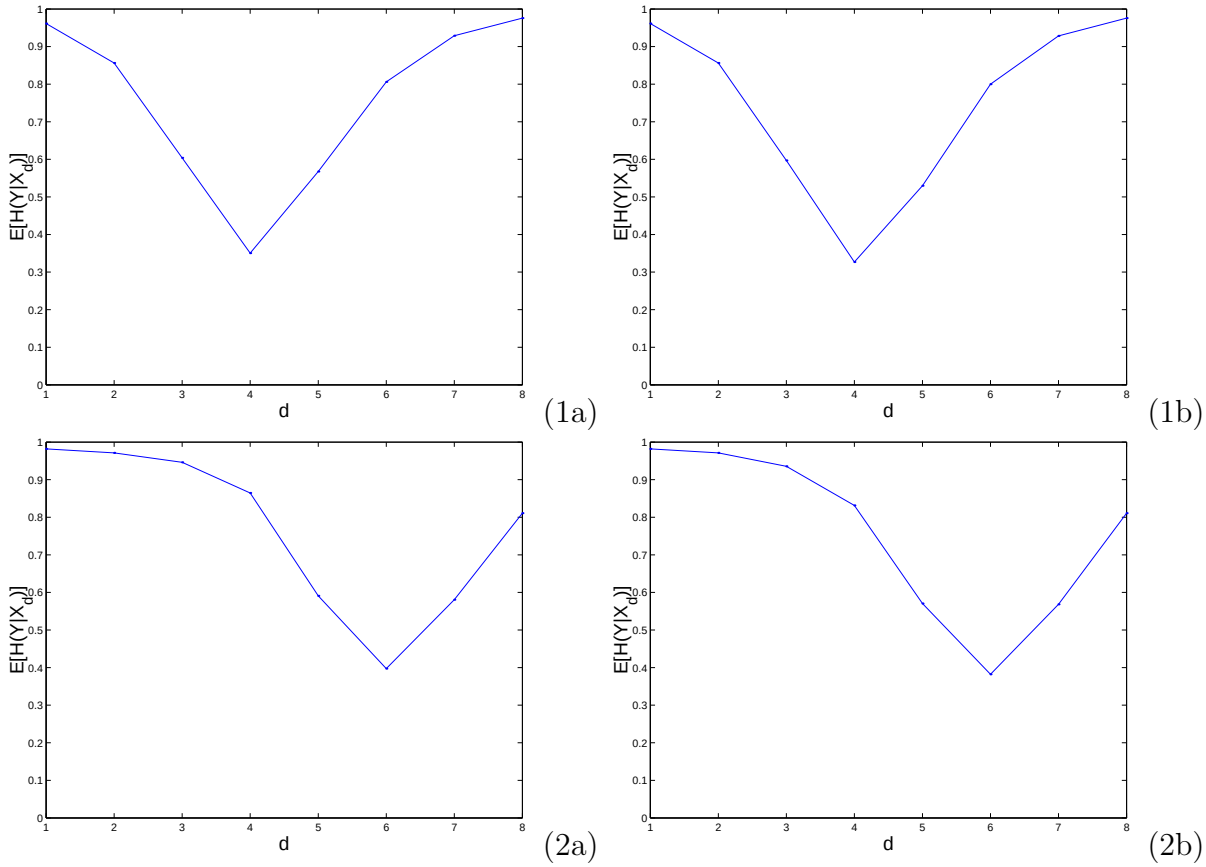


Figura 6.1: Gráficos de entropia condicional média em função da dimensão d de $\mathbf{X_d}$ sem características relacionadas. Cada característica pode assumir 3 valores possíveis, sendo que existem 3 classes possíveis. (1a) 81 amostras, SFS; (1b) 81 amostras, busca exaustiva; (2a) 729 amostras, SFS; (2b) 729 amostras; busca exaustiva.

Uma conseqüência desejável desse fato é que o espaço de busca se torna drasticamente reduzido, bastando verificar apenas os subespaços $\mathbf{X_d} \subseteq \mathbf{X}$ cujo tamanho seja menor ou igual a $log_p t$. Em problemas de bioinformática, por exemplo, o número de amostras é tipicamente muito menor que o tamanho de $\mathbf{X}$ (conjunto de genes), tornando factível esta abordagem. E mesmo em problemas de processamento de imagens que envolvem busca de bons W-operadores onde o número de amostras é muito maior que o número de características (tamanho da janela), ainda assim é factível utilizar essa abordagem, já que a máxima dimensão para que se obtenha o mínimo da "curva em U" cresce em escala logarítmica.

Características relacionadas

Este experimento assemelha-se ao anterior, exceto que agora há dois subespaços $\mathbf{X_r}$ e $\mathbf{X_{nr}}$, ambos contidos em $\mathbf{X}$. $\mathbf{X_r}$ é o subespaço de características relacionadas com Y e $\mathbf{X_{nr}}$ é o subespaço de características não relacionadas com Y. Para estabelecer esta relação, definimos que cada possível instância $\mathbf{x_r}$ de $\mathbf{X_r}$ determina um rótulo $Y = y$ com distribuição normal de probabilidade com média μ, $1 \leq \mu \leq c$ (μ sorteada com distribuição uniforme para cada $\mathbf{x_r}$) e desvio padrão pequeno σ ($\sigma = 0.2$ em nossos experimentos) da seguinte forma:

1. se o número aleatório j sorteado for menor que 0.5, faça $j \leftarrow j + c$ e vá para o passo 3;
2. se o número aleatório j sorteado for maior que $c + 0.5$, faça $j \leftarrow j - c$ e vá para o passo 3;
3. o rótulo $Y = y$ sorteado será tal que o número sorteado j esteja no intervalo $[y - 0.5, y + 0.5]$

Logo, isto quer dizer que toda instância $\mathbf{x_r}$ gera o rótulo $Y = \mu$ com uma probabilidade bem alta ($H(Y|\mathbf{x_r})$ possui valor baixo). Então, de fato, $\mathbf{X_r}$ é um bom preditor de Y.

Assim, cada amostra é gerada através do seguinte procedimento:

1. Sorteie, com distribuição uniforme, uma instância $\mathbf{x_r}$ de $\mathbf{X_r}$;
2. Obtenha o rótulo $Y = y$ aplicando a função de probabilidade de $\mathbf{x_r}$ descrita acima;
3. Sorteie, com distribuição uniforme, uma instância $\mathbf{x_{nr}}$ de $\mathbf{X_{nr}}$;
4. Concatene $\mathbf{x_r}$ com $\mathbf{x_{nr}}$ resultando em $\mathbf{x}$;
5. A amostra é o par $(\mathbf{x}, y)$.

Fixando T onde $\mathbf{X_r}$ tem dimensão 4, isto é, existem 4 características relacionadas com Y, aplicamos os algoritmos SFS e de busca exaustiva para cada uma das dimensões de um espaço total de 8 características, gerando os subespaços $\mathbf{X_1}$, $\mathbf{X_2}$, ... , $\mathbf{X_8}$, cada um com

o seu valor de entropia, como foi feito no experimento da seção anterior (6.1.1). Assim como na seção anterior, executamos esse tipo de experimento variando a quantidade de amostras, o número p de valores possíveis para cada característica e o número c de classes possíveis. Para cada uma execução, geramos dois gráficos com as dimensões d de $\mathbf{X_d}$ ($\forall d, 1 \leq d \leq n$) representadas pela abscissa e os valores $E[H(Y|\mathbf{X_d})]$ representados pela ordenada, resultantes da aplicação do SFS e da busca exaustiva.

A figura 6.2 mostra o resultado de duas execuções típicas, onde cada uma gerou dois gráficos: um decorrente da aplicação do SFS e o outro da aplicação da busca exaustiva. Em ambas as execuções, o número de valores que cada característica pode assumir é $p = 3$ e o número de classes distintas é $c = 3$. A diferença é que na execução 1 foram utilizadas $t = 3^4 = 81$ amostras, enquanto a execução 2 utilizou $t = 3^6 = 729$ amostras

Discussão

Diferentemente do experimento anterior, o aumento do número de amostras (incremento de t) tende a transformar a "curva em U" em uma "curva em L", onde o ponto de inflexão ocorre exatamente no ponto de dimensão de $\mathbf{X_r}$ (denote q o valor dessa dimensão). Neste caso, algumas dimensões posteriores a q podem resultar em valores de entropia bastante próximos ao valor dado por q, podendo até ser ligeiramente menores.

Um ponto importante a ser observado aqui é que eventualmente os métodos de busca não exaustivos (SFS por exemplo), podem englobar características que não pertencem ao subespaço relacionado com Y (efeito *nesting*). Caso isto ocorra, o ponto de mínimo da "curva em U" acaba sendo obtido em um ponto posterior a q (desde que haja um número suficiente de amostras) tal que todas as características que pertençam ao subespaço relacionado com Y estejam incluídas no conjunto solução do SFS. Caso o número de amostras não seja suficiente, é possível que características do subespaço relacionado não façam parte do conjunto solução do SFS. O surgimento desse efeito depende basicamente de três fatores: da quantidade e qualidade de amostras, do grau de predição do melhor subespaço e do grau de predição de cada uma das características do melhor subespaço.

Com relação aos valores de entropia para $\mathbf{X_r}$, percebe-se que há uma diminuição brusca de seu valor (ponto de mínimo) na figura 6.2 quando a quantidade de amostras passou de 81 para 6561 amostras. Seu valor de entropia caiu de cerca de 0.2 para cerca de 0.05. O mesmo não ocorre no experimento da seção anterior (ver figura 6.1), onde os valores de entropia para os melhores subespaços obtidos ficavam oscilando entre 0.3 e 0.4 indiferentes

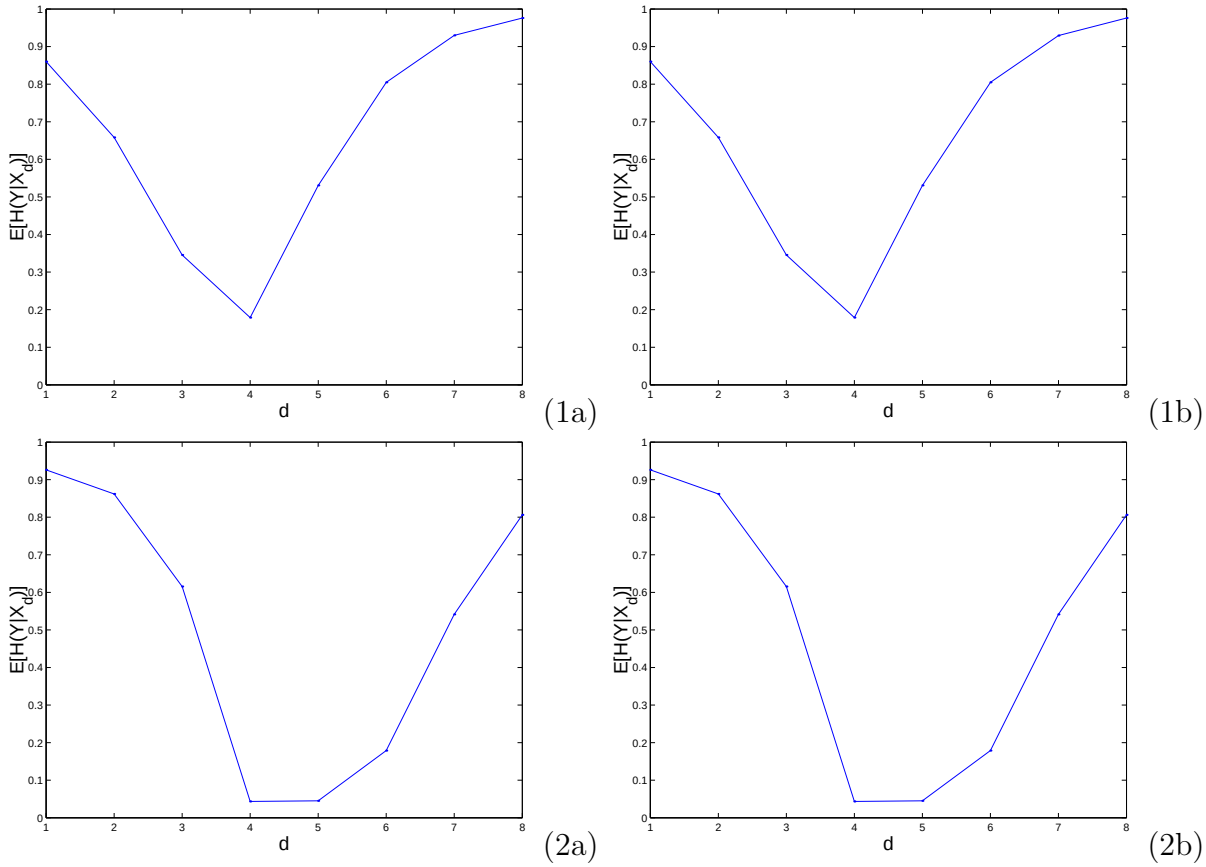


Figura 6.2: Gráficos de entropia condicional média em função da dimensão d de $\mathbf{X_d}$ com 4 características relacionadas. Cada característica pode assumir 3 valores possíveis, sendo que existem 3 classes possíveis. (1a) 81 amostras, SFS; (1b) 81 amostras, busca exaustiva; (2a) 729 amostras, SFS; (2b) 729 amostras; busca exaustiva.

ao número de amostras. A conclusão disso é que nos casos onde existem subespaços que são bons preditores de Y, a tendência de um classificador obtido por este método é de aprender cada vez mais sobre esses subespaços a medida em que se aumenta o número de amostras. Caso não haja subespaços desse tipo, o classificador tende a incluir o máximo de características que ele consegue até que comece a piorar o erro de estimação. Neste caso, as características formarão um subespaço que prediz muito pouco sobre os valores de Y.

Independência de geometria

Este experimento mostra que a abordagem desenvolvida não prioriza apenas as características que separam linearmente as classes. Para mostrar esta propriedade, geramos dados sintéticos com 2 características (x_1, x_2), em que cada uma pode assumir 3 valores (0, 1, 2) e $Y \in \{0, 1\}$. Logo, existem $3^2 = 9$ instâncias possíveis: {(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2)}. Elaboramos três exemplos nos quais a entropia condicional média vale zero, embora dois deles não sejam linearmente separáveis no espaço. Veja os exemplos a seguir:

1. **Grupos linearmente separáveis:** sejam 9 amostras, 1 para cada instância possível, em que $y = 0$ para as instâncias (0,0), (0,1), (0,2), (1,0), (1,1), (2,0) e $y = 1$ para as instâncias (1,2), (2,1), (2,2). A entropia de cada instância é zero já que toda instância corresponde a um valor único para y. Portanto, a entropia condicional média é zero (fig. 6.3).

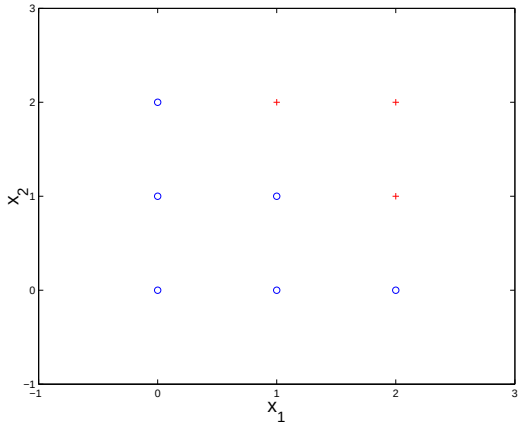


Figura 6.3: Exemplo de grupos linearmente separáveis em que $E[H(Y|\mathbf{X})] = 0$. Os símbolos "o" e "+" indicam as amostras das suas respectivas classes.

2. **Grupos côncavos:** sejam 9 amostras, 1 para cada instância, onde $y = 0$ para as instâncias (0,0), (1,0), (1,1), (2,0) e $y = 1$ para as instâncias (0,1), (0,2), (1,2), (2,1), (2,2). Estes grupos não são linearmente separáveis, pois não é possível traçar uma reta que os separe no plano (fig. 6.4). Mesmo assim, a entropia condicional média é zero, já que cada instância possível corresponde a uma única classe, não havendo mistura.

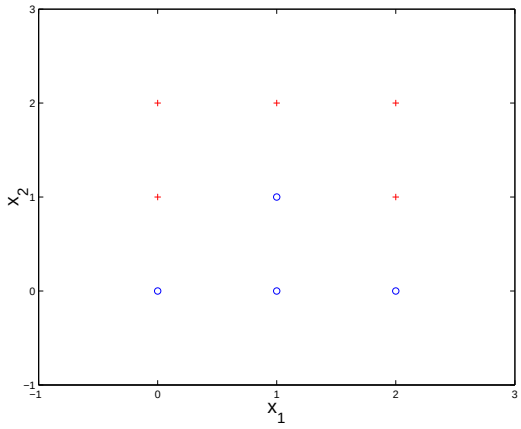


Figura 6.4: Exemplo de grupos côncavos em que $E[H(Y|\mathbf{X})] = 0$. Os símbolos "o" e "+" indicam as amostras das suas respectivas classes.

3. **Grupos envolventes:** sejam 9 amostras, 1 para cada instância, onde $y = 0$ para as instâncias (0,0), (0,1), (0,2), (1,0), (1,2), (2,0), (2,1), (2,2) e $y = 1$ para a instância (1,1). O primeiro grupo envolve o segundo, como pode ser visto na figura 6.5. Obviamente, estes grupos não são linearmente separáveis. Do mesmo modo que nos dois exemplos anteriores, a entropia condicional média é zero, pois cada instância corresponde a apenas uma única classe, não havendo mistura.

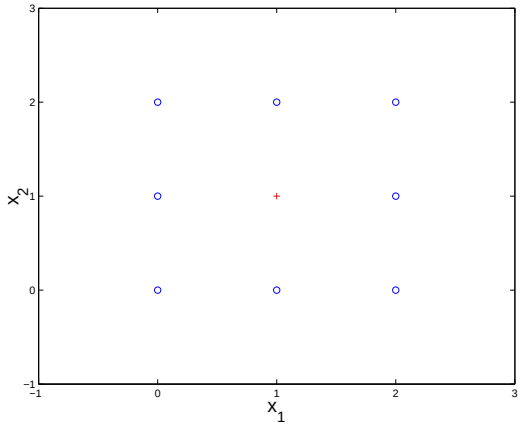


Figura 6.5: Exemplo de grupos envolventes em que $E[H(Y|\mathbf{X})] = 0$. Os símbolos "o" e "+" indicam as amostras das suas respectivas classes.

Discussão

Este experimento ilustra que o critério de entropia privilegia certos subespaços de acordo com a quantidade de informação que eles fornecem sobre o comportamento da variável de classe, mesmo que essas classes não sejam linearmente separáveis. Portanto este critério é mais geral que os critérios baseados em distância que costumam privilegiar apenas subespaços linearmente separáveis.

Mistura × Entropia

Seja $\mathbf{X}$ com 2 características (X_1, X_2), cada uma podendo assumir 3 valores $(0, 1, 2)$ e $Y \in \{0, 1\}$ como no experimento anterior. Geramos t amostras através do seguinte procedimento:

1. Para cada instância (i, j), $0 \leq i \leq 2$ e $0 \leq j \leq 2$, crie $t/9$ amostras da seguinte maneira: se $(i, j) \neq (1, 1)$ então a amostra será o par $((i, j), 0)$; senão a amostra será o par $((i, j), 1)$. Assim teremos um grupo interno a outro, como já foi ilustrado na figura 6.5 do experimento anterior.

2. Para cada amostra $((i, j), y)$ criada no passo anterior faça $(i, j) \leftarrow (i', j')$, tal que (i', j') esteja em $\Re^2$ e que seja obtido por distribuição normal com $(\mu'_i = i, \mu'_j = j)$ e um σ fixo.

3. Para cada amostra $((i, j), y)$ obtida do passo anterior, discretize i e j do seguinte modo. Se $i < 0.5$ então $i \leftarrow 0$; se $0.5 \leq i \leq 1.5$ então $i \leftarrow 1$; se $i > 1.5$ então $i \leftarrow 2$. Faça o mesmo para j.

Este procedimento aplicado a diferentes valores de σ influencia o valor de $E[H(Y|\mathbf{X})]$ do seguinte modo: quanto maior o valor de σ, maior será $E[H(Y|\mathbf{X})]$ (fig. 6.6). Esta relação não é bem definida para um número pequeno de amostras, mas quanto maior o número de amostras, melhor definida será a relação $\sigma_1 \leq \sigma_2 \rightarrow E[H(Y|\mathbf{X})]_1 \leq E[H(Y|\mathbf{X})]_2$. A figura 6.7 apresenta um gráfico representando uma superfície que ilustra esta relação. Neste gráfico, o número de amostras é dado no eixo X, o valor de σ é dado no eixo Y e a entropia condicional média é dada no eixo Z. Um aumento do número de amostras faz com que a curva no plano (σ, $E[H(Y|\mathbf{X})]$) tenha um comportamento crescente mais bem definido (com menos solavancos). Ou seja, quanto maior a mistura (σ), maior a entropia condicional média.

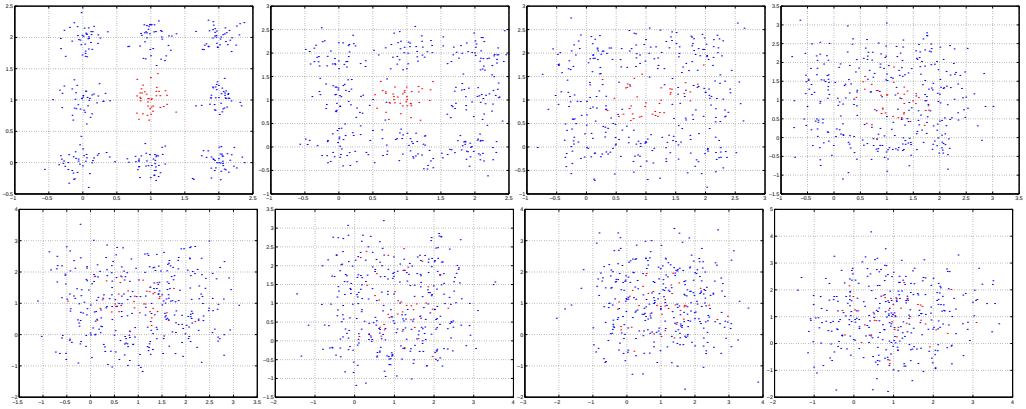


Figura 6.6: Gráficos ($x_1 \times x_2$) com 360 amostras e σ variável. As amostras pertencentes à primeira classe são representadas pela cor azul e as amostras da segunda classe pela cor vermelha. Valores de $(\sigma, E[H(Y|\mathbf{X})])$ respectivamente em ordem de varredura: (0.16, 0), (0.24, 0.0436), (0.32, 0.2263), (0.40, 0.2993), (0.48, 0.3933), (0.56, 0.4602), (0.64, 0.4639), (0.72, 0.4698).

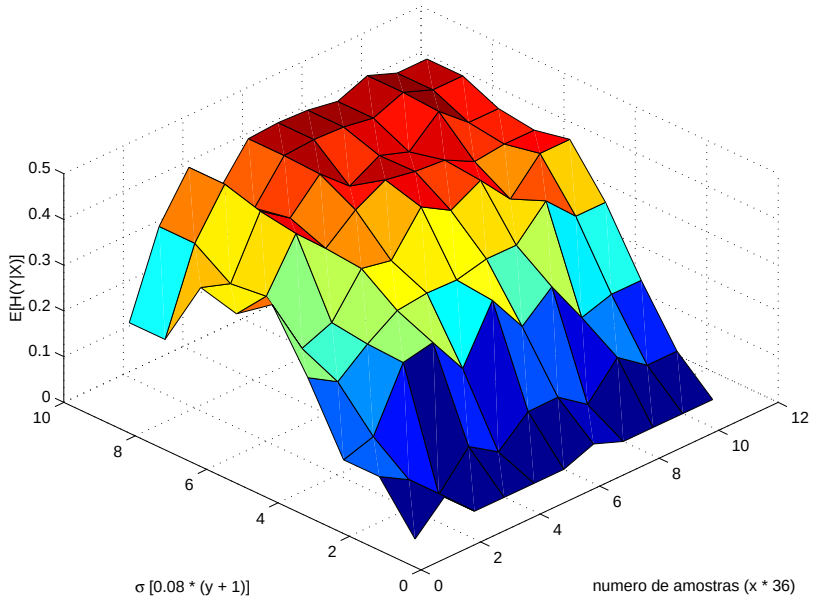


Figura 6.7: Superfície em que o número de amostras é dado no eixo X, o valor de σ é dado pelo eixo Y e $E[H(Y|\mathbf{X})]$ é representado pelo eixo Z.

Discussão

Este experimento mostrou claramente a forte relação da entropia com o grau de mistura entre as classes. Quanto mais misturadas as classes, maior a entropia. E, conseqüentemente, quanto maior a entropia, menos informação teremos sobre as classes utilizando o subespaço considerado. Por outro lado, subespaços que definem classes mais compactas e melhor separadas terão entropia menor, sendo fortes candidatas a serem selecionadas por um método que busque minimizar a entropia.

6.1.2 Análise de expressões gênicas

A técnica proposta está sendo aplicada no tratamento de dois problemas de análise de expressões gênicas:

- Identificação de subconjuntos de genes que melhor separam dois estados biológicos distintos;
- Identificação de arquitetura de redes de regulação gênica.

Identificação de genes que separam dois estados biológicos distintos

Dois experimentos que utilizam a função critério proposta neste trabalho (entropia condicional média) aplicados aos mesmos dados de SAGE provenientes de uma colaboração com a pesquisadora Helena Brentani do Ludwig Institute for Cancer Research foram realizados com o objetivo de validar os resultados obtidos pelo sistema de identificação de genes fortes através de MSV (ver seção 6.2.3 para maiores detalhes sobre esse sistema e sobre os dados de entrada). Ambos os experimentos foram realizados sobre o mesmo conjunto de dados e a mesma questão biológica formulada na seção 6.2.3: quais genes melhor distinguem tecidos com glioblastoma dos tecidos com astrocitoma graus II e III?.

Em ambos os experimentos, para calcular a entropia condicional média sobre um determinado subconjunto de genes, foi introduzida uma etapa de discretização das expressões logo após a etapa de normalização. Tal passo não é necessário caso seja aplicada a técnica de MSV, mas é fundamental no cálculo das entropias. Então, para cada gene, foi atribuído 0 às expressões normalizadas que tinham valor negativo (isto é, que tinham valor abaixo de suas próprias expressões), e atribuído 1 caso contrário. Portanto, a discretização adotada foi de grau 2 (dois valores possíveis para cada expressão). O resultado

final da discretização é uma matriz Booleana que serve como entrada para calcular a entropia condicional média. A fórmula utilizada no cálculo é dada pela equação 5.8.

Primeiro experimento

Foi feito o mesmo processo descrito na seção 6.2.3 a fim de gerar as 1000 melhores trincas através da técnica de MSV. Feito isso, a entropia condicional média foi calculada para cada uma das trincas, tendo como base a matriz discretizada de expressões como mencionada anteriormente.

Os valores mínimo, médio e máximo das 1000 entropias condicionais médias calculadas foram, respectivamente: 0, 0.1009, 0.4091 (tabela 6.1). Note que numa escala de 0 a 1, a média obtida possui um valor relativamente pequeno. Para se ter uma idéia de quão pequeno é este valor, sorteamos 1000 trincas ao acaso (distribuição uniforme) dentre todas as possíveis trincas e calculamos a entropia condicional média para cada uma delas. Os valores mínimo, médio e máximo destes valores foram, respectivamente: 0.0909, 0.7654, 0.9989. Como menores entropias condicionais médias resultam em melhores subespaços de características, estes números servem de validação da técnica de MSV aplicada nesse contexto. A figura 6.8 consolida essa validação, mostrando os gráficos para a freqüencia de trincas em função das entropias condicionais médias em ambos os casos.

	1000 melhores trincas por MSV	1000 trincas escolhidas ao acaso
ECM_{min}	0	0.0909
ECM_{med}	0.1009	0.7654
ECM_{max}	0.4091	0.9989

Tabela 6.1: Valores mínimo, médio e máximo dentre as entropias condicionais médias das 1000 melhores trincas obtidas pela técnica MVS e de 1000 trincas sorteadas uniformemente.

Segundo experimento

Este experimento consistiu em selecionar os 10 melhores genes individualmente pelo critério da entropia condicional média. Desses 10 genes, verificou-se que 8 estavam entre as 1000 melhores trincas escolhidas pela técnica MSV. E desses 8 genes, 2 separam completamente os dois tipos de tumor considerado (entropia condicional média nula). Além disso, o terceiro melhor gene de acordo com o critério de entropia é aquele que aparece

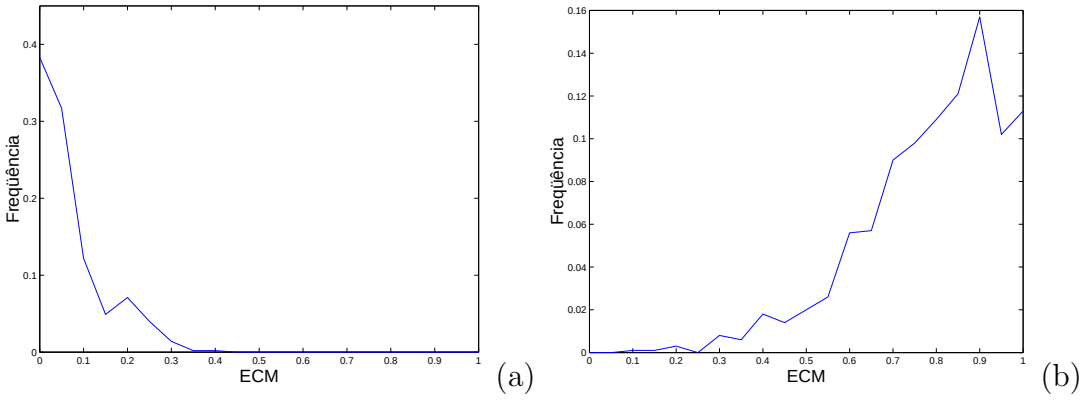


Figura 6.8: Gráficos da freqüência de trincas em função da entropia condicional média: (a) para as 1000 melhores trincas obtidas por MSV; (b) para 1000 trincas selecionadas ao acaso.

na melhor trinca, além de ser o mais freqüente entre as 1000 melhores trincas obtidas por MSV. A figura 6.27 da seção 6.2.3 mostra uma tabela das 10 melhores trincas, onde o gene mais freqüente (BC014549) aparece na coluna chamada "Y" da melhor trinca. Note que este gene aparece em 350 das 1000 trincas (coluna chamada "Py" para a melhor trinca).

A figura 6.28 mostra o gráfico tridimensional para a melhor trinca. Esse gráfico mostra claramente que o gene BC014549 é, de fato, o maior responsável pela separação dos grupos glioblastoma de astrocytoma graus II e III. Finalmente, é importante observar no gráfico da figura 6.32 que a classificação de 4 bibliotecas novas de astrocytoma grau II foi correta principalmente devido ao gene em questão.

A conclusão deste experimento e do anterior é que a técnica de entropia condicional média corroborou a técnica de seleção de genes fortes por MSV como uma boa ferramenta para identificar genes que distinguem dois estados biológicos.

Identificação de arquitetura de redes de regulação gênica

A técnica de seleção de características por entropia condicional média proposta neste trabalho está sendo aplicada em busca de identificar a arquitetura da rede de regulação gênica para os dados de *microarray* obtidos do genoma seqüenciado do *Plasmodium falciparum*, um agente parasita causador da malária [15]. Este trabalho está sendo desenvolvido em

conjunto com outros pesquisadores do IME-USP e também em conjunto com a equipe do Prof. Dr. Hernando Del Portillo do Instituto de Ciências Biomédicas (ICB-USP) [7].

O seqüenciamento completo do genoma do *Plasmodium falciparum* revelou que aproximadamente 60% do genoma anotado correponde a proteínas hipotéticas e que muitos genes, cujas vias metabólicas ou produtos biológicos são conhecidos bioquimicamente, não foram preditos. Recentemente, através do uso da técnica de Transformada Discreta de Fourier (DFT - Discrete Fourier Transform), foi sugerido que os parasitas seguem um rígido programa de relógio [15]. Então, uma nova lista de genes codificantes com funções biológicas semelhantes aumentaram significativamente novos alvos para vacinas e desenvolvimento de drogas. Nossa proposta é anotar genes sob uma diferente perspectiva: uma lista de propriedades funcionais é atribuída a redes de genes representando subsistemas do sistema regulatório de expressão do parasita [7].

O modelo adotado para representar redes gênicas é a Rede Gênica Probabilística (PGN - Probabilistic Genetic Network). Esta rede é uma cadeia de Markov com algumas propriedades adicionais. O modelo imita as propriedades de um gene como uma "porta" estocástica não-linear e os sistemas construídos pelo acoplamento dessas portas. O objetivo desta pesquisa é estimar uma PGN [27] representando um subsistema da rede de expressão gênica do parasita a partir de medidas de expressões de *microarray* (inicialmente obtidas do arquivo *QC dataset*[1] produzido pelo trabalho de DeRisi *et al* [15]). A estimativa do PGN é feita através da minimização da entropia condicional média (ou, equivalentemente, maximização da informação mútua) na busca de descobrir um subconjunto de genes que melhor prediz um determinado gene alvo Y no instante de tempo posterior.

O arquivo *QC dataset* consiste em uma matriz com 5080 genes por 46 instantes de tempo. Cada célula (i, j) da matriz representa a expressão do gene i no instante j. Para validar a metodologia proposta, estudamos a via glicolítica, uma via metabólica bem conhecida. Antes de aplicar a técnica de estimação de preditores (entropia condicional média), os sinais das expressões foram normalizados e discretizados.

Antes da discretização, os sinais são normalizados pela *transformação normal* [21] (ver equação 5.10 na seção 5.2). A discretização de um gene em um dado instante é um mapeamento do logaritmo da expressão contínua em três níveis de expressão (-1, 0,

[1]Este arquivo pode ser encontrado em `http://www.camda.duke.edu/camda04/datasets`

+1), respectivamente, sub-expresso, normal e super-expresso em relação à referência. A discretização do sinal de um gene g sobre todos os instantes de tempo t é realizado através de um mapeamento por limiares inferior (inf) e superior (sup) dado por:

- $g(t) \leftarrow -1$ se $g(t) < inf$
- $g(t) \leftarrow 0$ se $inf \leq g(t) \leq sup$
- $g(t) \leftarrow +1$ se $g(t) > sup$

A normalização e a discretização têm o efeito de criar classes de equivalência entre sinais amenizando os erros de estimação devido a falta de um maior número de amostras.

A capacidade preditória do modelo proposto foi testada para escolher genes alvo que codificam enzimas pertencentes à via glicolítica. Esses genes têm sinais quase senoidais e foram agrupados no estado de anel (*ring*) do ciclo de vida do parasita de acordo com o faseograma (ordenação dos sinais dos genes através de suas fases) produzido pelo DFT [15] (ver figura 6.9).

Neste experimento de predição, todos os 5080 elementos do QC dataset foram considerados como possíveis preditores de 8 genes alvos da glicólise. Para cada alvo, foi computada a informação mútua para a combinação de todas as duplas de genes do genoma e os 5 melhores foram selecionados, isto é, dentre as 12.900.660 duplas de genes (combinação de 5080, 2 a 2), as cinco melhores foram escolhidas. A figura 6.10(B) mostra três dos cinco genes (n132_136, j647_6 e c305) em que aparece o gene n132_136 fazendo parte das cinco melhores duplas que predizem i13056_1. A via metabólica da glicólise apresentada na figura 6.10(A) mostra que a predição está correta. Além disso, note que o outro gene que compõe a melhor dupla com n132_136 (j647_6) tem um sinal não senoidal como mostrado na figura 6.10(C), tendo sido descartado pela abordagem DFT [15].

Os outros alvos não foram preditos com a mesma precisão mas se o número de preditores considerados aumentar, eles aparecem logo. Os 400 melhores preditores individuais (isto é, apenas um gene predizendo um outro gene) para cada gene foram calculados. Todos os 8 genes alvos considerados foram verificados. É importante observar que o número de genes considerados necessários para encontrar o preditor certo de um gene alvo é relacionado com as suas posições no faseograma.

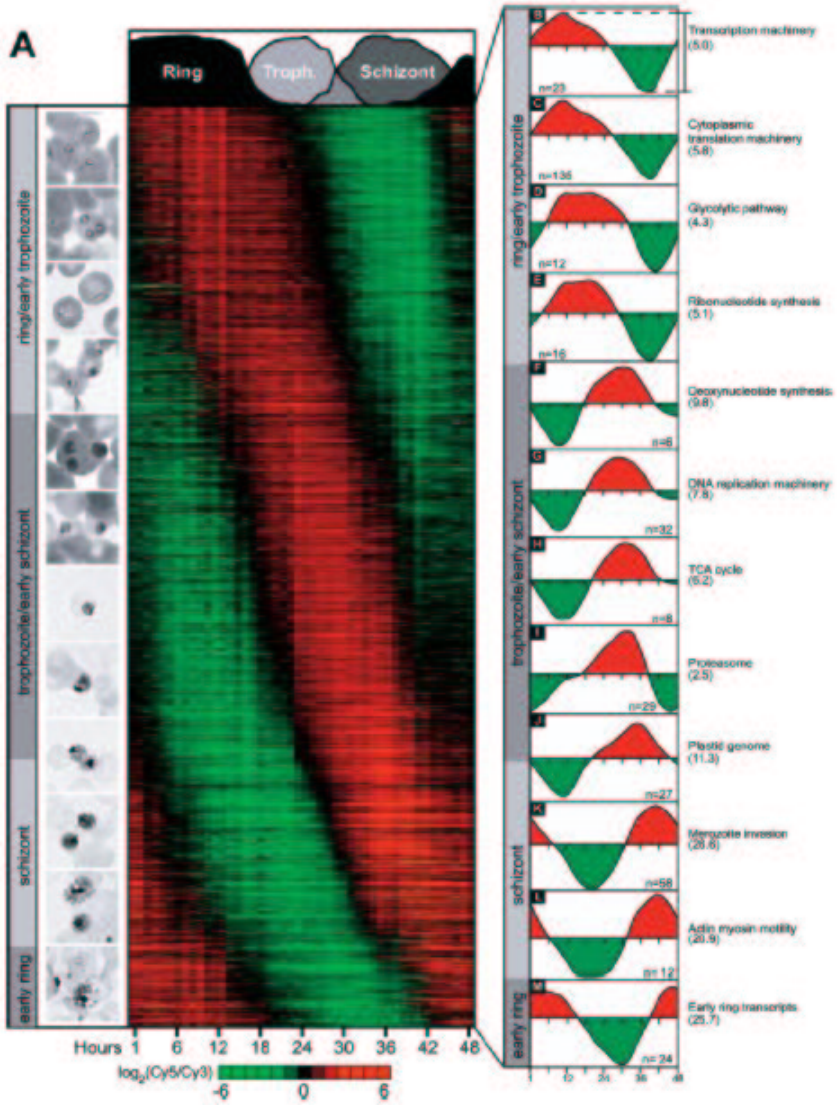


Figura 6.9: Faseograma produzido pela técnica DFT (reproduzido de [15]).

6.1.3 Imagens

Neste trabalho aplicamos a metodologia proposta (seleção de características por entropia condicional média) para estimar um operador de espaço restrito dos dados de treinamento [53]. A idéia é estimar uma sub-janela ótima W^* que maximiza a informação sobre a distribuição conjunta desconhecida de uma dada janela W e dos dados de treinamento disponíveis das formas obtidas de W. Escolher uma sub-janela é equivalente a agrupar exemplos dos dados de treinamento, já que formas diferentes podem se tornar a mesma

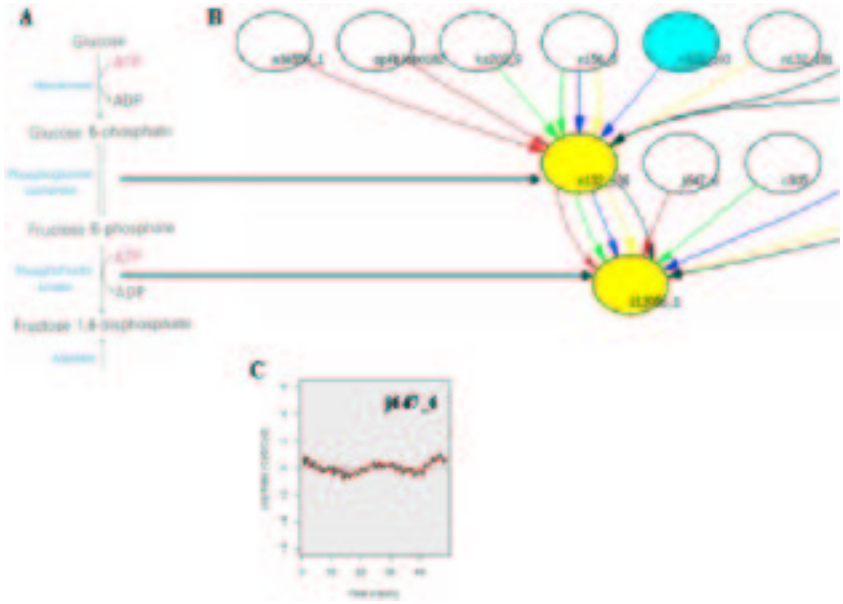


Figura 6.10: Capacidade preditória da PGN na via metabólica da glicólise. (A) Etapas iniciais da glicólise até a formação de Aldolase. (B) Grafo parcial mostrando as melhores duplas de combinações (setas vermelhas) que predizem phosphofructokinase. (C) Expressão temporal do gene PF10_0097 (oligo j647_6), não senoidal e que não foi incluído pela abordagem DFT [15].

quando vistas pela sub-janela. Isto, por um lado, diminui o erro de estimação das distribuições conjuntas pelo fato de tornar equivalentes formas raramente observadas e, por outro lado, aumenta o erro de estimação introduzindo ruído na classificação da forma. A melhor janela W^* deve balancear adequadamente esses efeitos.

A seção 4.3 explicou a importância de selecionar uma forma para a janela W na construção de um W-operador. Para explorar o conceito de entropia, as posições da janela que projetam a sua forma são tomadas como variáveis (características) que compõem um vetor aleatório $\mathbf{X}$. Então, construir W pode ser visto como um problema de seleção de características que usa $E[H(Y|\mathbf{X}_{\mathcal{Z}})]$ como uma função critério. A figura 6.11 ilustra esse conceito, mostrando duas formas possíveis para W. Nesta figura, as características selecionadas $\mathbf{X}_{\mathcal{Z}}$ são indicadas como células pretas, ou seja, 5 características foram selecionadas em 6.11(a) enquanto que 13 foram selecionadas em 6.11(b)

Nesta seção serão apresentados alguns resultados de filtragem de imagens e de reconhe-

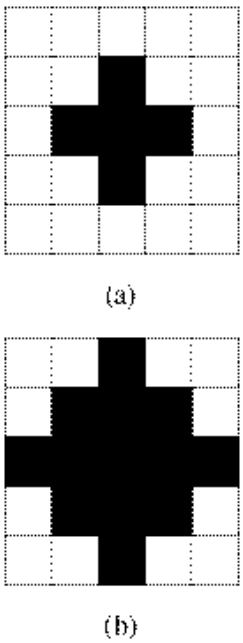


Figura 6.11: Janelas com (a) 5 características e (b) 13 características.

cimento de texturas obtidos da aplicação do algoritmo SFS de seleção de características, adotando a entropia condicional média como função critério para avaliar W-operadores. A rigor, utilizamos um algoritmo SFS levemente modificado. Ao invés de fornecer o parâmetro d (dimensão) como critério de parada para o algoritmo, fizemos com que ele incluísse características até que a entropia condicional média não pudesse ser melhorada, ou seja:

SFS-modificado($\mathbf{X}$)
$\mathbf{Z}' \leftarrow \phi$;
REPITA
  $\mathbf{Z} \leftarrow \mathbf{Z}'$;
  $\mathbf{Z}' \leftarrow \mathbf{Z} \cup \{X_i : \mathcal{F}(\mathbf{Z} \cup X_i) = min(min_{1 \leq j \leq n} \mathcal{F}(\mathbf{Z} \cup X_j), \mathcal{F}(\mathbf{Z})), \forall X_j \notin \mathbf{Z}\}$
ATÉ QUE ($|\mathbf{Z}'| = d$ OU $\mathbf{Z}' = \mathbf{Z}$)
DEVOLVA $\mathbf{Z}'$

Note, no procedimento acima, que a função critério $\mathcal{F}$ utilizada nos experimentos é a entropia condicional média $E[H(Y|\mathbf{X}_{\mathcal{Z}})]$ definida pela equação 5.8. Portanto, buscaremos minimizá-la para obter W. A estimação da equação 5.8 é baseada em um conjunto

de treinamento consistindo de pares de imagens original/ideal de uma forma análoga à construção de operadores de imagens PAC [10, 53].

Filtragem de imagens ruidosas

A aplicação da entropia condicional média para construir W-operadores foi cuidadosamente analisada em diversos experimentos de filtragem de imagens binárias, uma etapa fundamental em análise de formas [21]. Ruído sal e pimenta foi adicionado às imagens binárias, e W-operadores para filtrar essas imagens ruidosas foram gerados usando a metodologia criada. A figura 6.12 apresenta as imagens originais e um exemplo de suas respectivas versões com ruído sal e pimenta 10% (ou seja, cada pixel foi alterado com 10% de probabilidade) utilizados nos experimentos. A figura 6.13 mostra uma imagem de partitura musical e um exemplo dessa imagem com ruído sal e pimenta de 3%.

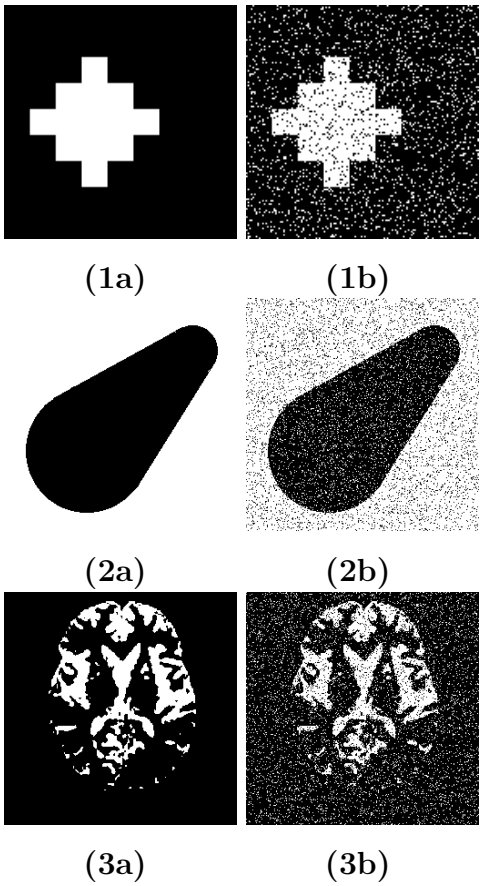
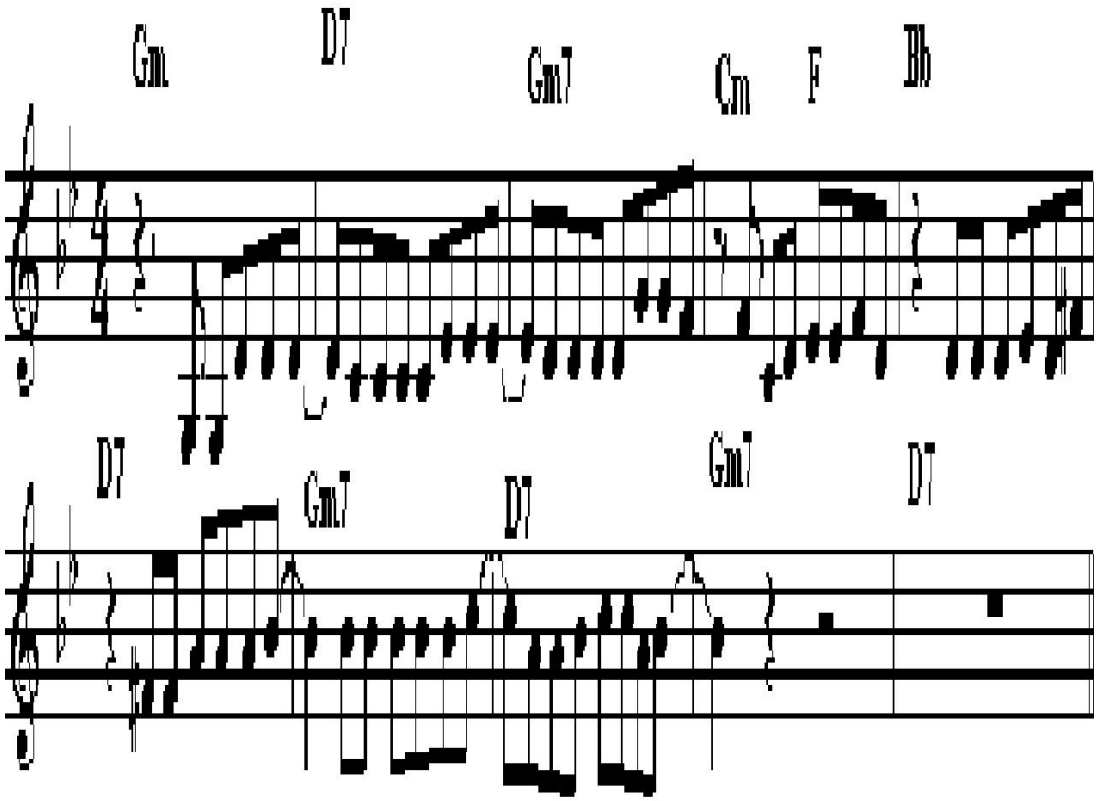
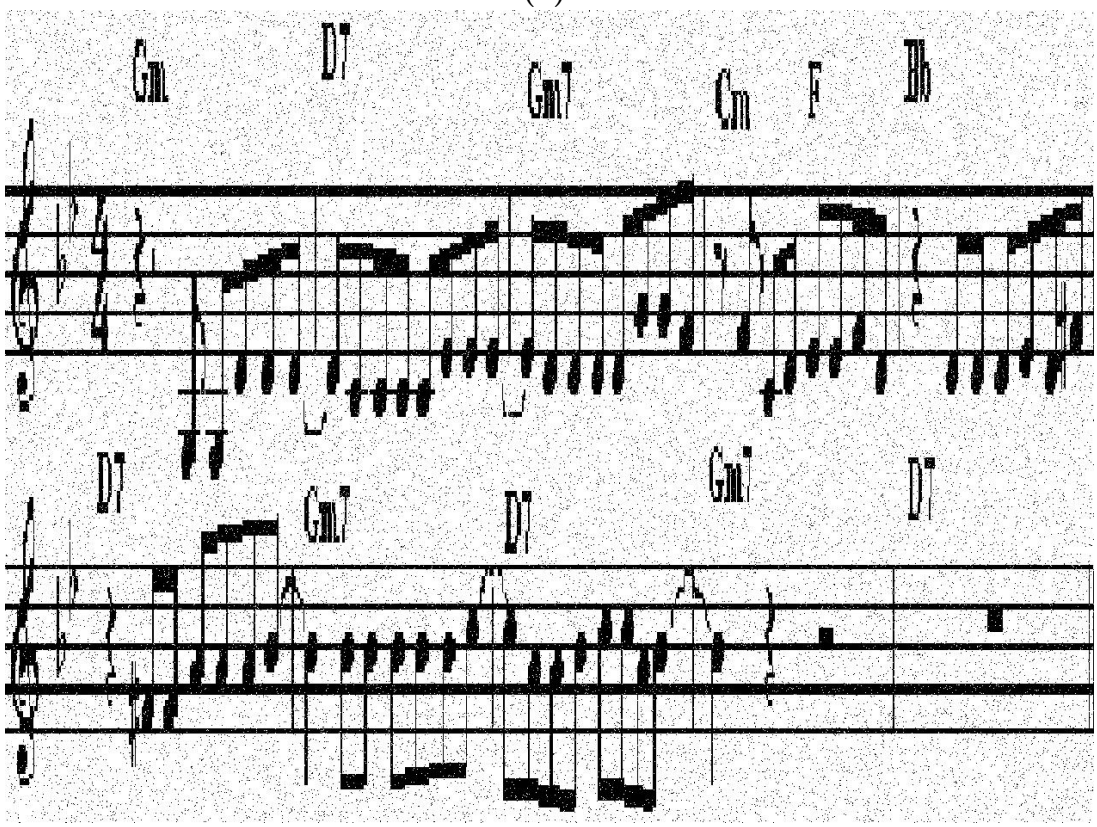


Figura 6.12: Imagens e exemplos com 10% de ruído sal e pimenta.



(a)

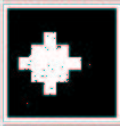
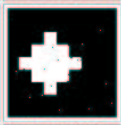


(b)

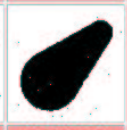
Figura 6.13: (a) Imagem de partitura; (b) exemplo com 3% de ruído sal e pimenta.

Primeiramente, analisamos o comportamento do erro médio absoluto (MAE - Mean Absolute Error) para as imagens da figura 6.12. O MAE nada mais é do que a razão da contagem dos pixels classificados incorretamente com relação ao número total de pixels da imagem. Selecionamos imagens simples com ruído adicionado para poder controlar todos os parâmetros com o objetivo de analisar o MAE como uma função de duas variáveis: (1) o tamanho do conjunto de treinamento usado para selecionar a melhor janela para o W-operador e (2) o tamanho do conjunto de treinamento usado para construir o W-operador. Utilizamos quatro tamanhos crescentes de amostras: 1/4 de uma imagem, 1/2 de uma imagem, 1 imagem e 3 imagens. Nos casos onde se retira menos de uma imagem como amostra, os pixels são obtidos de maneira aleatória (distribuição uniforme). Cada experimento consistiu em selecionar uma janela e treinar o W-operador com conjuntos de treinamento de tamanho crescente para depois aplicá-los a 10 imagens ruidosas geradas pelo mesmo modelo de ruído. A figura 6.14 resume os resultados de MAE nas imagens 1, 2 e 3 da figura 6.12. MAE mínimo se refere ao menor erro obtido entre as 10 execuções, enquanto o MAE médio indica a média dos erros dentre as 10 execuções. É importante notar que nos três casos o uso de maiores conjuntos de treinamento, tanto para selecionar a janela como para construir o W-operador, melhora a performance do filtro, como esperado.

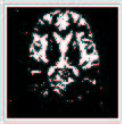
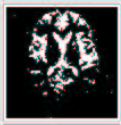
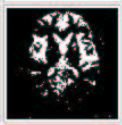
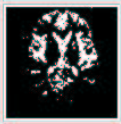
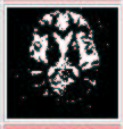
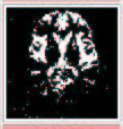
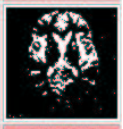
É importante analisar as janelas selecionadas para esses experimentos, que são apresentados na figura 6.15. Cada imagem nessa figura é associada a uma série de 10 execuções de seleção de janela para um respectivo tamanho de conjunto de treinamento. Uma matriz acumuladora foi criada, onde cada célula corresponde a uma variável da janela. Para cada execução, as variáveis selecionadas foram incrementadas na matriz acumuladora (procedimento análogo ao esquema de votação da transformada de Hough [21]). As imagens da figura 6.15 mostram as matriz acumuladoras correspondentes, sendo que os níveis de cinza codificam o número de votos que cada célula recebeu (os níveis de cinza mais escuros correspondem às celulas mais votadas). Como pode ser observado, as células bem votadas possuem um formato específico que é a melhor janela para construir um W-operador para o tipo considerado de imagem e ruído. Este efeito é usado para definir uma janela mais adaptada para construir o W-operador da seguinte forma: no i-ésimo experimento, o conjunto de d_i variáveis, correspondentes ao mínimo da "curva em U" da entropia condicional média, é selecionada. O número final de variáveis é a média arredondada para baixo de todos os d_i, denotado como $\bar{d}$, isto é, as $\bar{d}$ variáveis mais votadas são selecionadas.

	seleção 1/4	seleção 1/2	seleção 1	seleção 3
construção 1/4				
MAE mínimo	0.0056633	0.0080612	0.0057653	0.0099490
MAE médio	0.0070408	0.0087908	0.0071735	0.0118370
construção 1/2				
MAE mínimo	0.0046429	0.0048980	0.0036735	0.0050000
MAE médio	0.0055714	0.0058010	0.0052602	0.0067041
construção 1				
MAE mínimo	0.0036735	0.0031122	0.0029082	0.0040306
MAE médio	0.0050102	0.0038776	0.0036837	0.0045561
construção 3				
MAE mínimo	0.0038776	0.0024490	0.0027551	0.0020918
MAE médio	0.0047092	0.0035969	0.0035918	0.0030204

(1)

	seleção 1/4	seleção 1/2	seleção 1	seleção 3
construção 1/4				
MAE mínimo	0.0049288	0.0062516	0.0077318	0.0107390
MAE médio	0.0053099	0.0068641	0.0084562	0.0119360
construção 1/2				
MAE mínimo	0.0049918	0.0045037	0.0051020	0.0076688
MAE médio	0.0055130	0.0050942	0.0058469	0.0084546
construção 1				
MAE mínimo	0.0042989	0.0031809	0.0036691	0.0036218
MAE médio	0.0048123	0.0036958	0.0041163	0.0041887
construção 3				
MAE mínimo	0.0039210	0.0029447	0.0026612	0.0027242
MAE médio	0.0045257	0.0033557	0.0030518	0.0030408

(2)

	seleção 1/4	seleção 1/2	seleção 1	seleção 3
construção 1/4				
MAE mínimo	0.014928	0.017873	0.016755	0.021841
MAE médio	0.015568	0.018556	0.018345	0.022532
construção 1/2				
MAE mínimo	0.012598	0.015338	0.014755	0.018345
MAE médio	0.013801	0.016056	0.015360	0.018900
construção 1				
MAE mínimo	0.012125	0.012125	0.012708	0.013306
MAE médio	0.012725	0.013021	0.013103	0.014454
construção 3				
MAE mínimo	0.011621	0.011133	0.010944	0.011149
MAE médio	0.012172	0.011587	0.011552	0.011971

(3)

Figura 6.14: Resumo dos resultados da técnica de entropia condicional média sobre as imagens 1, 2 e 3 da figura 6.12.

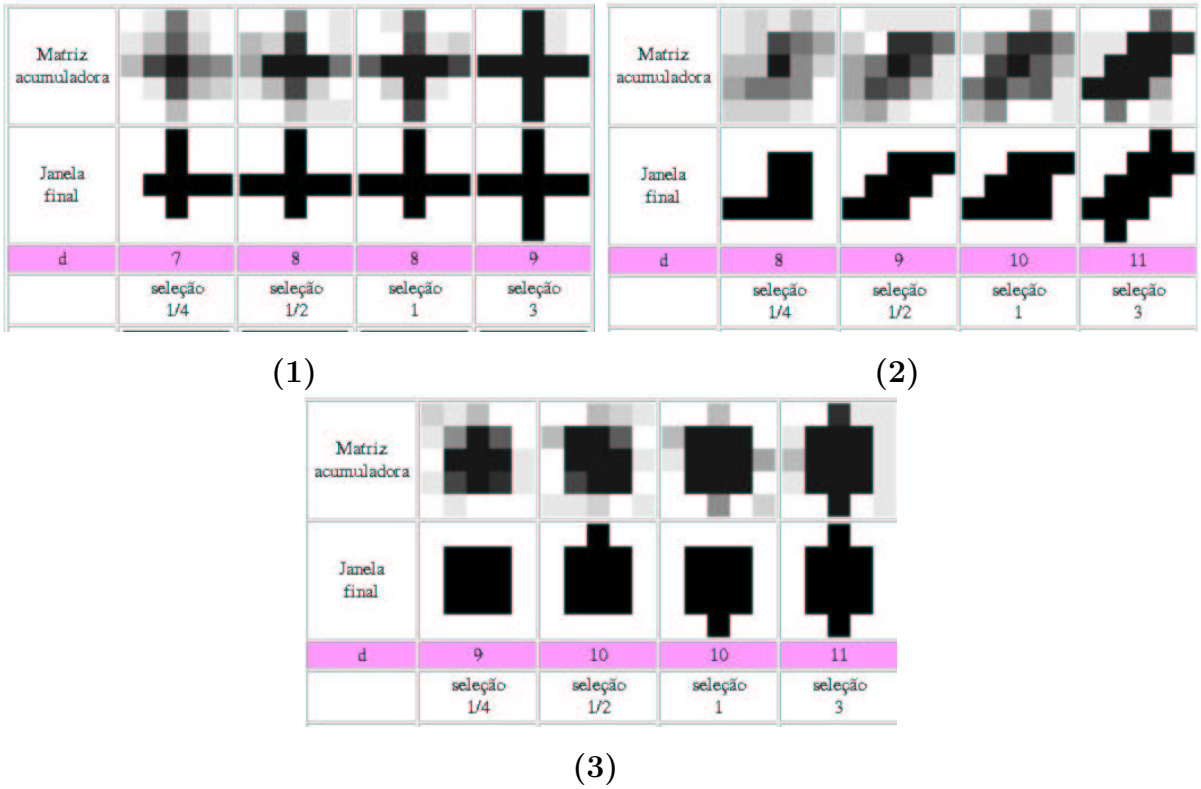


Figura 6.15: Matrizes acumuladoras e janelas W-operadoras obtidas para as imagens 1, 2 e 3 da figura 6.12 após 10 execuções.

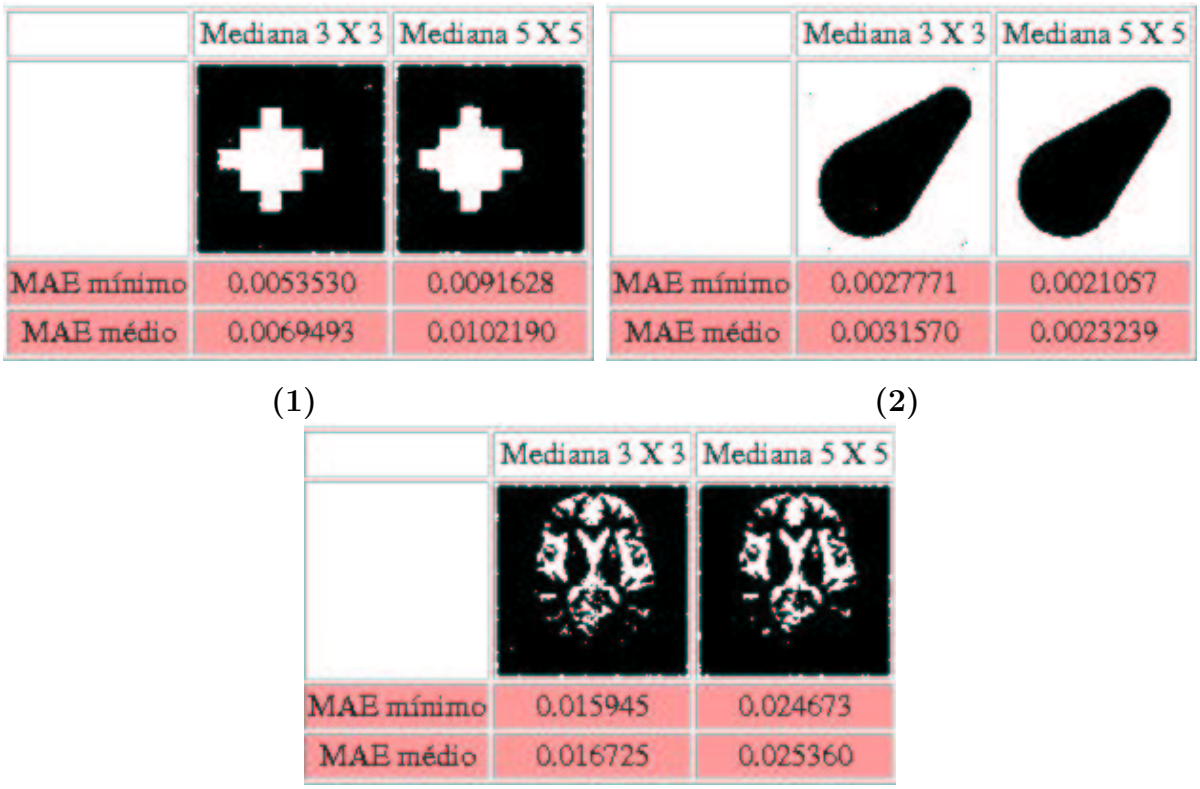


Figura 6.16: Resultados obtidos da aplicação do filtro da mediana 3×3 e 5×5 para as imagens 1, 2 e 3 da figura 6.12 após 10 execuções.

O ruído sal e pimenta é usualmente tratado em processamento de imagens pelo filtro da mediana. Comparamos os resultados dessa técnica tradicional com a técnica proposta. Aplicamos o filtro da mediana utilizando as janelas 3×3 e 5×5 sobre 10 imagens ruidosas para cada uma das imagens da figura 6.12 e obtivemos os respectivos MAE mínimo e o médio dentre as 10 execuções. Compare os resultados das tabelas da figura 6.16 com aqueles das tabelas da figura 6.14. Exceto para a imagem 2, utilizando pelo menos 1 imagem para selecionar os pixels da janela e pelo menos 1 imagem para construir o W-operador, a performance da técnica proposta foi superior às performances obtidas com os filtros da mediana.

Uma propriedade fundamental da técnica que utiliza entropia condicional média é que os filtros W-operadores gerados por ela têm a capacidade de melhorar significativamente a imagem resultado. Denominamos o processo de utilizar a imagem resultado como uma

nova imagem ruidosa para ser submetida a um mesmo filtro de *retroalimentação*. Mais precisamente, sendo o filtro W-operador uma função $W : I \to I$, onde I representa o espaço matricial das imagens, a retroalimentação é simplesmente a função $W(W(I))$.

Repetimos então os mesmos experimentos aplicando a retroalimentação tanto para a nossa técnica quanto para a técnica de mediana sendo que seus respectivos resultados podem ser vistos nas figuras 6.17 e 6.18. Note que a aplicação da retroalimentação com a técnica da mediana não garante melhora nos resultados. Já com a nossa técnica, a retroalimentação melhorou bastante a performance para as imagens 1 e 2 e melhorou um pouco (entre 5% e 10%) os resultados para a imagem 3.

Um último experimento foi realizado utilizando uma imagem que apresenta uma quantidade significativa de formas bem finas (largura de 1 ou 2 pixels). A figura 6.13 apresenta a imagem de uma partitura musical que possui essa característica e uma versão com 3% de ruído sal e pimenta. Utilizamos apenas 1 imagem para selecionar a janela e 1 imagem para construir o W-operador. O restante do processo é análogo ao que foi feito para as outras três imagens. A tabela 6.2 apresenta os valores MAE obtidos da aplicação do método de entropia condicional e da técnica da mediana com e sem retroalimentação.

	Sem retroalimentação	Com retroalimentação	
MAE mínimo	0.00054819	0.00047746	ECM
MAE médio	0.00060099	0.00050592	
MAE mínimo	0.00528850	0.00530190	Mediana 3×3
MAE médio	0.00533200	0.00532910	
MAE mínimo	0.01528500	0.01519800	Mediana 5×5
MAE médio	0.01528500	0.01527800	

Tabela 6.2: Tabela dos erros MAE para os resultados de 10 execuções das técnicas de entropia condicional média e mediana, com e sem retroalimentação, aplicadas na imagem de partitura da figura 6.13.

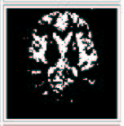
A tabela acima mostra que os resultados obtidos com o W-operador construído por entropia condicional foram cerca de 8 a 10 vezes melhor do que com o filtro da mediana 3×3 e cerca de 30 vezes melhor do que com o filtro da mediana 5×5. Essa diferença de performance é bastante perceptível na figura 6.19, onde são confrontadas a melhor

	seleção 1/4	seleção 1/2	seleção 1	seleção 3
construção 1/4				
MAE mínimo	0.0026020	0.0037245	0.0043367	0.0088776
MAE médio	0.0040357	0.0049235	0.0052551	0.010383
construção 1/2				
MAE mínimo	0.0019388	0.0020408	0.0014796	0.0026531
MAE médio	0.0025051	0.0030918	0.0022143	0.0044031
construção 1				
MAE mínimo	0.0012755	0.0011224	0.0013265	0.0010714
MAE médio	0.0023265	0.0019235	0.0018673	0.0017908
construção 3				
MAE mínimo	0.0015306	0.0005102	0.0010204	0.0004082
MAE médio	0.0023214	0.0011633	0.0020255	0.0009031

(1)

	seleção 1/4	seleção 1/2	seleção 1	seleção 3
construção 1/4				
MAE mínimo	0.0025983	0.0021101	0.0028817	0.0059366
MAE médio	0.0028896	0.0024266	0.0033447	0.0069964
construção 1/2				
MAE mínimo	0.0023778	0.0021573	0.0024408	0.0035903
MAE médio	0.0028140	0.0023085	0.0027526	0.0039872
construção 1				
MAE mínimo	0.0022361	0.0017164	0.0019054	0.0016062
MAE médio	0.0026203	0.0019085	0.0022046	0.0019936
construção 3				
MAE mínimo	0.0022361	0.0015905	0.0014645	0.0013857
MAE médio	0.002455	0.0018062	0.0017400	0.0016519

(2)

	seleção 1/4	seleção 1/2	seleção 1	seleção 3
construção 1/4				
MAE mínimo	0.013338	0.016723	0.016613	0.020472
MAE médio	0.013927	0.017175	0.017948	0.021421
construção 1/2				
MAE mínimo	0.011196	0.013102	0.013479	0.016802
MAE médio	0.012015	0.014018	0.013909	0.017750
construção 1				
MAE mínimo	0.010818	0.010393	0.011212	0.011527
MAE médio	0.011423	0.011724	0.012006	0.012902
construção 3				
MAE mínimo	0.010519	0.010598	0.010047	0.010850
MAE médio	0.010873	0.010965	0.010861	0.011588

(3)

Figura 6.17: Resumo dos resultados da técnica de entropia condicional média com retroalimentação sobre as imagens 1, 2 e 3 da figura 6.12.

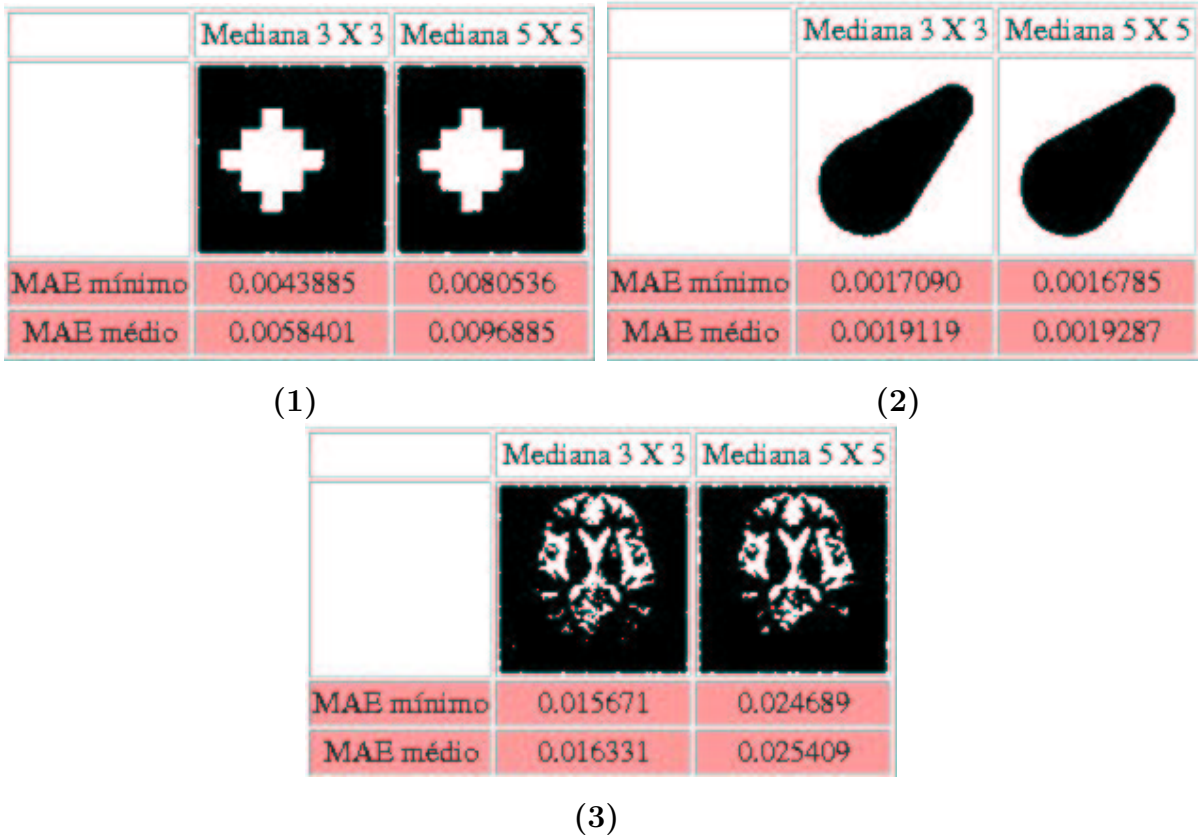


Figura 6.18: Resultados obtidos da aplicação do filtro da mediana 3×3 e 5×5 com retroalimentação para as imagens 1, 2 e 3 da figura 6.12 após 10 execuções.

imagem (MAE mínimo) obtida do W-operador por entropia condicional média por retroalimentação com a melhor imagem obtida do filtro de mediana utilizando janela 3×3 por retroalimentação. É importante observar que o filtro da mediana afeta severamente os padrões finos na imagem, muito mais do que o W-operador por entropia condicional. A matriz acumuladora e a janela obtidas estão representadas na figura 6.20

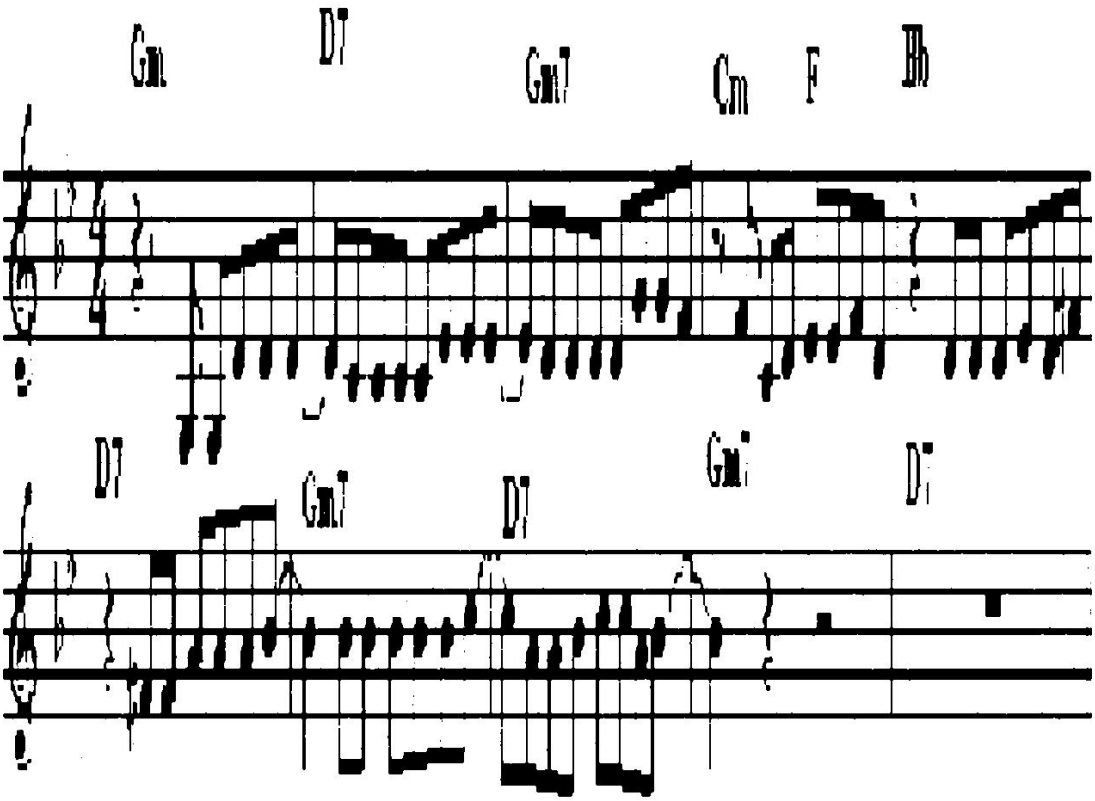
Reconhecimento de texturas

Aplicamos o nosso método proposto de construção de W-operadores também no contexto de reconhecimento de texturas. Considere a imagem da figura 6.21 representando uma concatenação de 4 texturas com 4 níveis de cinza.

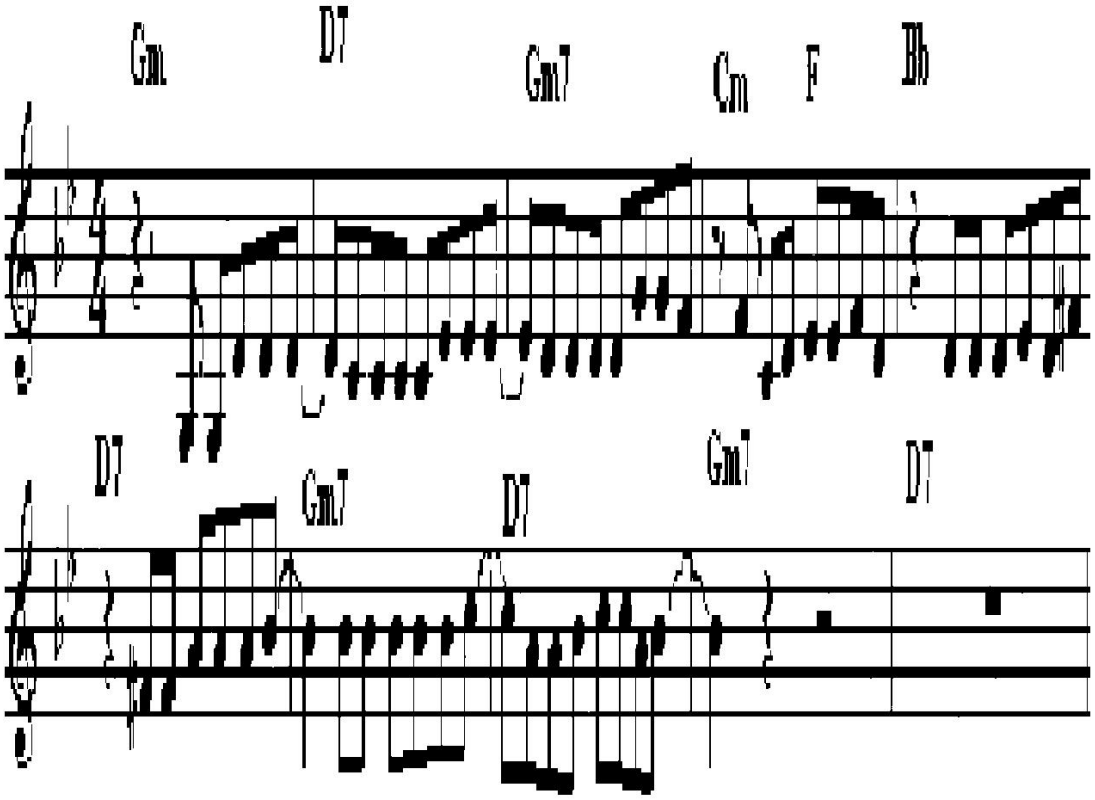
Utilizamos 1/4 de cada uma das regiões das texturas como conjunto de treinamento para obtenção da janela do W-operador (os pixels são obtidos aleatóriamente com distribuição uniforme). As classes possíveis de textura são: 0 (superior esquerda), 1 (superior direita), 2 (inferior esquerda) e 3 (inferior direita). A atribuição de classes e a legenda de classificação é ilustrada na figura 6.22. Note que, na legenda, existe uma classificação chamada de "indefinida". Este tipo de classificação ocorre quando existir instâncias $\mathbf{x}_{\mathcal{Z}}$ de $\mathbf{X}_{\mathcal{Z}}$ que não foram cobertas pelo conjunto de treinamento.

A figura 6.23 mostra a matriz acumuladora e a janela do W-operador obtida a partir de 10 execuções utilizando 1/4 das regiões correspondentes a cada textura como conjunto de treinamento.

Após a obtenção da janela, a construção do W-operador baseada nessa janela utilizou 1/4 de cada uma das regiões de textura. O resultado de aplicar esse W-operador para reconhecer as texturas da imagem da figura 6.21 pode ser observada na figura 6.24.

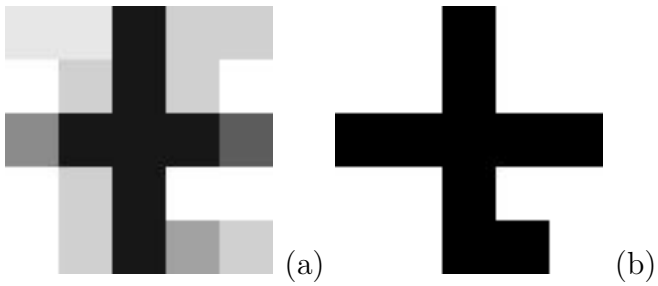


(a)



(b)

Figura 6.19: Resultados obtidos por retroalimentação na imagem de partitura da figura 6.13. (a) Mediana 3×3; (b) W-operador.



(a) (b)

Figura 6.20: (a) Matriz acumuladora; (b) Janela.

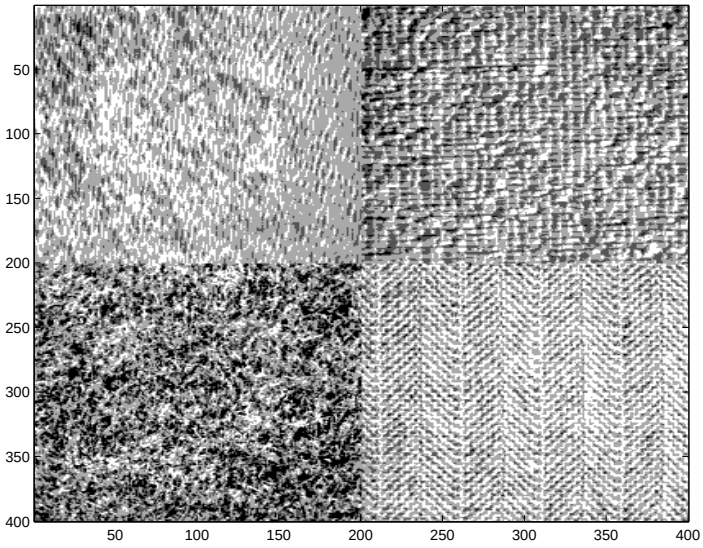
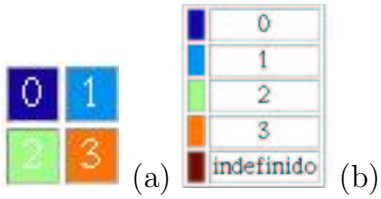


Figura 6.21: Concatenação de 4 texturas.

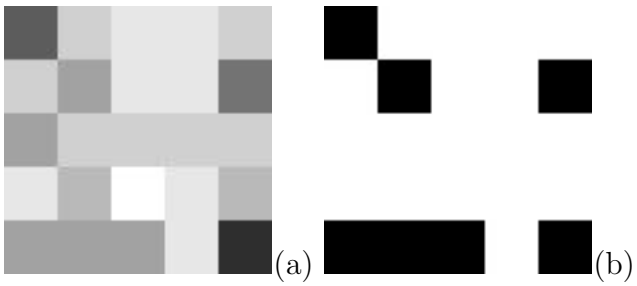


(a) (b)

Figura 6.22: (a) Atribuição das classes; (b) Legenda de classificação.

Discussão

Cada uma das regiões compreendidas pelas texturas apresentam preponderância de uma única classificação. Então, para atribuir uma textura a uma dada região basta



(a) (b)

Figura 6.23: (a) Matriz acumuladora; (b) Janela.

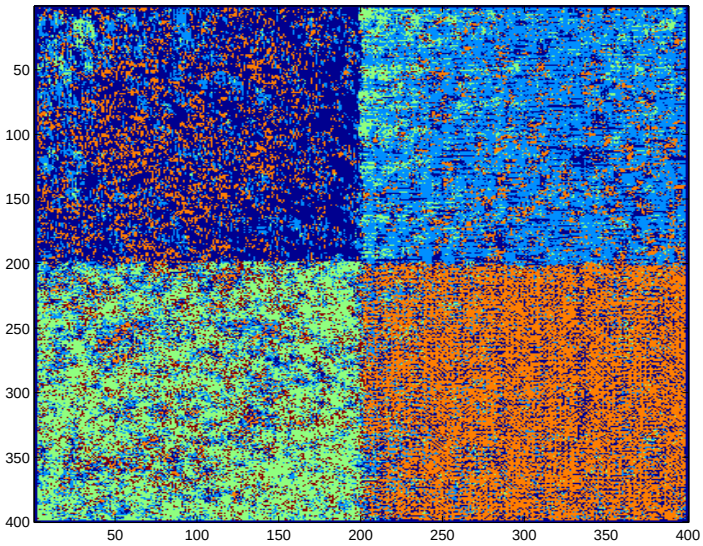


Figura 6.24: Resultado do reconhecimento das texturas da imagem original.

verificar o rótulo mais freqüente naquela região. Quanto maior a diferença entre o rótulo mais freqüente e os demais rótulos, mais confiança teremos sobre a textura de uma região. Os histogramas de freqüência da figura 6.25 mostram que, para as quatro regiões da figura 6.21, existe um rótulo que aparece em, no mínimo, 50% do número total de rotulações do W-operador. Além disso, o número de ocorrências do rótulo mais freqüente foi no mínimo duas vezes e meia maior que a freqüência do segundo rótulo mais freqüente para as quatro regiões. O número de rotulos indefinidos (valor 4 na legenda de classificação) não superou 15% do número total de rotulações em nenhuma das regiões.

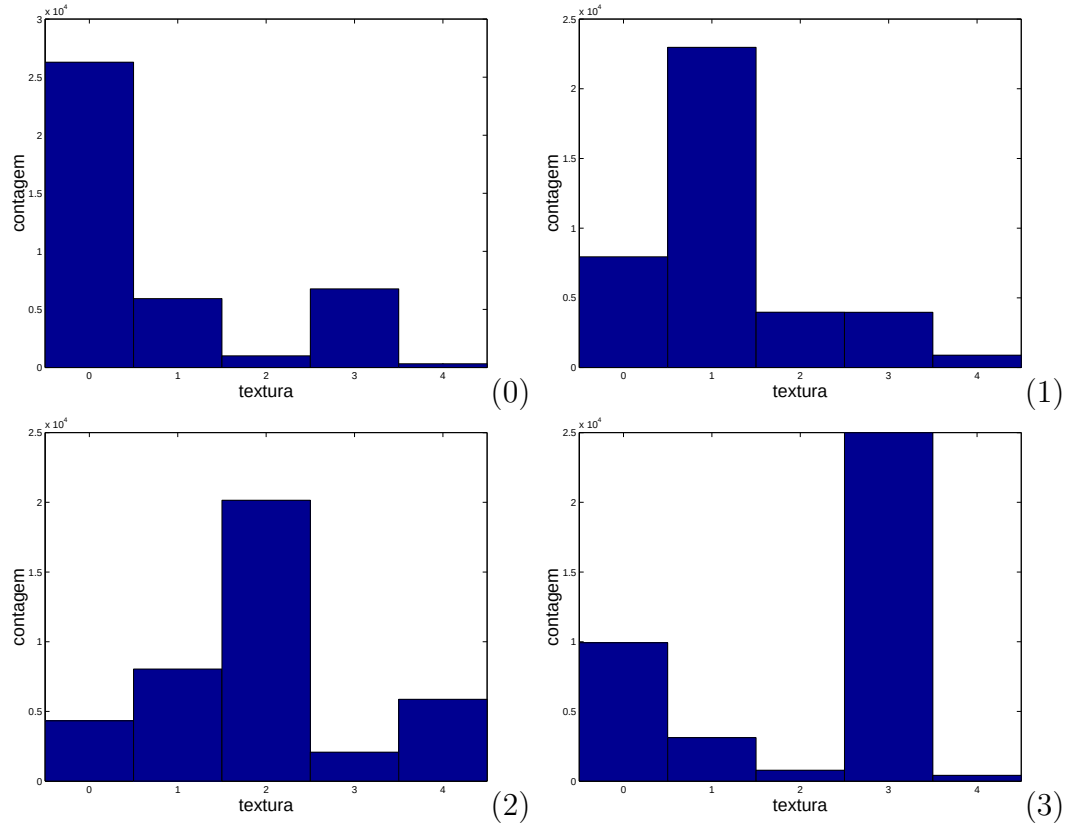


Figura 6.25: Histograma das freqüências de cada uma das rotulações realizadas pelo W-operador nas 4 regiões da figura 6.21.

6.2 Seleção de conjuntos de genes fortes através de MSV

6.2.1 Introdução

Os resultados descritos nesta seção foram obtidos pela aplicação de uma técnica elaborada pelo prof. Dr. Paulo J. S. Silva do IME-USP [64], tendo sido usada para análise de dados de *microarray* e, atualmente, esta técnica vem sendo utilizada para análise dos dados de SAGE [71]. Aplicamos o método de entropia condicional para validar esses resultados (ver seção 6.1.2).

As técnicas de *microarray* e SAGE lidam com milhares de expressões de genes ao mesmo tempo. Dentre esse enorme conjunto de genes, o interesse está em selecionar um

subconjunto pequeno de genes relativo ao conjunto inicial que melhor faz a distinção entre dois estados biológicos distintos (por exemplo: tecido normal × tecido com tumor) através de amostras de todas as expressões gênicas de cada tecido.

6.2.2 Genes fortes

Para selecionar genes diferencialmente expressos, foi utilizado o conceito de conjuntos de *genes fortes* [47]. Um conjunto de genes fortes é um pequeno grupo de genes que pode resistir a altas margens de erro na medida de expressão gênica.

Para procurar genes fortes, cada amostra deve ser considerada como o centro de uma distribuição de probabilidade. Um parâmetro de variância controla o efeito de dispersão. A separação dos dados em duas classes distintas é feita através de hiperplanos (fig. 6.26). Uma hipótese que geralmente é valida em dados de *microarray* e SAGE é que deve ser fácil separar os dados linearmente utilizando poucas amostras.

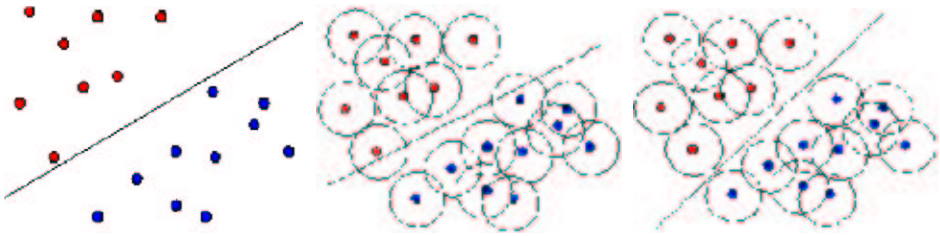


Figura 6.26: Hiperplanos separadores. A ilustração da direita mostra o hiperplano que separa os dados com a menor margem de erro possível.

Primeira estratégia para seleção de genes fortes

Procura-se inicialmente por planos que resistam ao máximo a erros nas medidas. Este problema é modelado e resolvido com o uso de programação linear e algoritmos de seleção de características através de máquinas de suporte vetorial (MSV). O problema é que o número de genes selecionados por esses planos é alto (15 a 25).

Segunda estratégia para seleção de genes fortes

Emprega-se a estratégia anterior para realizar uma pré-seleção de candidatos (entre 100 e 200). Os candidatos são então agrupados em grupos menores (3 ou 5 genes) de todas as combinações possíveis. Cada grupo é usado para classificar os dados e os melhores grupos são selecionados. Os genes mais importantes são aqueles que forem mais freqüentes nos melhores grupos. Os critérios de ordenação que podem ser usados são: freqüência, posição, índice de credibilidade ou uma combinação dos três.

A seção seguinte descreve uma das contribuições deste trabalho: um sistema de identificação e seleção de genes fortes que utiliza a implementação desta técnica como núcleo do sistema. Outra das contribuições deste trabalho, também discutida na próxima seção, foi a introdução da noção de índice de credibilidade como mais um critério de análise dos subconjuntos de genes devolvidos por esta técnica.

6.2.3 Sistema de identificação e seleção de genes fortes

Esse sistema vem sendo utilizado para análise de dados de SAGE provenientes de uma colaboração com a pesquisadora Helena Brentani do Ludwig Institute for Cancer Research. Tal colaboração tem por objetivo encontrar genes que são responsáveis por distinguir dois estados biológicos (por exemplo: tumor canceroso $\times$ normal ou dois tipos específicos de câncer).

O objetivo desse sistema *pipeline* é o de integrar todos os processos envolvidos na identificação e seleção de genes fortes em um único programa principal que cuida de receber e formatar os dados de entrada, executar os processos na ordem correta e coletar os dados de saída de cada um destes processos, gerando um relatório final dos resultados.

Geralmente, as tabelas de entrada possuem pouco mais de 200 bibliotecas de SAGE, cada uma delas correspondendo a uma amostra de tecido de algum orgão do corpo humano (por exemplo: cérebro, mama, estômago, pâncreas, além de outros). Algumas dessas bibliotecas são de controle normal e outras são de algum tipo de tumor. Cada uma delas contém da ordem de 20000 expressões gênicas ($n = 20000$). A parte responsável pelo pré-processamento do sistema cuida de ler uma tabela de entrada e dispor as informações sobre a forma de matriz, na qual cada biblioteca esteja disposta por linha e cada coluna

corresponda a um gene com suas expressões sobre as bibliotecas. Portanto, cada célula (i, j) da matriz de expressões corresponde à expressão do gene X_j na biblioteca Bib_i.

Com a matriz em mãos, os pesquisadores estão interessados em respostas para questões específicas, como: "Quais genes são responsáveis pelo tumor glioblastoma cerebral?". Para este exemplo, seleciona-se apenas as bibliotecas de glioblastoma e as bibliotecas de tecido cerebral normais, formando no total t amostras. Em seguida, deve-se rotular essas amostras de tal forma que as do primeiro grupo recebam o rótulo +1 (presença do tumor) e as amostras do segundo grupo recebam o rótulo -1 (ausência de tumor). Assim, obtém-se os elementos necessários para responder a essa questão: um conjunto de treinamento composto por uma matriz de expressões $t \times n$ e um vetor de t rótulos. Portanto, uma questão biológica pode ser representada por uma seleção das bibliotecas que irão compor dois grupos distintos.

Cada valor de expressão é um número inteiro positivo que indica o número de vezes que uma determinada *tag* (gene) foi observada em uma determinada biblioteca[2]. Cada biblioteca tem um número total de observações de *tags*. Portanto, o próximo passo de pré-processamento é dividir cada expressão pelo número total de observações das *tags* da biblioteca correspondente, ou seja, para $1 \leq i \leq t$ e para $1 \leq j \leq n$:

$$x_{ij} \leftarrow x_{ij}/nBib_i$$

em que x_{ij} é a expressão do gene X_j na biblioteca Bib_i e $nBib_i$ é o número total de tags observadas na biblioteca i. Portanto, a matriz de contagens passa a ser uma matriz de freqüências.

Em seguida, aplica-se um passo de normalização sobre a matriz através da *transformação normal* [21] (vide equação 5.10 na seção 5.2) para que todos os genes tenham a mesma amplitude de valores. Após este passo, se x_{ij} tiver valor negativo, isso significa que o gene X_j tem um valor de expressão na biblioteca i que está abaixo da média das suas próprias expressões entre todas as bibliotecas. Caso contrário (valor positivo) X_j se expressou acima da média.

Esta matriz e o vetor de rótulos são fornecidos ao núcleo do sistema, que devolve uma tabela com os mil melhores subconjuntos de genes ordenados pelo erro. Desta tabela

[2]Para uma breve discussão sobre a tecnologia de SAGE, ver seção 3.2.2

obtém-se facilmente o número de ocorrências de cada um dos genes nesses subconjuntos, servindo como um critério adicional para seleção final dos subconjuntos. Esta seleção final (normalmente menos de 10 subconjuntos) não é realizada de forma automática, mas de forma parcialmente subjetiva com base no erro e na freqüência dos genes. Ou seja, a seleção final fica a cargo dos pesquisadores.

Em geral, o número de bibliotecas envolvidas em uma determinada questão biológica gira em torno de 10 a 30. Para este número relativamente pequeno de amostras, normalmente trabalha-se com subconjuntos de 3 genes (trincas), pois o erro de estimação que se comete ao levar em conta subconjuntos de 4 ou mais genes passa a ser muito alto para tal quantidade limitada de amostras. Levando trincas em consideração, é relevante notar que o número de subconjuntos devolvidos (1000) pela técnica é ínfimo se comparado ao universo de todas as trincas possíveis (combinação da ordem de 20000, 3 a 3 $\sim$ 1,3 trilhão de trincas).

Para facilitar o trabalho de seleção final, o sistema gera uma tabela em formato html na qual cada linha contém informações sobre um determinado subconjunto de genes, contendo o nome de cada gene, o número total de bibliotecas, o número de bibliotecas do primeiro grupo, o erro, a distância entre os grupos, a freqüência com que cada gene aparece na tabela e um índice de credibilidade $\mathcal{C}$ que se obteve desse subconjunto (figura 6.27). Caso os subconjuntos considerados sejam trincas, cada linha terá também apontadores para dois gráficos tridimensionais que mostram os valores das expressões da trinca considerada em cada biblioteca representados por um ponto no espaço formado pelos 3 genes nos eixos X, Y e Z. Esses valores podem ser obtidos tanto da matriz de freqüências quanto da matriz normalizada, mas para facilitar a interpretação, o sistema adota os valores de freqüencia ($x_{ij}/nBib_i$). Cada ponto terá cor azul ou vermelha, dependendo do grupo do qual a biblioteca faz parte (-1 ou +1 respectivamente) (figura 6.28). Um desses gráficos apresenta também os intervalos de credibilidade com índice de credibilidade $\mathcal{C}$ em torno de cada ponto (figura 6.29). Segue a seguir uma breve explicação sobre a geração dos intervalos de credibilidade em torno de cada ponto e o cálculo do índice de credibilidade.

#	X	Y	Z	Total libs	Libs 1	Erro	Distância	Fx	Fy	Fz	Cred
1	AL833702	BC014549	BC003091	14	6	0.003598	1.977421	186	350	51	0.85156
2	BC014549	NM_014169	AF200478	14	6	0.006132	1.874602	350	147	28	0.625
3	BC033872	NM_004092	AK093630	14	6	0.006460	1.945676	145	32	187	0.64844
4	BC014549	NM_014169	U96136	14	6	0.006606	1.875805	350	147	68	0.73438
5	AL833702	AK093630	AK054724	14	6	0.006928	1.870132	186	187	46	0.53906
6	AL833702	BC014549	BC033872	14	6	0.006980	1.423012	186	350	145	0.85156
7	BC014549	AL133585	BC033872	14	6	0.007066	1.955371	350	21	145	0.70312
8	BC014549	BC033872	AK093630	14	6	0.007232	1.577272	350	145	187	0.70312
9	BC014549	AK074494	NM_013293	14	6	0.007585	1.710351	350	57	44	0.58594
10	BC014549	NM_014169	X65724	14	6	0.008020	1.680727	350	147	32	0.58594

Figura 6.27: Exemplo de tabela gerada mostrando 10 melhores trincas.

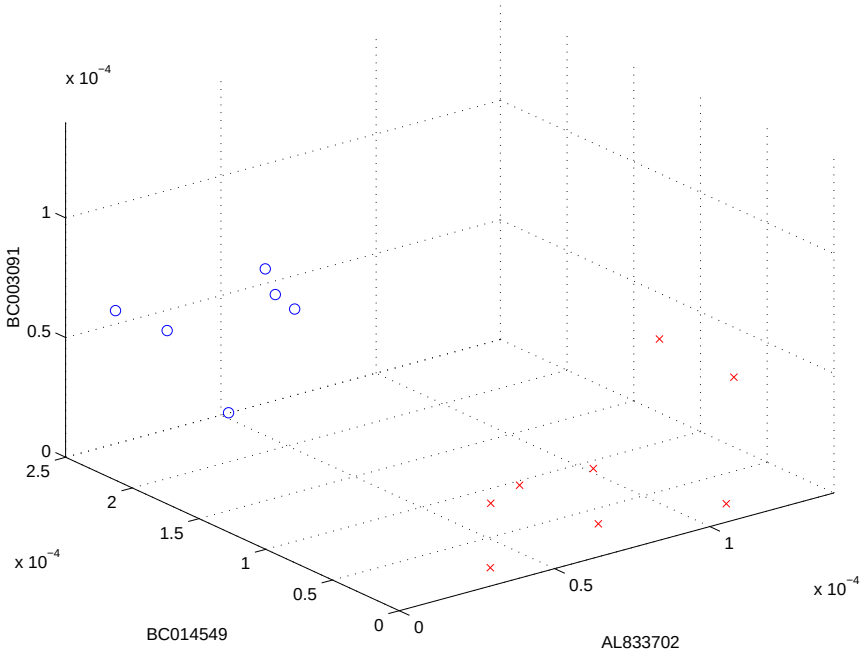


Figura 6.28: Exemplo de gráfico 3D dos valores de expressão da melhor trinca obtida para a tabela da figura 6.27.

6.2.4 Intervalo e índice de credibilidade

O conceito de índice de credibilidade para uma determinada trinca é outra das contribuições deste trabalho, realizada em conjunto com Ricardo Z. N. Vêncio, aluno de mestrado em estatística do IME-USP.

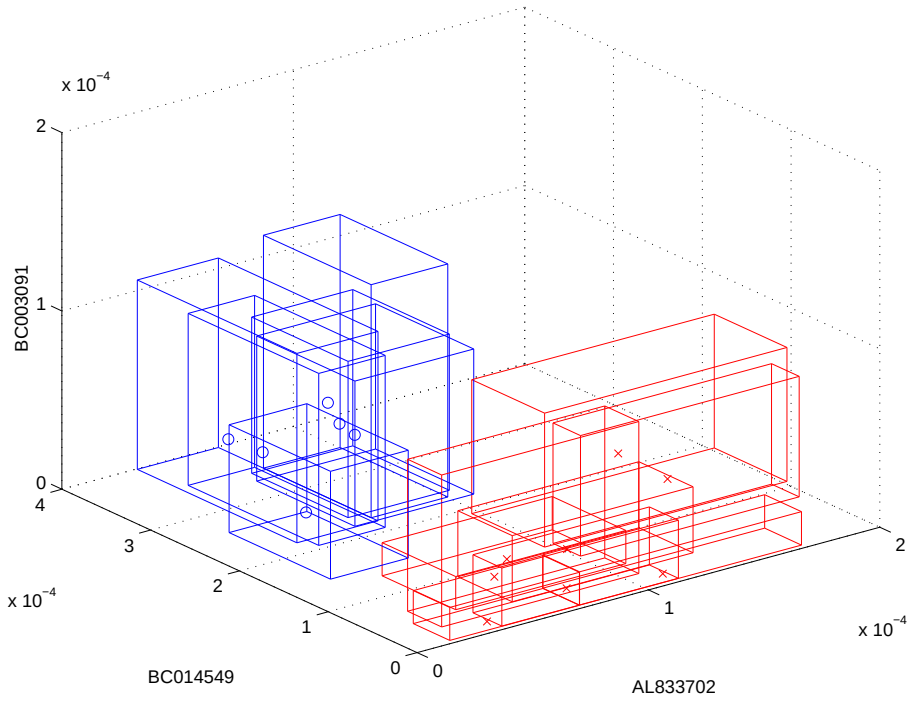


Figura 6.29: Exemplo de gráfico 3D com intervalos de credibilidade sobre a figura 6.28.

Como os dados de expressão do SAGE são obtidos por amostragem (contagem de *tags* observadas), é importante obter uma margem de credibilidade para cada uma das expressões com base no total de tags observadas em uma biblioteca. É amplamente conhecido que este problema pode ser modelado por uma função densidade de probabilidade *beta* (β) [13]. Então, dados dois números inteiros a e b sendo a contagem de ocorrências de uma *tag* em uma determinada biblioteca e o número total de *tags* observadas nessa biblioteca respectivamente, o intervalo de credibilidade em torno de a é calculado a partir da função densidade beta dada pela seguinte fórmula:

$$f(x) = \frac{x^a(1-x)^{b-a}}{\int_0^1 t^a(1-t)^{b-a}dt} \tag{6.1}$$

Escolhido um índice de credibilidade $\mathcal{C}$, tal que $0 < \mathcal{C} < 1$, os valores extremos do intervalo de credibilidade, t_1 e t_2, são obtidos através da integração da curva formada por $f(x)$ a partir de sua moda (pico, ponto de máximo da curva) (figura 6.30(a)) de tal forma que a densidade de probabilidade nos pontos t_1 e t_2 sejam iguais ($f(t_1) = f(t_2)$) e a área debaixo da curva no intervalo $[t_1, t_2]$ seja igual a $\mathcal{C}$. Nem sempre isto é possível,

pois há casos onde a fronteira de um dos lados de $f(x)$ é atingida antes da integração ter alcançado o valor $\mathcal{C}$, já que $f(x)$ não é necessariamente simétrica em torno de sua moda (figura 6.30(b)). Se isto ocorrer, deve-se integrar o restante no outro lado da moda em que a sua fronteira não foi atingida. Usando a função densidade de probabilidade *a priori* não informativa, a moda da função densidade de probabilidades ocorre em $x = a$, ou seja, a moda ocorre no ponto de contagem de ocorrências da *tag* considerada.

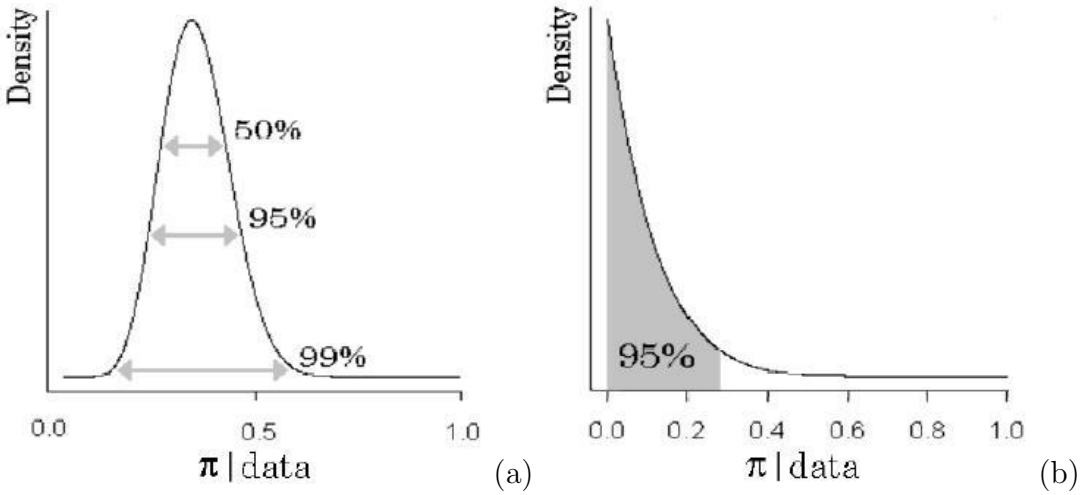


Figura 6.30: Construção dos intervalos de credibilidade. (a) 3 exemplos de intervalos de credibilidade. (b) Exemplo de assimetria de uma função densidade de probabilidades beta.

Assim, o cálculo do índice de credibilidade de uma determinada trinca de genes é feito da seguinte maneira. Fixada uma biblioteca e fixados os valores de uma dada trinca de genes sobre ela formando um ponto no espaço tridimensional, calcula-se os intervalos com o índice de credibilidade $\mathcal{C}$ dado para cada uma das expressões que o compõem. Em seguida, constrói-se um paralelepípedo envolvendo o ponto em questão. Após a projeção dos paralelepípedos sobre todos os pontos utilizando o mesmo $\mathcal{C}$, realiza-se o seguinte teste: se nenhum dos paralelepípedos do grupo 1 cruzar com nenhum paralelepípedo do grupo 2, significa que tal trinca possui índice de credibilidade maior que $\mathcal{C}$, caso contrário, possui índice de credibilidade menor que $\mathcal{C}$.

Conseqüentemente, para descobrir o índice de credibilidade aproximado de uma determinada trinca, aplica-se uma busca binária no intervalo $[0, 1]$, estipulando inicialmente $\mathcal{C} = 0.5$ e testando se um dos paralelepípedos de um grupo cruza com um de outro grupo.

Se sim, aplica o mesmo teste para $\mathcal{C} = 0.25$. Senão, aplica para $\mathcal{C} = 0.75$. Quanto maior a profundidade da busca binária, mais preciso será o índice de credibilidade obtido. O sistema adota profundidade 7, ou seja, o mesmo procedimento é aplicado 7 vezes.

Resultados

Realizamos um teste de classificação com a trinca de menor erro obtido para o cruzamento dos tumores cerebrais glioblastoma × astrocytoma graus II e III através de 6 bibliotecas de glioblastoma e 8 bibliotecas de astrocytoma (4 de grau II e 4 de grau III), ou seja, as 6 primeiras pertencentes ao primeiro grupo (azul) e as 8 últimas pertencentes ao segundo grupo (vermelho). As figuras 6.27, 6.28 e 6.29 são relativas a esse cruzamento. É importante observar que além de ser a melhor trinca pelo critério de erro, ela também tem o melhor índice de credibilidade se comparado com as outras 10 trincas (ver figura 6.27). A figura 6.31 ilustra esse fato através de um gráfico (credibilidade × erro) para as 1000 melhores trincas. O valor (credibilidade, erro) da trinca escolhida (a de menor erro) está representada por um "+" vermelho no gráfico.

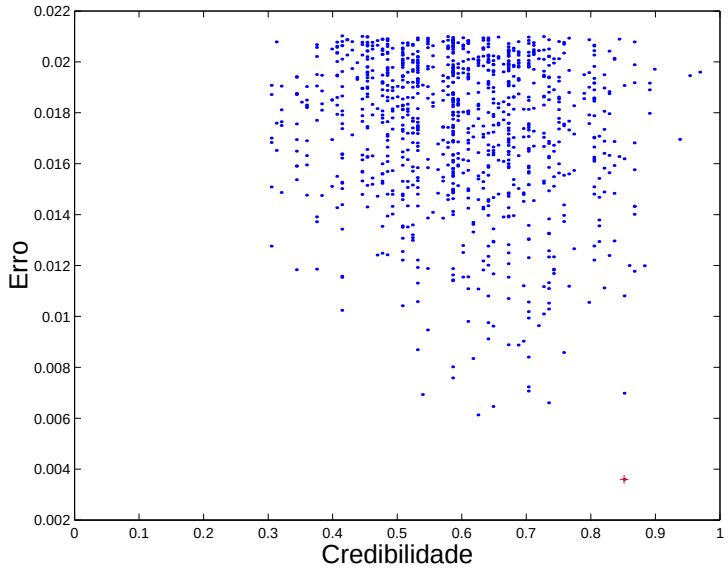


Figura 6.31: Gráfico (credibilidade × erro de cada uma das 1000 melhores trincas). O valor da credibilidade e do erro da trinca escolhida para classificação está representada com o símbolo "+".

Foram adicionados aos mesmos gráficos, 4 valores da mesma trinca em questão referentes a outras 4 bibliotecas de astrocytoma grau II que não faziam parte do cruzamento. Estes valores estão representados em verde na figura 6.32. Note que a trinca considerada classificou as 4 bibliotecas de teste corretamente porque todos os novos valores estão localizados bem próximos dos pontos vermelhos (segundo grupo), confirmando a sua capacidade de separar bem essas duas classes. Mais informações poderão ser obtidas com o artigo [8], que encontra-se em fase de preparação.

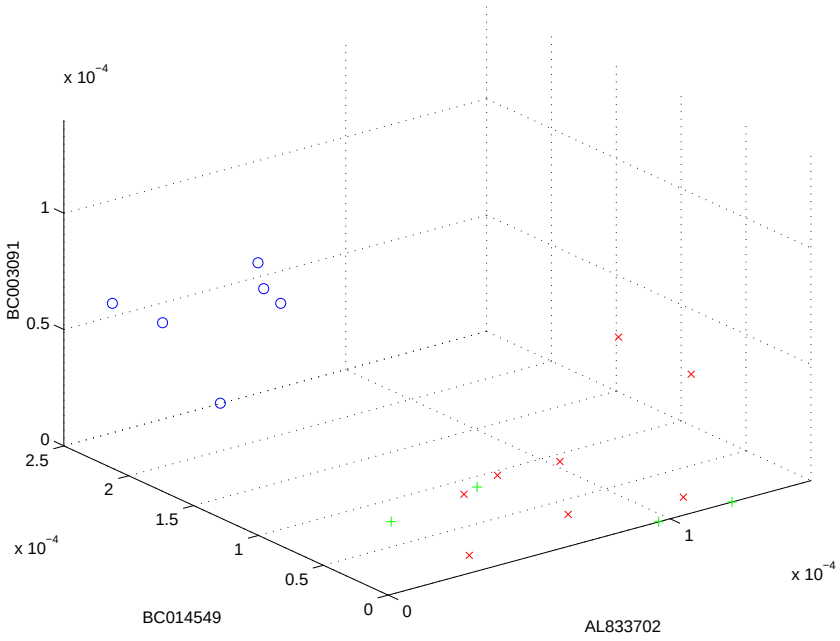


Figura 6.32: Resultado da classificação das novas bibliotecas de astrocytoma grau II (representadas com o símbolo "+").

A seção 6.1.2 mostrou a realização de dois experimentos nos quais aplicamos o método de entropia condicional para validar esses resultados.

Capítulo 7

Conclusões

Neste texto, foram apresentadas as contribuições deste mestrado, sendo a principal delas, uma função critério para seleção de características adequada para separar duas ou mais classes distintas e que não privilegie subespaços que separem as classes linearmente. Ela baseia-se nas entropias condicionais da variável classe dadas as instâncias de um subespaço de características. A proposta de realizar seleção de características através de conceitos de informação mútua e entropia não é nova [14, 35, 39, 50, 72, 77]. Porém, a alteração da fórmula da entropia condicional média para comportar a constante α como uma forma de atribuir pesos às instâncias não observadas é, sem dúvida, uma contribuição importante e fundamental para evitar os erros de estimação que se comete ao selecionar subespaços com dimensão muito grande através de conjuntos de treinamento relativamente pequenos. Sem este valor α, quaisquer subespaços de dimensão maiores teriam sempre entropias condicionais médias menores do que os subespaços menores contidos neles. Agindo assim, estaríamos realizando redução de dimensionalidade desprezando a "curva em U" característica do problema da dimensionalidade.

Um problema em aberto que ficou deste trabalho está em como estimar corretamente o valor de α. Embora o valor $\alpha = 1$ tenha sido adotado pelos nossos experimentos com resultados satisfatórios, estimar o valor de α, talvez com base no tamanho do conjunto de treinamento, pode resultar em subespaços ainda melhores.

Estudar o valor da entropia condicional média também é outro desafio para descobrir se um subespaço de características selecionado realmente é um bom preditor dos rótulos. É fato que, fixado α, o subespaço de características selecionado será o melhor preditor

dentre todos os subespaços, caso o algoritmo de busca tenha testado todas as combinações possíveis. O problema é que existem situações nas quais não existe um subespaço de características que seja um bom preditor dos rótulos. No problema de identificar redes de regulação gênica, por exemplo, resolver esta questão é fundamental pois pode ser que existam genes cujas expressões não sejam influenciadas por nenhum outro gene (nós disjuntos da rede).

As aplicações abordadas neste trabalho foram bastante diversificadas, abrangendo duas áreas da computação: bioinformática e processamento de imagens. A função critério proposta para seleção de características vem atendendo satisfatoriamente as exigências de cada área. Isto comprova a genericidade do método proposto. Segue abaixo algumas considerações sobre trabalhos em andamento e futuros.

- Estimação de α e estudo da significância do valor da entropia condicional média com base no conjunto de treinamento.
- Na análise de dados de SAGE sobre tecidos humanos mencionada anteriormente, os resultados da questão biológica discutida neste texto estão em fase de validação experimental [8]. Caso tais resultados sejam validados com sucesso, o próximo passo será a análise de outras questões biológicas já propostas através da mesma técnica de MVS combinada com a técnica de entropia condicional média.
- No caso da identificação de redes de regulação gênica em dados de *microarray* de malária, o projeto encontra-se em fase de ajuste de parâmetros e validação com dados simulados. Algumas variantes do modelo proposto estão sendo discutidas para posterior implementação e testes com dados simulados para, em seguida, aplicá-lo aos dados reais visando a produção de subsistemas de regulação mais confiáveis [7].
- No contexto de processamento de imagens, a técnica proposta para encontrar um W-operador minimal é bastante genérica, podendo ser utilizada em diversas outros aplicações, por exemplo, na análise de documentos. Um refinamento possível da técnica considerada é a construção de W-operadores no contexto de análise multi-resolução [28].
- Um projeto tido como conseqüência do critério proposto com este mestrado é referente ao desenvolvimento de um algoritmo *branch and bound* de seleção de características que explora a propriedade da "curva em U" formada pelos valores das

entropias condicionais médias em função da dimensão das características. Este projeto vem sendo desenvolvido com o objetivo inicial de obter janelas W-operadoras minimais para reconhecimento de texturas [59]. Comparação deste algoritmo com algoritmos clássicos da literatura, como o SFS ou o SFFS, estão previstos.

Referências Bibliográficas

[1] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing*, volume 4, pages 17–28, 1999.

[2] U. Alon, N. Barkal, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *Proceedings of National Academy of Sciences*, volume 96, pages 6745–6750, USA, Jun 1999.

[3] H. A. Armelin, J. Barrera, E. R. Dougherty, M. D. Gubitoso, J. E. Ferreira, N. S. T. Hirata, and E. J. Neves. A simulator for gene expression networks. In *SPIE Microarrays: Optical Technologies and Informatics*, volume 4266, pages 248–259, San Jose, January 2001.

[4] P. Baldi. On the convergence of a clustering algorithm for protein-coding regions in microbial genomes. *Bioinformatics*, 16:367–371, 2000.

[5] P. Baldi and A. D. Long. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.

[6] J. Barrera, R. M. Cesar-Jr, D. O. Dantas, D. C. Martins-Jr, and N. W. Trepode. From microarray images to biological knowledge. In R. Mondaini, editor, *Proceedings of the Second Brazilian Symposium on Mathematical and Computational Biology*, pages 209–239, 2002.

[7] J. Barrera, R. M. Cesar-Jr, D. C. Martins-Jr, E. F. Merino, R. Z. N. Vencio, F. G. Leonardi, M. M. Yamamoto, C. A. B. Pereira, and H. A. Portillo. Abstract: A new

annotation tool for malaria based on inference of probabilistic genetic networks. In *Proceedings of CAMDA*, 2004.

[8] J. Barrera, R. M. Cesar-Jr, D. C. Martins-Jr, P. J. S. Silva, R. Z. N. Vêncio, and H. P. Brentani. Strong gene sets obtained by SVM to separate glioblastoma from astrocitoma grade II and III tumors. 2004. (em preparação).

[9] J. Barrera, P. A. Martin, E. R. Dougherty, M. D. Gubitoso, N. S. T Hirata, and N. W. Trepode. Identification of input-free finite lattice dynamical systems under envelope constraints. In *International Symposium on Mathematical Morphology*, pages 337–345, Sydney, 2002.

[10] J. Barrera, R. Terada, R. Hirara-Jr, and N. S. T. Hirata. Automatic programming of morphological machines by PAC learning. *Fundamenta Informaticae*, 41:229–258, Jan 2000.

[11] M. Bittner, P. Meltzer, Y. Chen, Y. Jlang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature 406*, pages 536–45, 2000.

[12] I. Bloch. On fuzzy distances and their use in image processing under imprecision. *Pattern Recognition*, 11(32):1873–1895, 1999.

[13] H. Bolfarine and M. C. Sandoval. *Introdução à inferência estatística*. Rio de Janeiro, 2001.

[14] B. V. Bonnlander and A. S. Weigend. Selecting input variables using mutual information and nonparametric density estimation. In *Proc. of the 1994 Int. Symp. on Artificial Neural Networks*, pages 42–50, Tainan, Taiwan, 1994.

[15] Z. Bozdech, M. Llinás, B. L. Pulliam, E. D. Wong, J. Zhu, and J. L. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *Plos Biology*, 1(1), Oct 2003.

[16] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S Furey, M. Ares-Jr, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. In *Proceedings of National Academy of Sciences*, volume 97, pages 275–286, USA, Jan 2000.

[17] T. E. Campos. Técnicas de seleção de características com aplicações em reconhecimento de faces. Master's thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, Rua do Matão, 1010, May 2001.

[18] T. E. Campos, I. Bloch, and R. M. Cesar-Jr. Feature selection based on fuzzy distances between clusters: First results on simulated data. In *Lecture Notes in Computer Science*, volume 2013, pages 186+, Rio de Janeiro, Brasil, Mar 2001. Springer-Verlag Press.

[19] T. E. Campos, R. S. Feris, and R. M. Cesar-Jr. Improved face $\times$ non-face discrimination using fourier descriptors through feature selection. In *Proceedings of 13th SIBGRAPI*, pages 28–35. IEEE Computer Society Press, October 2000.

[20] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. In *Pacific Symposium on Biocomputing*, volume 4, pages 29–40, 1999.

[21] L. F. Costa and R. M. Cesar-Jr. *Shape analysis and classification: theory and practice*. CRC Press, Boca Raton, 2001.

[22] T. M. Cover and J. A. Thomas. Elements of information theory. In *Wiley Series in Telecommunications*. John Wiley & Sons, New York, NY, USA, 1991.

[23] P. D'haeseleer. *Reconstructing Gene Networks from Large Scale Gene Expression Data*. PhD thesis, University of New Mexico, 2000.

[24] P. D'haeseleer, S. Liang, and Roland Somgyi. Tutorial: Gene expression data analysis and modeling. In *Pacific Symposium on Biocomputing*, Hawaii, January 1999.

[25] E. R. Dougherty and J. Barrera. Pattern recognition theory in nonlinear signal processing. *Journal of Mathematical Imaging and Vision*, 16(3):181–197, 2002.

[26] E. R. Dougherty, J. Barrera, G. Mozelle, S. Kim, and M. Brun. Multiresolution analysis for optimal binary filters. In *Journal of Mathematical Imaging and Vision*, volume 14, pages 53–72, 2001.

[27] E. R. Dougherty, M. L. Bittner, Y. Chen, S. Kim, K. Sivakumar, J. Barrera, P. Meltzer, and J. M. Trent. Nonlinear filters in genomic control. In *IEEE-NSIPP*, pages 10–15, Turkey, 1999. Antalya.

[28] E. R. Dougherty, S. Kim, G. Mozelle, J. Barrera, and M. Brun. Multiresolution filter design. In *Non Linear Image Processing XI. Electronic Imaging '2000*, San Jose, 2000.

[29] E. R. Dougherty and R. A. Lotufo. *Hands-On Morphological Image Processing*. SPIE, June 2003.

[30] A. Drawid and M. Gerstein. A bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of Molecular Biology*, 301:1059–1075, 2000.

[31] R. O. Duda, P. E. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, NY, 2000.

[32] B. Dutilh. Analysis of data from microarray experiments, the state of the art in gene network reconstruction. *Theoretical biology and Bioinformatics*, October 1999.

[33] M. B. Eisen, P. T. Spellman, P. O Brown, and D. Botstein. Cluster analysis and display of genoma-wide expression patterns. In *Proceedings of National Academy of Sciences*, volume 95, pages 14863–14868, USA, Dec 1998.

[34] K. M. Eyster and R. Lindahl. Molecular medicine: a primer for clinicians. In *Part XII: DNA microarrays and their application to clinical medicine*, pages 57–61. S D J Med, 2001.

[35] F. Fleuret. Binary feature selection with conditional mutual information. *Rapport de Recherche*, (4941), October 2003.

[36] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[37] D. Ghosh. Mixture modeling of gene expression data from microarray experiments. *Bioinformatics*, 18(2):275–286, 2001.

[38] D. K. Gifford. Blazing pathways through genetic mountains. *Science*, 293:2049–2051, 2001.

[39] M. A. Hall and L. A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *FLAIRS*, 1999.

[40] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Pacific Symposium on Biocomputing*, pages 422–433, 2001.

[41] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, Louis Staudt, W. C. Chan, D. Botstein, and P. Brown. "gene shaving"as a method for identifying distinct sets of genes with similar expression pattern. *Genome Biology*, 1:1–21, 2000.

[42] R. Hirata-Jr, J. Barrera, R. F. Hashimoto, D. O. Dantas, and G. P. Esteves. Segmentation of microarray images by mathematical morphology. *Real Time Imaging*, 8:491–505, 2002.

[43] T. E. Ideker, V. Thorsson, and R. M. Karp. Discovery of regulatory interactions through perturbation: Inference and experimental design. In *Pacific Symposium on Biocomputing*, volume 5, pages 302–313, 2000.

[44] A. K. Jain, P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000.

[45] A. K. Jain and D. Zongker. Feature-selection: Evaluation, application, and small sample performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):152–157, February 1997.

[46] S. Kim, E. R. Dougherty, J. Barrera, Y. Chen, M. Bittner, and J. M. Trent. Finding robust linear expression-based classifiers. In *SPIE Microarrays: Optical Technologies and Informatics*, San Jose, 2001.

[47] S. Kim, E. R. Dougherty, J. Barrera, Y. Chen, M. Bittner, and J. M. Trent. Strong feature set from small samples. *Journal of Computational Biology*, 2002. Accepted.

[48] S. Knudsen. *A Biologist's Guide to Analysis of DNA Microarray Data*. John Wiley & Sons, 2002.

[49] S. Kullback. *Information Theory and Statistics*. Dover, 1968.

[50] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 212–217, San Mateo, California, 1992. Morgan Kaufmann.

[51] M. Li, B. Wang, Z. Momeni, and F. Valafar. Pattern recognition techniques in microarray data analysis. In F. Valafar, editor, *Annals of New York Academy of Sciences*, volume 980, pages 41–64, December 2002.

[52] S. Liang, S. Furhman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Simposium on Biocomputing*, volume 3, pages 18–29, 1998.

[53] D. C. Martins-Jr, R. M. Cesar-Jr, and J. Barrera. W-operator window design by maximization of training data information. In *Proceedings of 17th SIBGRAPI*. IEEE Computer Society Press, 2004. (in press).

[54] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. In *IEEE Trans. Computers*, volume 26, pages 917–922, 1977.

[55] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, 2002.

[56] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17:215–224, 2001.

[57] P. Pudil, J. Novovicová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, November 1994.

[58] J. Quackenbush. Computational analysis of microarray data. *Nature Genetics*, 2:418–427, 2001.

[59] M. Ris, J. Barrera, R. M. Cesar-Jr, and D. C. Martins-Jr. Mean conditional entropy based branch and bound algorithm. 2004. (em preparação).

[60] R. Sasik, T. Hwa, N. Iranfar, and W. F. Loomis. Percolation clustering: a novel approach to the clustering of gene expression patterns in dictyostelium development. In *Pacific Symposium on Biocomputing*, volume 6, pages 335–347, 2001.

[61] D. Shalon, S. J. Smith, and P. O. Brown. A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome Res*, pages 639–45, 1996.

[62] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[63] C. E. Shannon and Warren Weaver. *The mathematical theory of communication*. Univ. of Illinois Press, 1963.

[64] P. J. S. Silva, R. Hashimoto, S. Kim, J. Barrera, L. Brandão, E. Suh, and E. R. Dougherty. Using linear programming to find strong feature sets for small samples. (submitted).

[65] R. Somogyi and S. Fuhrman. Distributivity, a general information theoretic network measure, or why the whole is more than sum of its parts. In M. Holcombe and R. Paton, editors, *Information Processing in Cells and Tissues*, pages 273–283, 1997.

[66] R. Somogyi, S. Fuhrman, M. Askenazi, and A. Wuensche. The gene expression matrix: towards the extraction of genetic network architectures. In *Nonlinear Analysis, Theory, Methods and Applications*, volume 30, pages 1815–1824, 1997.

[67] P. Somol, P. Pudil, J. Novovicová, and P. Paclík. Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20:1157–1163, 1999.

[68] E. S. Soofi. Principal information theoretic approaches. *Journal of the American Statistical Association*, (95):1349–1353, 2000.

[69] Z. Szallasi. Genetic network analysis in light of massively parallel biological data acquisition. In *Pacific Symposium on Biocomputing*, volume 4, pages 5–16, 1999.

[70] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, NY, 1999.

[71] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

[72] P. A. Viola. Alignment by maximization of mutual information. Technical Report AITR-1548, 1995.

[73] M. Wahde and J. Hertz. Coarse-grained reversed engineering of genetic regulatory networks. *BioSystems*, (55):129–136, 2000. Elsevier.

[74] S. Watanabe. *Pattern Recognition: Human and Mechanical*. Wiley, 1985.

[75] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. In *Proceedings of National Academy of Sciences*, volume 95, pages 334–339, USA, Jan 1998.

[76] A. Wuensche. Genomic regulation modeled as a network with basins of attraction. In *Pacific Simposium on Biocomputing*, pages 89–102, 1998.

[77] M. Zaffalon and M. Hutter. Robust feature selection by mutual information distributions. In *18th International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 577–584, 2002.