Capítulo

2

Reconhecimento de Padrões

2.1 O que é reconhecimento de padrões ?

Há duas maneiras de se reconhecer e/ou classificar um padrão [CONNEL, S. D. & JAIN, A. K. (2001)]: (i) classificação supervisionada: o padrão de entrada é identificado como um membro de uma classe pré-definida, ou seja , a classe é definida pelo projetista do sistema, ou (ii) classificação não supervisionada: o padrão é determinado por uma "fronteira" de classe desconhecida.

Um problema de reconhecimento de padrão consiste de uma tarefa de classificação ou categorização, onde as classes são definidas pelo projetista do sistema (classificação supervisionada) ou são "aprendidas" de acordo com a similaridade dos padrões (classificação não supervisionada).

O interesse na área de reconhecimento de padrões tem crescido muito devido as aplicações que, além de serem desafiantes, são também computacionalmente cada vez mais exigentes. A Tabela 2.1 mostra exemplos de domínios do problema com suas respectivas classes de padrões.

Com o avanço e a disponibilidade de vários recursos computacionais, tornou-se fácil o projeto e a utilização de elaborados métodos de análise e classificação de padrões. Em muitas aplicações, não existe somente uma única abordagem para

classificação que seja "ótima" e, por isso, a combinação de várias abordagens de classificadores é uma prática bastante usada.

O projeto de um sistema de reconhecimento de padrões, envolve, essencialmente, três etapas:

(i) aquisição de dados (extração de características) e pré-processamento (seleção das características mais discriminativas);

(ii) representação de dados;

(iii) tomada de decisão (construção de um classificador ou descritor).

A escolha de sensores, técnicas de pré-processamento, esquema de representação e método para a tomada de decisão, depende do domínio do problema. Um problema bem definido e suficientemente detalhado, onde se tem pequenas variações intra-classes e grandes variações inter-classes, produzirá representações compactas de padrões e, consequentemente a estratégia de tomada de decisão será simplificada. Aprender, a partir de um conjunto de exemplos (conjunto de treinamento), é um atributo importante desejado na maioria dos sistemas.

Domínio do Problema	Aplicação	Padrão de Entrada	Classes de Padrões
Bioinformática	Análise de Seqüência	DNA/Seqüência de Proteína	Tipos conhecidos de genes/padrões
Mineração de dados	Busca por padrões significantes	Pontos em um espaço multi-dimensional	Compactar e separar grupos
Classificação de documentos	Busca na Internet	Documento texto	Categorias semânticas, (negócios, entre outros)
Análise de documento de imagem	Máquina de leitura para cegos	Imagem de Documento	Caracteres alfa-numéricos, palavras
Automação industrial	Inspeção de placas de circuito impresso	Intensidade ou alcance de imagem	Natureza do produto (defeituosa ou não)
Recuperação de base de dados multimídia	Busca Internet	Vídeo *clip*	Gêneros de vídeo (p.e. ação, diálogo, entre outros.)
Reconhecimento biométrico	Identificação pessoal	Face, íris, impressão digital	Usuários autorizados para controle de acesso
Sensoriamento remoto	Prognóstico da produção de colheita	Imagem multi-espectral	Categorias de aproveitamento de terra, desenvolvimento de padrões de colheita
Reconhecimento de voz	Inquérito por telefone sem assistência de operador	Voz em forma de onda	Palavras faladas

Tabela 2.1: Exemplos de aplicações para o reconhecimento de padrões [JAIN, A. K. et al. (2000b)].

A escolha de uma abordagem para o reconhecimento de padrões não é uma tarefa simples e muitas vezes ela conta com a experiência do projetista. Na próxima seção, várias abordagens para o reconhecimento de padrões são apresentadas. Vale observar que elas não são necessariamente independentes, pois desde os primórdios da pesquisa em reconhecimento de padrões, várias são as tentativas para o projeto de sistemas híbridos [FU, K. S. (1983)]. E na literatura de reconhecimento de padrão, às vezes a mesma abordagem possui diferentes interpretações.

2.2 Algumas Técnicas para reconhecimento de padrões

Esta seção apresenta as principais técnicas para reconhecimento de padrões.

2.2.1 "Casamento" de modelos (*Template Matching*)

Uma das primeiras e mais simples abordagens para reconhecer padrões é a técnica de casamento de modelos. O "casamento" é uma operação genérica usada para determinar a similaridade entre duas entidades do mesmo tipo. O modelo é tipicamente um protótipo.

O padrão a ser reconhecido é comparado com os modelos armazenados, observando todas as variações possíveis em termos de: translação, rotação e mudanças de escalas. A medida de similaridade é frequentemente uma correlação ou uma função de distância. Muitas vezes o modelo, por si mesmo, é "aprendido" a partir do conjunto de treinamento. Esse método é computacionalmente exigente, mas a atual disponibilidade de recursos computacionais permite com que essas abordagens viabilizem-se mais facilmente [JAIN, A. K. et al. (2000a)].

O casamento de modelos faz parte das abordagens de decisão teórica que se baseiam na utilização de funções de decisão (ou discriminantes). Seja $x = (x_1, x_2, ..., x_n)^T$ um vetor de padrão n-dimensional. Para M classes de padrões $w_1, w_2, ..., w_M$, o problema básico é encontrar M funções de decisão $d_1(x), d_2(x), ..., d_M(x)$, com a propriedade de que, se o padrão x pertencer à classe w_i, então:

$$d_i(x) > d_j(x) \quad j = 1,2,...,M; \; j \neq i. \tag{2.1}$$

ou seja, um padrão desconhecido *x* pertencerá à i-ésima classe de padrões se a substituição de *x* em todas as funções de decisão fizer com que $d_i(x)$ tenha o maior valor numérico. Empates são resolvidos arbitrariamente.

A fronteira de decisão que separa as classes w_i e w_j é dada pelos valores de *x* para os quais $d_i(x) = d_j(x)$ ou, equivalentemente, pelos valores de *x* para os quais

$$d_i(x) - d_j(x) = 0 \tag{2.2}$$

É comum identificar a fronteira de decisão entre duas classes pela função $d_{ij}(x) = d_i(x) - d_j(x) = 0$. Portanto, $d_{ij}(x) > 0$ para padrões de classe w_i e $d_{ij}(x) < 0$ para padrões de classe w_j [GONZALEZ, R. C. & WOODS, R. E. (1992)].

Muitos pesquisadores atualmente se utilizam da abordagem de casamento de modelos em diversas áreas de aplicações: i) para determinar a presença de uma imagem ou objeto dentro de uma cena [CHOI, M.S. & KIM, W.Y. (2000)] e ii) para reconhecimento de caracteres [CONNEL, S. D. & JAIN, A. K. (2001)]. O aspecto da segurança em sistemas que utilizam técnicas de casamento de modelos, em aplicações de reconhecimento de pessoas, é investigado em [BOLLE, R. M. et al. (2001)], pois eles são mais vulneráveis a ataques de força bruta. Isto resulta em invasões de privacidade que acarretam grandes problemas, pois o usuário tem registrado uma imagem de parte de seu corpo no banco de dados do sistema.

2.2.2 Classificador de distância mínima

Suponha que cada classe de padrões seja representada por um vetor protótipo (ou médio):

$$m_j = \frac{1}{N_j} \sum_{x \in w_j} x \quad j = 1,2,...,M \tag{2.3}$$

em que N_j é o número de vetores de padrões de classe w_i e a soma é realizada sobre esses vetores. Uma maneira de definir a pertinência de um vetor padrão *x* desconhecido é atribuí-lo à classe de seu protótipo mais próximo. A distância Euclidiana, ou a de

Hamming, pode ser usada para determinar a proximidade, reduzindo o problema à computação das distâncias:

$$D_j(x) = \| x - m_j \| \quad j = 1,2,...,M \tag{2.4}$$

em que $\|a\| = (a^T a)^{\frac{1}{2}}$ é a norma Euclidiana. Atribui-se então, x à classe w_i se $D_i(x)$ for a menor distância. Ou seja, a menor distância implica no melhor casamento nessa formulação. Não é difícil mostrar que isso é equivalente a avaliar a função

$$d_j(x) = x^T m_j - \frac{1}{2} m_j^T m_j \quad j = 1,2,...,M \tag{2.5}$$

e atribuir x à classe w_i se $d_i(x)$ levar ao maior valor numérico. Essa formulação está de acordo com o conceito de função de decisão, como definido na Equação (2.1).

A partir das Equações (2.2) e (2.5), pode-se ver que a fronteira de decisão entre as classes w_i *e* w_j para o classificador de distância mínima é

$$d_{ij}(x) = d_i(x) - d_j(x) = x^T (m_i - m_j) - \frac{1}{2}(m_i - m_j)^T (m_i - m_j) = 0 \tag{2.6}$$

A superfície dada pela Equação (2.6) é a bisseção perpendicular do segmento de linha entre m_i *e* m_j. Para n=2 a bisseção perpendicular é uma linha, para n=3 é um plano e para n>3 é chamado de hiperplano [GONZALEZ, R. C. & WOODS, R. E. (1992)].

2.2.3 Casamento de modelos por correlação

Segundo [GONZALEZ, R. C. & WOODS, R. E. (1992)], o conceito básico de correlação de imagens é considerado como a base para encontrar casamentos de uma sub-imagem $w(x,y)$ de tamanho *J* x *K* dentro de uma imagem *f(x,y)* de tamanho *M* x *N*, supondo-se que $J \leq M$ *e* $K \leq N$. Embora a abordagem por correlação possa ser

formulada na forma vetorial, o tratamento direto com uma imagem ou sub-imagem é mais intuitivo.

Em sua forma mais simples, a correlação entre *f(x,y)* e *w(x,y)* é

$$c(s,t) = \sum_x \sum_y f(x,y)w(x-s,y-t) \tag{2.7}$$

em que *s=0,1,2,...,M-1* e *t=0,1,2,...,N-1*, e a soma é realizada sobre a região da imagem em que *f* e *w* se sobreponham. A Figura 2.1 ilustra este procedimento, sendo considerada a origem de *f(x,y)* o topo à esquerda e a de *w(x,y)* a região de seu centro. Para qualquer valor de *(s,t)* dentro de *f(x,y)*, a aplicação da Equação (2.7) leva a uma valor *c*. Na medida que *s* e *t* são varridos, *w(x,y)* é movido na área da imagem, fornecendo uma função *c(s,t)*. O valor máximo de *c(s,t)* indica a posição em que *w(x,y)* melhor se casa com *f(x,y)*. Note que se perde precisão para valores de *s* e *t* perto das bordas de *f(x,y)*, com a amplitude de erro sendo proporcional ao tamanho de *w(x,y)*.

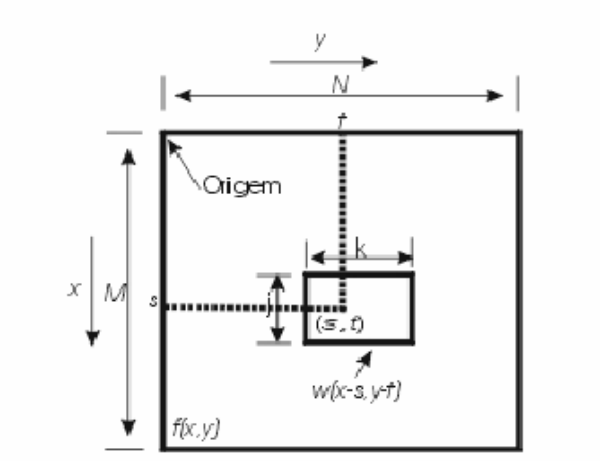

Figura 2.1: Esquema para se obter a correlação de f(x,y) e w(x,y) no ponto (s,t) [GONZALEZ, R. C. & WOODS, R. E. (1992)].

A função de correlação dada na equação (2.7) possui a desvantagem de ser sensível a mudanças na amplitude de *f(x,y)* e de *w(x,y)*. Por exemplo, dobrando-se todos os valores de *f(x,y)*, dobrar-se-ão os valores de *c(s,t)*. Uma abordagem frequentemente usada para evitar essa dificuldade é realizar o casamento através do coeficiente de correlação, que é definido como

$$g(s,t) = \frac{\sum_x \sum_y [f(x,y) - \overline{f}(x,y)][w(x-s,y-t) - \overline{w}]}{\left\{ \sum_x \sum_y [f(x,y) - \overline{f}(x,y)]^2 \sum_x \sum_y [x(x-s,y-t) - \overline{w}]^2 \right\}^{\frac{1}{2}}} \quad (2.8)$$

em que *s=01,2,3,...,M-1* e *t=0,1,2,..,N-1*, $\overline{w}$ é o valor médio dos pixels em *w(x,y)* (computado apenas 1 vez), $\overline{f}(x,y)$ é o valor médio de *f(x,y)* na região coincidente com a posição corrente de *w*, e as somas são realizadas sobre as coordenadas comuns, tanto a *f* como a *w* . O coeficiente de correlação $g(s,t)$ tem sua escala no intervalo *-1* a *1*, independentemente de mudanças na amplitude de *f(x,y)* e *w(x,y)*.

Embora a função de correlação possa ser normalizada para mudanças de amplitude através do coeficiente de correlação, a obtenção da normalização para mudanças de tamanho e rotação pode ser difícil. A normalização em relação ao tamanho envolve mudança de escala espacial, um processo que acrescenta um custo computacional considerável. Se uma pista em relação à rotação puder ser extraída de *f(x,y)*, então bastará rotacionar *w(x,y)* de maneira que ela mesma se alinhe com o grau de rotação de *f(x,y)*. Entretanto se a natureza da rotação for desconhecida, as busca pelo melhor casamento requererá rotações exaustivas de *w(x,y)*. Esse procedimento é impraticável e, por conseguinte, a correlação é raramente usada em casos em que rotação arbitrária ou sem restrições estejam presentes [GONZALEZ, R. C. & WOODS, R. E. (1992)].

2.2.4 Técnicas estatísticas

Em reconhecimento de padrões com abordagem estatística, um padrão é representado por um conjunto de características chamado de vetor de características *d*-dimensional. Os conceitos da teoria de decisão estatística são utilizados para estabelecer fronteiras de decisão entre classes e padrões. O sistema de reconhecimento é operado em dois modos: treinamento (aprendizagem) e classificação (teste) (de acordo com a Figura 2.2) [JAIN, A. K. et al. (1996)].

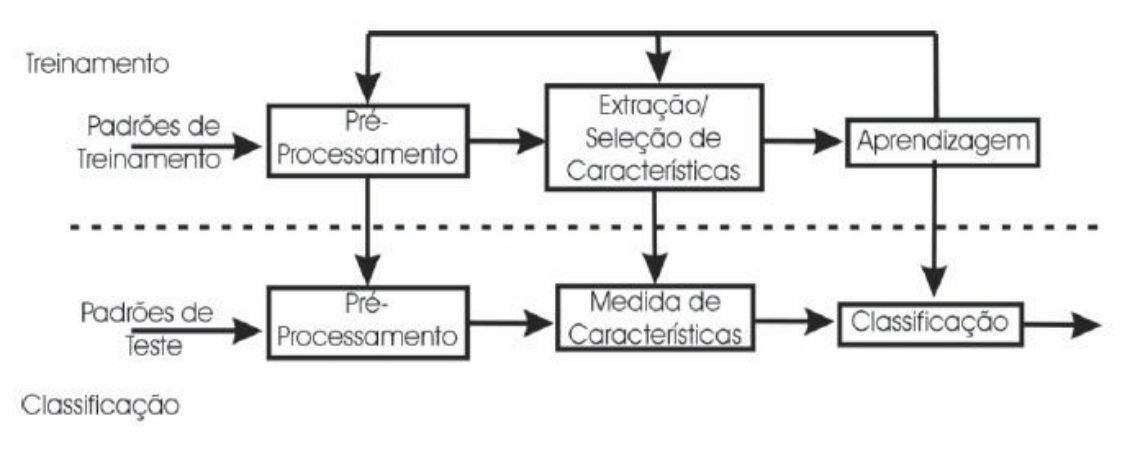

Figura 2.2: Blocos funcionais para o reconhecimento de padrão na abordagem estatística [JAIN, A. K. et al. (2000a)].

Na Figura 2.2 a função do módulo de pré-processamento é capturar o padrão de interesse, remover ruído, normalizar e realizar qualquer outra operação que contribua para a definição de uma representação compacta do padrão.

Um dos problemas óbvios encontrados, principalmente quando o padrão trata-se de uma imagem, é a alta dimensionalidade dos dados de entrada. Técnicas que combinam as variáveis (características) de entrada mais próximas para produzir um menor número das mesmas, ajudam a aliviar tais problemas. Essas técnicas podem ser construídas "manualmente", com base no problema particular, ou podem ser derivadas dos dados, a partir de procedimentos automáticos [BISHOP, C. M. (1996)]. Esses métodos são chamados de extração e seleção de características e serão vistos com mais detalhes nas seções seguintes. Eles estão presentes no módulo de treinamento, parte superior da Figura 2.2. Para encontrar características apropriadas às representações de padrões de entrada o classificador é treinado para particionar o espaço de características. Otimizações do pré-processamento e das estratégias de extração e seleção de características são realizados no caminho recorrente da Figura 2.2. No modo classificação, o classificador treinado mapeia o padrão de entrada em uma das classes de padrões sob consideração, baseado nas características medidas.

O processo de tomada de decisão estatística em reconhecimento de padrões pode ser sintetizado como segue: seja um padrão representado por um vetor $x = (x_1, x_2, ..., x_d)$ com d características, ele será determinado a uma das c classes $w_1, w_2, ..., w_c$. Supõe-se que cada característica apresente uma densidade de probabilidade (dependendo das características serem contínuas ou discretas)

condicionada à cada classe. Assim, um padrão x pertencente a uma classe w_i é visto como uma observação extraída aleatoriamente a partir de uma função de probabilidade classe-condicional $p(x|w_i)$. As regras de decisão, incluindo a regra de decisão de Bayes, a regra da probabilidade máxima (que pode ser vista como um caso particular da regra de Bayes) e a regra Neyman-Peason são eficazes para definir a fronteira de decisão. A regra de decisão "ótima" de Bayes para a minimização do risco condicional pode ser declarada como segue:

$$R(w_i \mid x) = \sum_{j=1}^{c} L(w_i, w_j).P(w_j \mid x) \tag{2.9}$$

Ela determina a classe w_i para o padrão de entrada x onde o risco condicional é mínimo, $L(w_i, w_j)$ é a função de perda causada na decisão de w_i quando a classe verdade é w_j *e* $P(w_i \mid x)$ é a probabilidade posterior [JAMISON, T. A. & SCHALKOFF, R. J. (1998)]. No caso da função perda ser 0/1, como definido na equação 2.10, o risco condicional torna-se a probabilidade condicional de falsa classificação.

$$L(w_i, w_j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \tag{2.10}$$

Para a escolha da função de perda, a regra de decisão de Bayes pode ser simplificada como mostrado na equação 2.11. Ela determina o padrão de entrada x para a classe w_i se

$$P(w_i \mid x) > P(w_j \mid x), \ \textit{para todo} \ \ j \neq i \tag{2.11}$$

Várias estratégias são utilizadas para projetar um classificador para o reconhecimento de padrões com abordagem estatística, dependendo da espécie de informação disponível a respeito de densidades de classe-condicional. Se todas elas são especificadas, então a regra de decisão ótima de Bayes pode ser usada para a classificação. Entretanto, densidades de classe-condicional são frequentemente desconhecidas na prática e devem ser aprendidas a partir dos padrões de treinamento disponíveis. Se a forma da densidade classe-condicional é conhecida, por exemplo uma Gaussiana multivariada, mas alguns dos parâmetros de densidades, por exemplo,

vetores médio e matrizes de covariância, são desconhecidos, então tem-se um problema de decisão paramétrico. Uma estratégia comum para esses tipos de problemas é substituir os parâmetros desconhecidos na função densidade por seus valores estimados. Se a forma da densidade classe-condicional não é conhecida, então opera-se em um modo não paramétrico. Neste caso, estima-se a função densidade (ex: abordagem janela Parzen) ou constrói-se diretamente a fronteira de decisão baseada nos dados de treinamento (ex: *k*-ézimo vizinho mais próximo). O perceptron multicamada pode ser visto como um método supervisionado não paramétrico que constrói uma fronteira de decisão.

Outra dicotomia na abordagem estatística para o reconhecimento de padrões é a do aprendizado supervisionado *versus* o aprendizado não supervisionado. Em um problema de aprendizado não supervisionado, algumas vezes o número e as estruturas de classes devem ser aprendidas mediante o conjunto de exemplos de treinamento. As várias dicotomias são mostradas na árvore de estruturas da Figura 2.3.

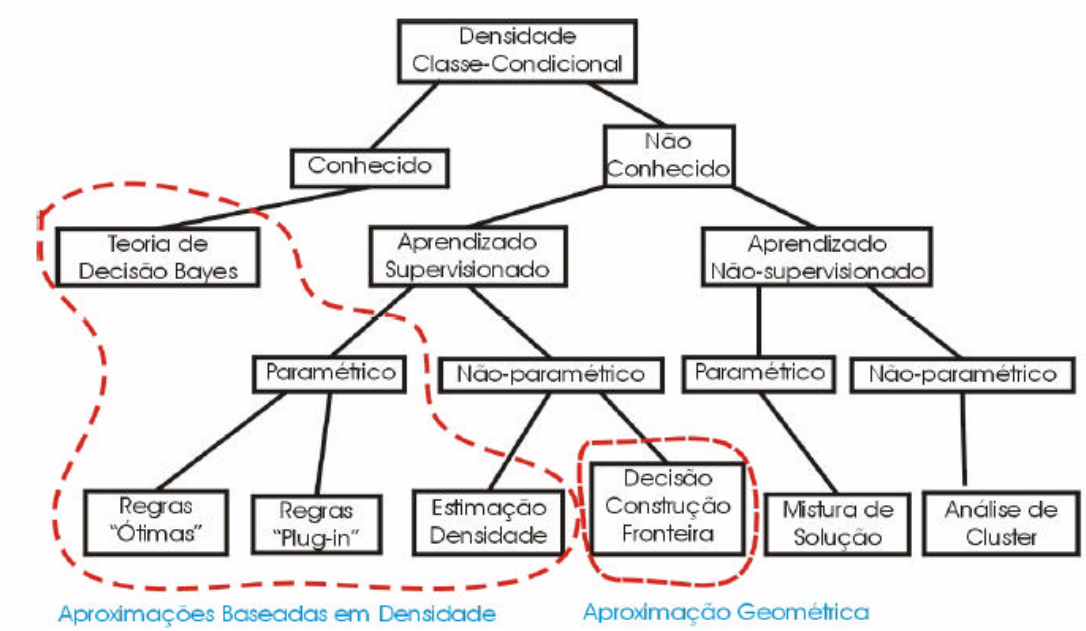

Figura 2.3: As várias abordagens estatísticas para o reconhecimento de padrão. [JAIN, A. K. et al. (2000a)].

À medida que se percorre a árvore de cima para baixo e da esquerda para a direita, menos informações a respeito das características e classes de padrões são disponíveis e, como resultado, a dificuldade de classificação aumenta. Em alguns casos, a maioria dos métodos (nas folhas da árvore da Figura 2.3) são tentativas de

implementar a regra de decisão de Bayes. A análise de agrupamentos (*cluster*) trata de problemas de tomada de decisão no modo não paramétrico e aprendizado não supervisionado [JAIN, A. K. & DUBES, R. C. (1998)], onde o número de categorias ou *clusters* não é especificado; a tarefa é descobrir uma categorização razoável dos dados (se existir alguma). Algoritmos de análise de agrupamentos junto com várias técnicas para visualização e projeção de dados multi-dimensionais são também referidas como métodos de análise exploratória de dados.

Ainda outra dicotomia baseia-se na maneira como as fronteiras de decisão são obtidas: direta (abordagem geométrica) ou indireta (abordagem baseada em densidade probabilística). A abordagem probabilística requer primeiro que a função de densidade seja estimada, para então construir as funções discriminantes que especificam as fronteiras de decisão. Por outro lado a abordagem geométrica frequentemente constrói fronteiras de decisão diretamente, através de funções de custo fixo.

Não importa qual seja a regra de classificação ou decisão usada, ela deve ser treinada usando os exemplos de treinamento disponíveis e o desempenho de um classificador dependerá disso e da quantidade desses exemplos. Ao mesmo tempo, o objetivo principal de um sistema de reconhecimento é classificar exemplos de testes futuros, os quais são provavelmente diferentes dos exemplo vistos durante o treinamento.

2.3 Super-treinamento e super-adaptação

Otimizar um classificador para maximizar seu desempenho sobre o conjunto de treinamento pode nem sempre resultar no desempenho desejado para um conjunto de teste. A habilidade de generalização de um classificador refere-se ao seu desempenho em classificar padrões de testes que não foram usados durante o estágio de treinamento. Uma habilidade pobre de generalização pode ser atribuída a qualquer um dos seguintes fatores:

(i) número de características muito grande relativo ao número de exemplos de treinamento;

(ii) grande número de parâmetros desconhecidos associados ao classificador (ex: classificadores polinomiais ou uma rede neural com número excessivo de neurônios na camada intermediária);

(iii) um classificador ser intensivamente otimizado no conjunto de treinamento (super-treinamento).

O super-treinamento, também é análogo ao fenômeno de super-adaptação em regressão, quando existem muitos parâmetros livres. Esses fenômenos são teoricamente investigados através de classificadores que minimizam a taxa de erro aparente (o erro no conjunto de treinamento). Há várias fases no fenômeno de super-treinamento, por exemplo, dependendo da relação entre o numero t de exemplos e o número m de parâmetros modificáveis. Quando t é menor ou quase igual a m, os exemplos podem em princípio, ser memorizados e a sobre-adaptação é elevada nesta fase, principalmente quando $t \approx m$.

O super-treinamento pode ser dividido em duas categorias:

(i) **Absoluto**, quando o desempenho de classificação degrada para todas as categorias de padrões e

(ii) **Relativo**, quando o desempenho de classificação degrada para algumas categorias, enquanto para outras permanece inalterado ou até mesmo melhora.

Às vezes, há dominância de padrões de algumas categorias no conjunto de treinamento, ocasionando um super-treinamento do classificador que se adaptará às mesmas. Isso é considerado um super-treinamento relativo. O super-treinamento absoluto ocorre principalmente devido ao conjunto de treinamento ser um limiar representativo para o conjunto de teste. Por outro lado, o super-treinamento ocorre usualmente devido ao conjunto de treinamento apresentar padrões "confusos" nas regiões do envoltório da fronteira de decisão [CHOI, M.S. & KIM, W.Y. (2000)].

Os estudos clássicos de [GROHMAN, W. M. & DHAWAN, A. P. (2001)], sobre a capacidade de complexidade de classificadores, provêem um bom entendimento dos mecanismos que levam ao super-treinamento. Classificadores complexos, por exemplo, aqueles tendo muitos parâmetros independentes, podem ter uma grande capacidade, isto é, eles são hábeis para representar muitas dicotomias para um dado conjunto de dados.

As armadilhas da super-adaptação em estimadores, para um dado conjunto de treinamento, são observadas em muitos estágios de um sistema de reconhecimento de padrões, tais como na redução de dimensionalidade, estimativa de densidade, e construção do classificador. O conceito de super-adaptação refere-se à demasiada

adaptação e ajuste do classificador a exemplos específicos, perdendo assim sua capacidade de generalização. Em alguns casos consiste de uma distorção local da fronteira de decisão, ou seja, não cabe supor que sua ocorrência é simultânea em todo o espaço de características e a distorção pode ocorrer em diferentes locais em diferentes momentos. Isto implica que em alguns locais a fronteira de decisão é continua, enquanto em outras áreas a super-adaptação já está presente [ROSIN, P. L. & FIRENS, F. (1995)]. Uma solução certa é sempre usar um conjunto teste independente do conjunto de treinamento para avaliação. Para evitar a necessidade de muitos conjuntos de testes independentes, estimadores são frequentemente baseados em subconjuntos dos dados rotacionados, preservando diferentes partes dos dados para otimização e avaliação.

2.4 O problema da dimensionalidade e o fenômeno de máximo

O desempenho de um classificador depende do inter-relacionamento entre o tamanho do conjunto de exemplos, o número de características dos padrões e a sua complexidade. Seja o exemplo de uma simples técnica de tabela de consulta, onde se particiona o espaço de características em células e se associa um nome de classe a cada célula. Isso requer que o número de exemplos de treinamento seja uma função exponencial da dimensão de características [CHAMP, P. (1994)]. Esse fenômeno é chamado de “maldição da dimensionalidade”, que conduz ao “fenômeno de máximo” em um projeto de classificador [JAIN, A. K. et al. (2000a)].

A probabilidade de classificação falsa de uma regra de decisão não aumenta na mesma proporção que aumenta o número de características, dado que as densidades classe-condicional sejam completamente conhecidas. Entretanto, tem-se frequentemente observado que, na prática, o aumento de características pode degradar o desempenho de um classificador se o número de exemplos de treinamento que foi usado para projetar o classificador é relativamente pequeno em relação ao número de características. Este é um comportamento paradoxal referido como “fenômeno de máximo” [SUNG, K. K. & POGGIO, T. (1998)]. Uma simples explanação sobre este fenômeno é dada a seguir. A maioria dos classificadores paramétricos geralmente usados estima parâmetros não conhecidos e liga-os a parâmetros verdadeiros nas densidades de classe-condicional. Em uma amostra de tamanho fixo, quando o número de características cresce (à medida que aumenta o número de parâmetros desconhecidos) a confiança dos parâmetros estimados

decresce. Consequentemente, o desempenho dos classificadores, para uma amostra de tamanho fixo, pode degradar com um aumento no número de características.

Todos os classificadores geralmente usados, incluindo redes neurais diretas, podem sofrer o problema da dimensionalidade, pois é muito difícil estabelecer um exato relacionamento entre a probabilidade de falsa classificação, o número de exemplos de treinamento, o número de características e os parâmetros verdadeiros das densidades de classe-condicional. Algumas linhas de direção são sugeridas com base no tamanho do conjunto de exemplos para dimensionalidade. É geralmente aceitável que o número de exemplos de treinamento por classe seja, pelo menos, dez vezes o número de características *(n/d>10)*. Isto seria uma boa prática a se seguir no projeto de um classificador [SUNG, K. K. & POGGIO, T. (1998)], maior deveria ser a proporção do tamanho de exemplos para ser evitado o problema da dimensionalidade.

2.5 Redução da dimensionalidade

As vantagens em reduzir a dimensionalidade da representação do padrão refletem-se na medida de custo e precisão do classificador. Além disso, uma pequena quantidade de características pode aliviar o problema da dimensionalidade, quando o número de exemplos de treinamento é pequeno. Porém, um reduzido número de características pode levar a uma fraca discriminação e consequentemente a uma precisão inferior no sistema de reconhecimento resultante. Mas a redução de dimensionalidade é necessária quando, por exemplo, é possível construir dois padrões arbitrários similares, codificando-os a partir de um grande número de características redundantes [WATANABE, S. (1985)]. No entanto, toda redução de dimensionalidade implica numa perda de informação, e esta última pode vir a ser fundamental para discriminação dos padrões. Por isto, o objetivo principal das técnicas de redução de dimensionalidade é preservar o máximo possível da informação relevante dos dados.

Existem diferenças entre seleção e extração de características, embora na literatura elas sejam usadas indistintamente. O termo seleção refere-se a algoritmos que procuram selecionar o melhor subconjunto de um conjunto de características de entrada. Já algoritmos de extração são métodos que criam novas características a partir de transformações ou combinações do conjunto de características original. Frequentemente, a extração precede a seleção, pois primeiro as características são extraídas a partir do sentido dos dados (usando componente principal ou análise

discriminante) e então algumas características extraídas, com baixa habilidade de discriminação, são descartadas.

A escolha entre seleção e extração depende do domínio de aplicação e dos dados específicos de treinamento disponíveis. A seleção conduz à economia na medida de custo quando algumas características são descartadas e as que foram selecionadas, retém suas interpretações físicas originais. Além do mais, as mesmas podem ser importantes para o entendimento do processo físico que gera os padrões. Por outro lado, transformações geradas por extração podem prover uma melhor habilidade discriminativa do que o melhor subconjunto de características originais, mas estas novas características podem não ter um claro sentido físico.

O ponto principal da redução de dimensionalidade é a escolha de uma função de critério. Um critério geralmente usado é o erro de classificação segundo um subconjunto de características. Porém, o erro de classificação, por si só, não é confiável quando a quantidade de exemplos de padrões é pequena em relação ao número de características. E ainda mais, para a escolha de uma função critério, é necessário determinar a dimensionalidade apropriada do espaço de características reduzido. E em reposta a isto surge a noção de dimensionalidade intrínseca dos dados, que consiste em determinar se os padrões *d*-dimensionais originais podem ser descritos adequadamente em um subespaço de dimensionalidade menor do que *d.* Por exemplo, padrões *d*-dimensionais ao longo de uma curva aplainada tem uma dimensionalidade intrínseca de um, independente do valor de *d*. Deve-se perceber que dimensionalidade intrínseca não é o mesmo que dimensionalidade linear, que consiste de uma propriedade global dos dados, envolvendo o número de autovalores significativos da matriz de covariância dos dados. Apesar de haver muitos algoritmos disponíveis para estimar a dimensionalidade intrínseca [TIBBALDS, A. D. (1998)], eles não indicam quão facilmente um subespaço de dimensionalidade pode ser identificado.

2.6 **Extração de características**

Segundo [JAIN, A. K. et al. (2000a)], um método de extração de características determina um subespaço apropriado de dimensionalidade *m* (de uma maneira linear ou não–linear) no espaço de características original de dimensionalidade $d (m \leq d)$. A transformada linear, assim como a análise de componentes principais (PCA) ou

expansão Karhunen-Loève computam os *m* maiores autovetores da matriz de covariância *d* x *d* de *n* padrões d–dimensionais. A transformação linear é definida como

$$Y_{nxm} = X_{nxm} H_{nxm}^{T} \tag{2.12}$$

onde *X* é a matriz de padrão *n* x *d*, *Y* é a matriz derivada *n* x *m*, e *H* é a matriz de transformação linear *m x d*, cujas colunas são auto-vetores. Visto que PCA usa as características mais expressivas (auto-vetores com os maiores autovalores), ele efetivamente aproxima os dados para um subespaço linear usando o critério do erro quadrático médio. Existem outros métodos que são mais apropriados para distribuições não-Gaussianas.

Enquanto que PCA é um método de extração de características linear e não supervisionado, análise discriminante usa a informação de categoria associada com cada padrão para extração (linear) da maioria das características discriminatórias. Nela a separação inter-classes é feita por uma medida de separabilidade que resulta no encontro de auto-vetores de $S_w^{-1} S_b$ (o produto do inverso da matriz de espalhamento do interior da classe S_w e a matriz de espalhamento entre as classes S_b) [MARR, D. (1982)].

Existem muitas maneiras de definir técnicas de extração de características não lineares. Um método semelhante e diretamente relacionado ao PCA é chamado de Kernel PCA [HOPCROFT, J. E. & ULLMAN, J. D. (1979)]. A idéia básica do kernel PCA é primeiro mapear os dados de entrada dentro de algum novo espaço de característica *F*, via uma função não linear Φ (por exemplo, polinomial de grau *p;p>1*) e então executar um PCA linear no espaço mapeado.

Escalonamento multidimensional (MDS) é outra técnica de extração de características não linear. Seu objetivo é representar um conjunto de dados multidimensional em 2 ou 3 dimensões semelhantes onde a matriz distância, no espaço de característica *d*-dimensional original é preservada tão fielmente quanto possível no espaço projetado. Um problema com MDS é que ele não possui uma função de mapeamento explícita. Assim, não é possível estabelecer um novo padrão em um mapa já computado por um dado conjunto de treinamento, sem ter que repetir o mapeamento. Muitas técnicas tem sido investigadas para tratar essa deficiência que abrange desde interpolação linear até o treinamento de uma rede neural.

Uma rede neural direta oferece um procedimento integrado para extração de características e classificação. A saída de cada camada intermediária pode ser interpretada como um conjunto de novas características, frequentemente não lineares, apresentadas à camada de saída para classificação. Nesse sentido, redes multi-camadas servem como extratores de características [ZIMMERMANN, A. C. et al. (2000)]. Por exemplo, as redes que apresentam as então chamadas "camadas de pesos compartilhados", são de fato filtros para extração de características em imagens bi-dimensionais. Durante o treinamento, os filtros são direcionados para os dados de maneira a maximizar o desempenho da classificação.

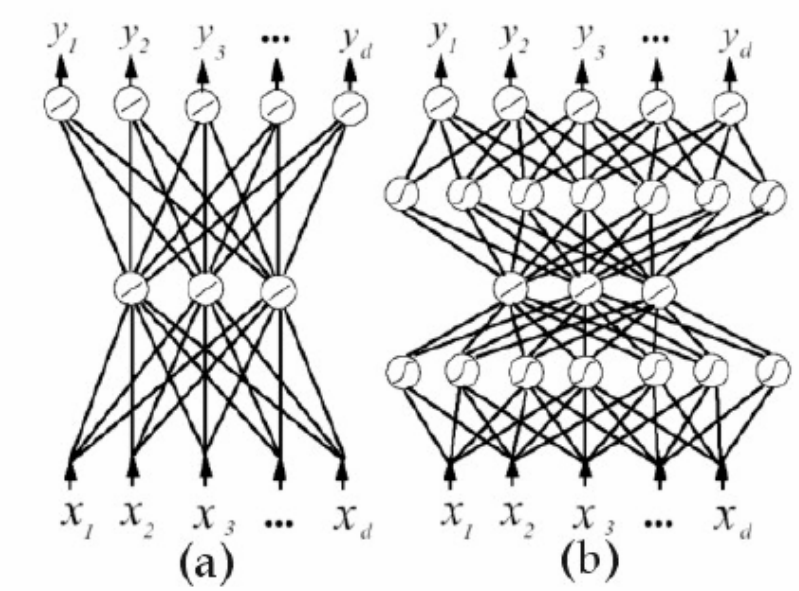

Figura 2.4: Redes Auto-associativas para encontrar um subespaço. (a) linear, (b) não-linear (nem todas as conexões são mostradas) [JAIN, A. K. et al. (2000a)].

Redes neurais também podem ser usadas diretamente para extração de características em um modo não supervisionado. A Figura 2.4 mostra a arquitetura de uma rede que é hábil para encontrar o subespaço PCA. Ao invés de funções sigmóides, os neurônios tem funções de transferência lineares. Esta rede tem *d* entradas e *d* saídas, onde *d* é o número de características dado. As entradas são também usadas como saídas desejadas, forçando a camada de saída a reconstruir o espaço de entrada usando somente uma camada intermediária. Os três nós na camada intermediária capturam os primeiros três componentes principais [CHAMP, P. (1994)]. Se duas camadas não lineares, com unidades intermediárias contendo funções de transferência sigmoidais, são incluídas

também (veja Figura 2.4(b)), então um subespaço não linear é encontrado na camada intermediária (também chamada de camada de gargalo). A não linearidade é limitada pelo tamanho dessas camadas adicionais. Estas, então chamadas redes auto-associativas, ou redes PCA não lineares , oferecem uma poderosa ferramenta para treinar e descrever subespaços não lineares [ZADEH, L.A. (1973)].

O mapa de Kohonen [KOHONEN, T. (1995); YAO, X. & LIU, Y. (1998)], pode também ser usado para extração de características não lineares. Nesta rede, conhecida como rede SOM, os neurônios são dispostos em um espaço *m*-dimensional, onde *m* é geralmente *1,2* ou *3*. Cada neurônio é conectado a todas às *d* unidades de entrada. Os pesos das conexões de cada neurônio formam um vetor de pesos *d*-dimensional. Durante o treinamento, padrões são apresentados à rede de forma aleatória. A cada apresentação o vencedor, que é o vetor peso mais próximo do vetor de entrada, é identificado primeiro. Então, todos os neurônios na vizinhança do vencedor são atualizados de modo que seus vetores de pesos movam-se em direção ao vetor de entrada. Depois que o treinamento é feito, os vetores de pesos dos neurônios da vizinhança tornam-se bem parecidos com os padrões de entrada que estão próximos no espaço de características original. Assim, uma mapa de "preservação de topologia" é formado, ou seja, a rede SOM oferece um mapa *m*-dimensional com uma conectividade espacial, que pode ser interpretada com a extração de características.

2.7 Seleção de Características

O problema da seleção é: para um dado conjunto de *d* características, selecionar um subconjunto de tamanho *m* que conduza ao menor erro de classificação. O interesse da aplicação de métodos de seleção deve-se ao grande número de características encontradas nas seguintes situações: (i) união de multi-sensores e (ii) integração de múltiplos modelos de dados [JAIN, A. K. et al. (2000a)].

Seja *Y* o conjunto de características dado, com cardinalidade *d* e seja *m* o número de características desejado no subconjunto selecionado $X, X \subseteq Y$. Seja *J(X)* a função de critério de seleção para o conjunto *X*. Supõe-se que o maior valor de *J* indique um melhor subconjunto de características; a escolha natural da função critério é $J = (1 - P_e)$, onde P_e denota o erro de classificação. O uso de P_e na função critério faz o procedimento de seleção depender do classificador usado e dos tamanhos dos conjuntos

de treinamento e teste. A maioria das abordagens diretas para o problema de seleção irá requerer (i) exame de todos os possíveis $\binom{d}{m} = \frac{d!}{m!\,(d-m)!}$ subconjuntos de tamanho m e (ii) seleção do subconjunto com o maior valor de $J(.)$. Entretanto o número de subconjuntos possíveis cresce combinatorialmente, fazendo desta uma busca exaustiva impraticável mesmo para valores pequenos de m e d. O único método de seleção ótimo que evita a busca exaustiva pelo uso de resultado intermediários para o valor final de critério, está baseado no algoritmo de ramificação e fronteira [JAIN, A. K. et al. (2000a)].

Dado que os procedimentos de extração e seleção de características tenham encontrado uma representação apropriada para os padrões, deve-se escolher a abordagem na qual o classificador estatístico será projetado, que na prática é um problema difícil e na maioria das vezes esta escolha é frequentemente baseada na experiência do projetista e nos acontecimentos ocorridos entre classificador e usuário [JAIN, A. K. et al. (2000a)].

2.8 Considerações Finais

Neste capítulo foram discutidos métodos de reconhecimento de padrões, considerando os métodos de extração de características que determinam um subespaço apropriado de dimensionalidade m (de uma maneira linear ou não–linear) no espaço de características original de dimensionalidade $d (m \leq d)$. Também foi abordado o problema da seleção, no qual dado conjunto de d características, selecionar um subconjunto de tamanho m que conduza ao menor erro de classificação.

Foi também mostrado o problema da dimensionalidade no qual a probabilidade de classificação falsa de uma regra de decisão não aumenta na mesma proporção que aumenta o número de características, dado que as densidades classe-condicional sejam completamente conhecidas e com isto foi mostrado também as vantagens em reduzir a dimensionalidade da representação do padrão que se refletem na medida de custo e precisão do classificador.