Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 6757-6764.

Support Vector Machines na Classificação de Imagens Hiperespectrais

Rafaela Andreola[1]
Victor Haertel[1]

[1]Universidade Federal do Rio Grande do Sul – UFRGS
Centro Estadual de Pesquisas em Sensoriamento Remoto e Meteorologia - CEPSRM
Caixa Postal 15052 - CEP 91501-970 - Porto Alegre - RS, Brasil
rafaela.andreola@gmail.com; victor.haertel@ufrgs.br

Abstract. In this paper we investigate the performance of the Support Vector Machine (SVM) classifier when applied to high dimensional image data, depicting natural scenes. As the SVM classifier deals with a pair of classes at a time, a multi-stage classifier, structured as a binary tree is proposed in this study. At each node in the tree we search for the pair of classes showing the largest separability. Samples of these two classes are then used to train the SVM classifier at this node and the remaining classes are allocated to one of the two descending nodes, or replicated in both. This procedure is repeated at every node in the tree, until the terminal nodes are reached. The RBF kernel is used in this study. Tests are performed using AVIRIS hyperspectral image data covering a test area which includes classes spectrally very similar, separable in high-dimensional spaces only.

Palavras-chave: support vector machines, binary tree classifier, high-dimensional image data

1. Introdução

O pequeno número de bandas que caracteriza os sensores multiespectrais, tradicionalmente utilizados em sensoriamento remoto, em geral é suficiente para discriminar a maioria das classes que ocorrem em cenas naturais (florestas, culturas agrícolas, corpos de água, rochas e solos, áreas urbanas, etc). Entretanto, essa capacidade de discriminar é limitada quando estão presentes na cena classes espectralmente muito semelhantes entre si, isto é, classes cujos vetores de médias são muito próximos entre si. Sensores hiperespectrais podem ser usados para auxiliar nesse problema. Pode-se mostrar que classes espectralmente muito semelhantes entre si, ou mesmo idênticas, isto é, classes que compartilham do mesmo vetor de médias podem, não obstante, serem separadas com alta acurácia em espaços de alta dimensionalidade, desde que suas matrizes de covariância sejam suficientemente distintas (Fukunaga, 1990). Do ponto de vista metodológico, classificação de dados em alta dimensionalidade não é uma tarefa trivial. No caso de classificadores paramétricos, um problema consiste na estimação dos valores para os parâmetros do classificador, a partir de um número de amostras de treinamento geralmente pequeno quando comparado com a dimensionalidade dos dados. No inicio do processo de classificação, a acurácia da imagem temática produzida por um classificador paramétrico tende a crescer, na medida em que novas bandas, isto é, informações adicionais, são incluídas no processo. Em um determinado ponto o valor da acurácia atinge um máximo, para em seguida passar a declinar, na medida em que novas bandas continuam a ser adicionadas ao processo. Esta configuração, conhecida como "fenômeno de Hughes" resulta da impossibilidade de se obter estimativas confiáveis para um número crescente de parâmetros, a partir de um número fixo (e limitado) de amostras de treinamento.

Muitos trabalhos têm sido publicados na literatura para reduzir este problema. De acordo com Melgani e Bruzzone (2004), quatro técnicas principais podem ser identificadas:
(a) Técnicas em Análise Discriminante Regularizada (regularização das matrizes de covariância);
(b) Incremento no número de amostras de treinamento, com a introdução das chamadas amostras semi-rotuladas, juntamente com as amostras de treinamento disponíveis (amostras rotuladas);

(c) Redução na dimensionalidade dos dados, com perda mínima de informação (técnicas de seleção ou extração de variáveis);

Neste contexto, desperta o interesse a utilização de classificadores não-paramétricos, como é o caso de SVM, que apresenta a vantagem de não ser afetado por este tipo de problema (Huang et al, 2002). Algumas investigações experimentais apontam a eficácia do SVM para a análise desses dados (espaço característico hiper-dimensional), sem a necessidade de nenhum procedimento de seleção ou extração de variáveis (Melgani e Bruzzone, 2004). Investigações com respeito a aplicações de técnicas de SVM no processo de classificação de dados de imagens em alta dimensionalidade como aquelas obtidas por satélites de sensoriamento remoto apresentam um interesse tanto no desenvolvimento de novos conhecimentos quanto no de aplicações práticas.

1.1 SVM

Support Vector Machines (SVM) é uma técnica de aprendizado de máquina, fundamentada nos princípios da Minimização do Risco Estrutural (*Structural Risk Minimization* – SRM). Esta técnica busca minimizar o erro com relação ao conjunto de treinamento (risco empírico), juntamente com o erro com relação ao conjunto de teste. A motivação para esse princípio surgiu da necessidade de desenvolver limites teóricos para a capacidade de generalização dos sistemas de aprendizagem. Uma maior generalização normalmente implica em um número maior de acertos na fase de teste. Quanto mais ajustada for superfície de decisão aos dados do conjunto de treinamento, isto é, quanto mais complexo for o hiperplano de decisão dessas funções no espaço de entrada dos dados, maior será o risco estrutural (Cherkassky e Mulier (1998) *apud* Netto (2005)). O objetivo de SVM consiste em obter um equilíbrio entre ambos os erros, minimizando o excesso de ajustes (*overfitting*) e melhorando a capacidade de generalização.

A função de decisão que maximiza a habilidade de generalização é determinada pelo problema de duas classes. Assumindo que as amostras de treinamento das diferentes classes são linearmente separáveis, a função de decisão mais adequada é aquela para a qual a distância entre os conjuntos das amostras de treinamento é maximizada. Neste contexto, a função de decisão que maximiza esta separação é denominada de *ótima* (Figura 1).

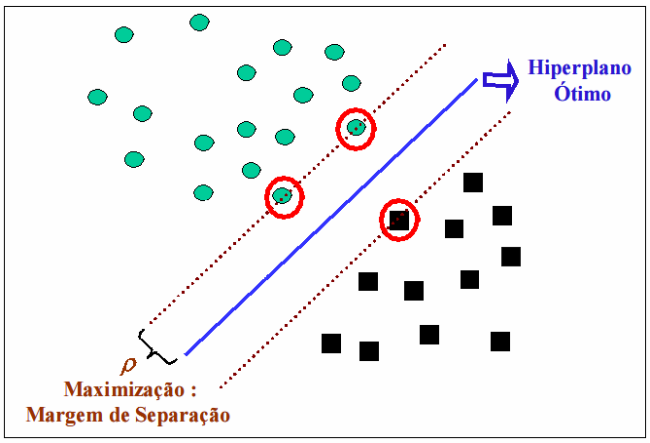


Figura 1. O hiperplano ótimo separando os dados com a máxima margem ρ. Os *support vectors* (circulados em vermelho) e uma distribuição dos dados no R^2. Fonte: Adaptado de Semolini (2002).

Seja $\boldsymbol{X}_i$ ($i=1, 2, ..., M$) um conjunto de treinamento em um problema que consiste de duas classes linearmente separáveis (ω_1 e ω_2). A cada amostra fica associado um rótulo: $y_i=1$ se $\boldsymbol{X}_i \in \omega_1$, $y_i=-1$ se $\boldsymbol{X}_i \in \omega_2$. Nesse caso a função de decisão linear adquire a forma:

$$D(x) = w^T x + b \qquad (1)$$

onde w é um vetor m-dimensional (pesos) e b é o termo independente, para $i=1, 2, ..., M$.

Porém, freqüentemente as duas classes não são linearmente separáveis, isto é, a separação entre as amostras de treinamento das duas classes requer uma função não-linear. A solução mais simples nestes casos consistiria na adoção de polinômios de grau mais elevado. Entretanto, esta abordagem apresenta, segundo Duda et al (2000), o risco de excesso de ajuste (*overfitting*), o qual resulta em perda de generalização do classificador.

Para tratar dos casos linearmente não-separáveis, se introduz a variável de folga (*slack variable*) $\xi_i(\geq 0)$ (Figura 2). Neste caso, amostras de treinamento x_i, para as quais $0<\xi_i<1$, são corretamente classificados, embora sem a margem de separação máxima. Amostras x_i para as quais $\xi_i \geq 1$, são classificados erroneamente pelo hiperplano ótimo. Neste caso, o hiperplano de separação ótimo pode ser obtido seguindo-se uma abordagem semelhante àquela adotada para o caso de amostras linearmente separáveis.

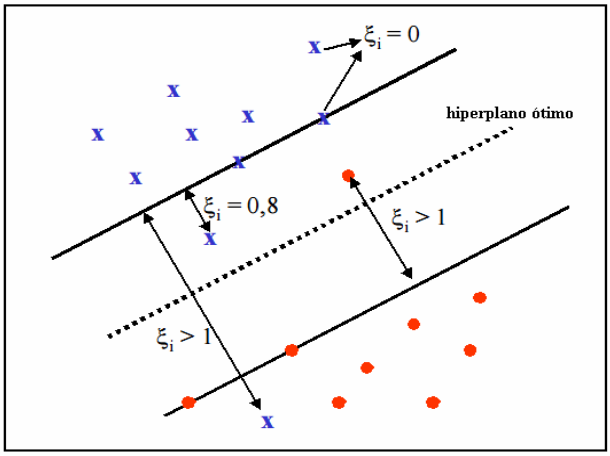


Figura 2. Exemplos de valores e situações da variável de folga ξ. Distribuição dos dados no R^2.
Fonte: Adaptado de Semolini (2002).

Uma alternativa, nestes casos, consiste em mapear os dados para um espaço de dimensão mais alta, no qual os dados passam a ser linearmente separáveis, segundo Fukunaga (1990). No contexto de SVM, esse espaço é denominado de espaço característico *(feature space)*.

Representando por $g(x)=(g_1(x), ..., g_l(x))^T$ uma função de transformação que mapeia as amostras x_i do espaço original para um espaço característico de dimensão mais elevada (l), a nova função de decisão neste novo espaço passa a ser dada por:

$$D(x) = w^T g(x) + b \qquad (2)$$

onde w é um vetor l-dimensional e b é o termo independente (*bias*).

De acordo com a teoria de Hilbert-Schmidt, se uma função simétrica $H(x,x')$ satisfaz a seguinte condição:

$$\sum_{i,j=1}^{M} h_i h_j H(x_i, x_j) \geq 0 \qquad (3)$$

para todo M, x_i e h_i, onde M é um número natural e h_i é um número real, então existe uma função de mapeamento $g(x)$, que mapeia x no espaço característico, tal que:

$$H(x, x') = g^T(x) g(x') \qquad (4)$$

A condição (3) é chamada condição de *Mercer*, e a função que satisfaz essa condição chama-se *Mercer kernel* ou simplesmente *kernel* (Abe, 2005). O teorema de Mercer permite saber quando uma função candidata a *kernel* é de fato um produto interno em algum espaço.

Este teorema, entretanto, não indica como obter $H(x, x')$. A vantagem do uso de *kernels* é que não se precisa lidar com o espaço característico de alta-dimensão explicitamente: usa-se *H(x,x')* no treinamento e classificação ao invés de *g(x).*

Usando o *kernel*, o problema de separação de um par de classes no espaço pode ser resolvido maximizando:

$$Q(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j H(x_i, x_j) \quad (5)$$

sujeito às restrições:

$$\sum_{i=1}^{M} y_i \alpha_i = 0 \quad \text{e} \quad 0 \leq \alpha_i \leq C \quad \text{para} \quad i=1, ..., M \quad (6)$$

Pode-se mostrar que neste caso, a função de decisão assume a seguinte forma:

$$D(x) = \sum_{i \in S} \alpha_i y_i H(x_i, x) + b \quad (7)$$

sendo o coeficiente linear *b* dado por:

$$b = \frac{1}{|U|} \sum_{j \in U} (y_j - \sum_{i \in S} (\alpha_i y_i H(x_i, x_j)) \quad (8)$$

e *U* representa o sub-conjunto composto pelos *support vectors* denominados de *unbounded*, isto é, aqueles para os quais ($0 \leq \alpha_i \leq C$).

A forma da função discriminante depende do *kernel* adotado. Um exemplo comum de *kernel* é a Função Base Radial (RBF), dado por:

$$H(x, x') = \exp(-\gamma \| x - x' \|^2) \quad (9)$$

onde γ é um parâmetro positivo para controle.

A regra de classificação é, então:

$D(x)>0 \quad x_i \in \omega_1$ (10)
$D(x)<0 \quad x_i \in \omega_2$

Se *D(x)=0*, então ***x*** está sobre o hiperplano separador e não é classificado. Quando as amostras de treinamento são linearmente separáveis, a região {*x | 1>D(x)>-1*} é a região de generalização.

Pode-se mostrar que SVM apresenta vantagens com respeito a classificadores convencionais, especialmente quando o número de amostras de treinamento é pequeno e a dimensionalidade dos dados é grande, devido ao fato de que os classificadores convencionais não têm mecanismos para maximizar a margem (distância entre os dois hiperplanos extremos). A maximização da margem permite aumentar a capacidade de generalização do classificador (Abe, 2005).

Finalmente, deve-se mencionar que o classificador SVM só pode ser utilizado na separação de um par de classes a cada vez. Dados de sensoriamento remoto de cenas naturais envolvem a presença de um número maior de classes. Desta forma, aplicações de técnicas SVM na classificação de imagens de sensoriamento remoto requerem abordagens adequadas. Neste estudo é proposto o emprego da função de decisão SVM (Equação 7) em um classificador em estágio múltiplo, tratando um par de classes de cada vez.

2. Materiais e Métodos

Nesta seção são descritos os experimentos realizados implementando a função de decisão SVM em um classificador em estágio múltiplo estruturado como uma árvore binária. São empregados nestes experimentos dados em alta dimensionalidade (hiperespectrais) coletados pelo sistema sensor AVIRIS (*Airbone Visible Infrared Imaging Spectrometer*) sobre uma área teste denominada de *Indian Pines*, localizada no noroeste do Estado de Indiana, USA. O sistema sensor AVIRIS coleta dados em 224 bandas espectrais, no intervalo (0.4 – 2.5μm) do espectro eletromagnético (Landgrebe, 2003). Deste conjunto foram removidas bandas ruidosas (vapor de água na atmosfera), restando 190 bandas. Esta cena compreende culturas de soja e milho, empregando técnicas de cultivo distintas (cultivo tradicional, cultivo direto e cultivo mínimo), além áreas de pastagem e florestais.

Como as escalas das bandas da cena em questão são muito diferentes, decidiu-se padronizar estes dados de acordo com a Equação (11), em Johnson e Wichern (1982).

$$Z = (V^{1/2})^{-1}(X - \mu) \tag{11}$$

onde μ é o vetor de médias, X é o espaço original, Z é o espaço normalizado e $V^{1/2}$ é dado por:

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix} \tag{12}$$

Das classes disponíveis na cena foram selecionadas sete classes levando-se em consideração o número maior de amostras disponíveis (Ver Tabela 1). Dentre elas estão cinco que apresentam alta semelhança espectral e, portanto, de difícil discriminação sem o uso de sensores hiperespectrais. Os algoritmos para os experimentos realizados foram programados em MATLAB.

Tabela 1. Relação das classes utilizadas.

Classes	Amostras disponíveis
ω_1 – milho cultivo mínimo	834
ω_2 – milho plantio direto	1434
ω_3 – pastagem e árvores	747
ω_4 – soja cultivo convencional	614
ω_5 – soja cultivo mínimo	2468
ω_6 – soja plantio direto	968
ω_7 – floresta	997

Conforme mencionado na seção anterior, SVM como classificador, considera um par de classes a cada vez. Nas aplicações a problemas envolvendo mais de duas classes, deve-se adotar metodologias adequadas, tais como:

(a) Formulando diretamente SVM como um problema de otimização multi-classe. Por causa do número de classes que devem ser discriminadas simultaneamente, o número de parâmetros a serem estimados cresce consideravelmente tornando esta abordagem menos estável e, conseqüentemente, afetando a performance da classificação;

(b) Empregando uma estrutura em estágios múltiplos, em uma arquitetura composta por um conjunto de classificadores binários. A decisão é tomada combinando-se as decisões parciais de cada membro do conjunto.

O classificador proposto, desenvolvido em forma de árvore binária, leva em consideração esta segunda abordagem. Para o treinamento do classificador, em cada nó da árvore, aplica-se

o algoritmo que pode ser visto na Figura 3a. Como mostra o fluxograma em questão, as amostras de treinamento são, primeiramente, atribuídas ao nó raiz. Em seguida, supondo-se que os dados sejam normalmente distribuídos, escolhe-se as duas classes que originarão os nós filhos pelo critério distância de Bhattacharyya, dado em Duda et al (2000):

$$B = \frac{1}{8}(\mu_1 - \mu_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \ln \left(\frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}} \right) \quad (13)$$

onde μ_1 e μ_2 são os vetores de médias das classes ω_1 e ω_2 respectivamente, e Σ_1 e Σ_2 as matrizes de covariância.

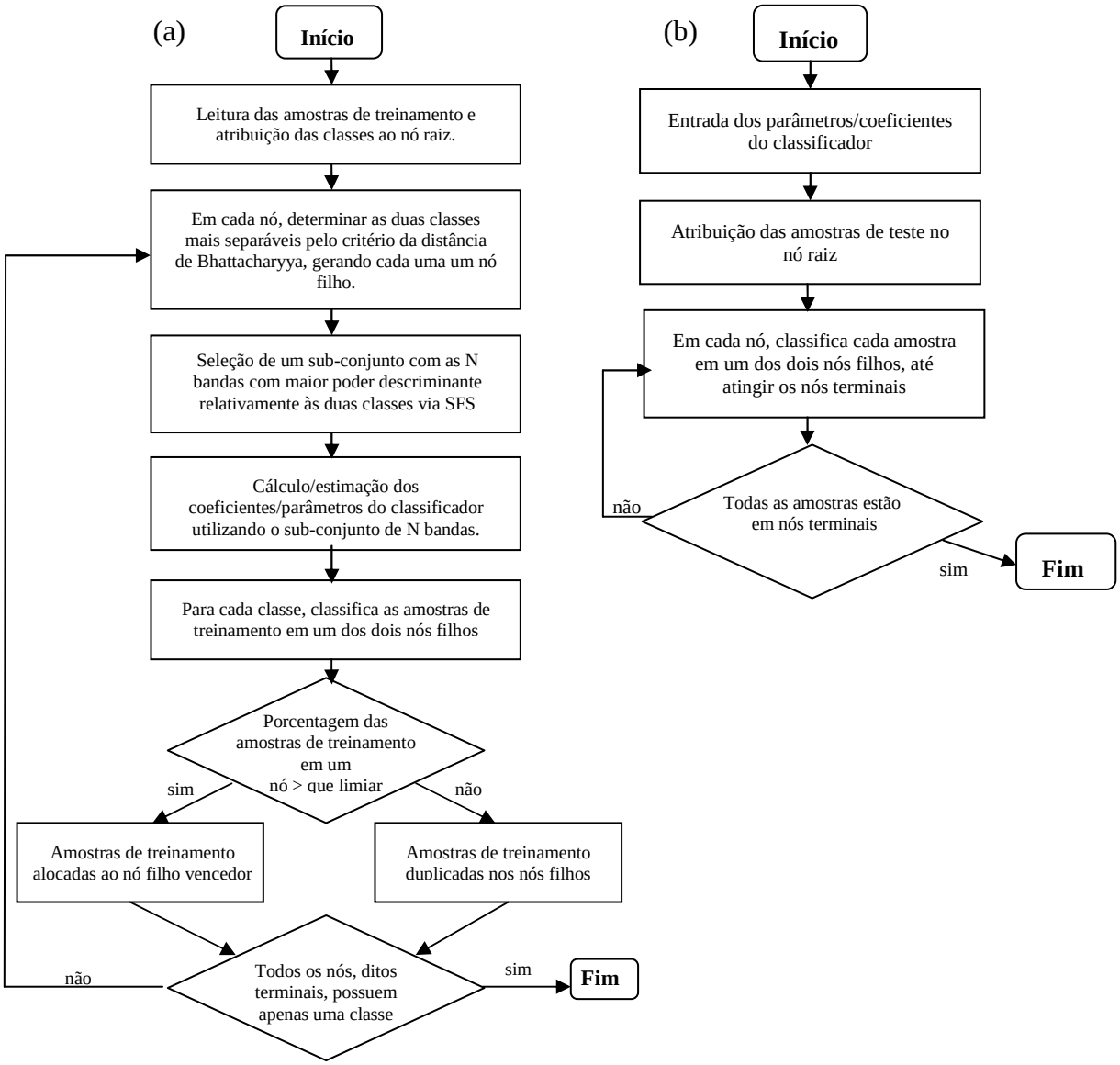


Figura 3: (a) Fluxograma do algoritmo de treinamento do classificador; (b) Fluxograma do algoritmo de teste do classificador.

O uso do algoritmo SFS (*Sequential Forward Selection*) tem por objetivo selecionar, em cada nó, o sub-conjunto das N bandas com maior poder discriminante (Serpico *et al.*, 2003). Estas serão usadas para o cálculo dos coeficientes no caso do uso do classificador SVM ou para a estimação dos parâmetros no caso do uso do classificador Máxima Verossimilhança Gaussiana (MVG) – cujas acurácias serão comparadas. Utilizando-se as respectivas funções

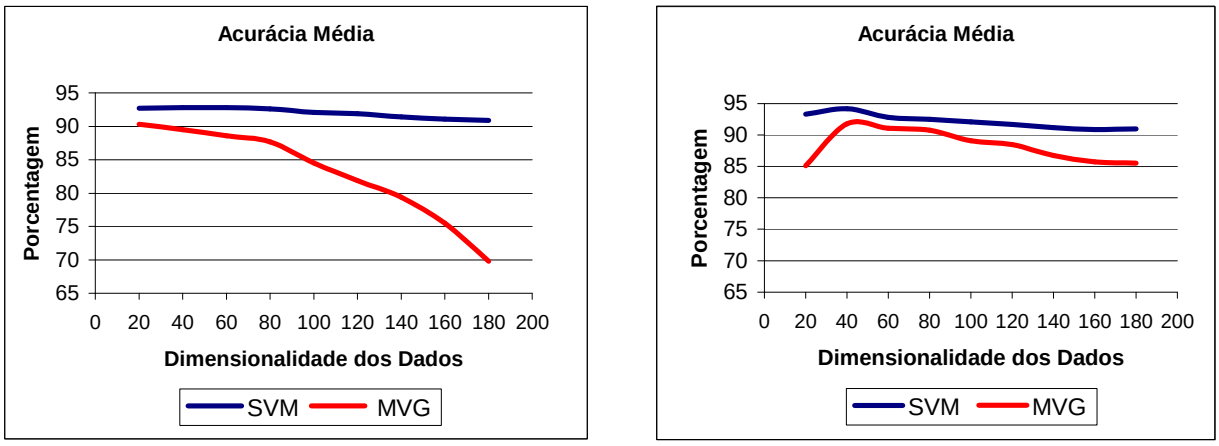
de decisão, classifica-se as amostras de treinamento das demais classes em um dos dois nós filhos. Caso a porcentagem das amostras de treinamento de uma dada classe classificada em um dos nós filhos seja maior que limiar, todas as amostras serão atribuídas a este nó filho. Caso contrário, as amostras de treinamento desta classe são copiadas em ambos os nós filhos. Esse processo será repetido até que cada nó contenha apenas uma classe.

Na Figura 3b, pode-se observar o fluxograma do algoritmo de teste do classificador. Entra-se com as amostras de teste no nó raiz. Com base nos parâmetros estimados (caso do classificador MVG) ou nos coeficientes calculados (caso do classificador SVM) na fase de treinamento, em cada nó decide-se em qual nó filho a amostra de teste será classificada.

Este processo é repetido para cada amostra, ao longo dos vários níveis na árvore binária, até que um nó terminal seja atingido, atribuindo desta forma um rótulo a cada uma das amostras.

3. Resultados e Discussões

Os resultados obtidos com o classificador SVM foram comparados com aqueles obtidos pelo classificador paramétrico Máxima Verossimilhança Gaussiana (MVG), ambos implementados em árvore binária, como citado na metodologia. Para exemplificar a diferença existente entre os resultados produzidos pelos dois classificadores, elaborou-se dois estudos, que serão avaliados segundo a acurácia média. Primeiramente, dividiu-se o conjunto das amostras disponíveis em dois sub-conjuntos, um para fins de treinamento (200 amostras) e o outro para fins de teste (400 amostras). Nos experimentos a acurácia média produzida por cada um dos dois classificadores foi estimada variando a dimensionalidade dos dados de 20 à 180 bandas (com passo de 20), empregando-se o método SFS. O limiar adotado na árvore binária foi de 99%. Usou-se nesses experimentos, para o classificador SVM, o *kernel* RBF (Equação 9), com γ igual à 0.6 e C (cujo valor influencia nas Equações 6 e 8) igual a 10 (Figura 4a). Em um segundo experimento, foram utilizadas 300 amostras para treinamento e outras 300 para fins de teste, mantendo-se o mesmo critério para variação na dimensionalidade dos dados e para os valores dos parâmetros. (Figura 4b).



(a) (b)

Figura 4: (a) Acurácia média para os classificadores SVM e MVG para 200 amostras de treinamento e 400 amostras de teste. (b) Acurácia média para os classificadores SVM e MVG para 300 amostras de treinamento e 300 amostras de teste.

A análise dos resultados dos dois experimentos (Figura 4) evidencia a principal vantagem do classificador SVM (não-paramétrico) em relação ao classificador MVG (paramétrico). No caso de um classificador paramétrico, um número crescente bandas espectrais resultam em um número também crescente de parâmetros a serem estimados. Neste caso, se o número de

amostras de treinamento permanece reduzido, os valores estimados para os parâmetros tornam-se pouco confiáveis, causando uma degradação na performance do classificador. Esta condição resulta no conhecido "Fenômeno de Hughes", claramente visível, especialmente no experimento envolvendo um número menor de amostras de treinamento (Figura 4a). Estes experimentos servem, portanto, para ilustrar a adequação de um classificador implementando a abordagem SVM na classificação de dados em alta dimensionalidade (imagens hiperespectrais), quando o número de amostras disponíveis é limitado, fato este que freqüentemente ocorre em situações reais.

4. Conclusões

Neste trabalho são relatados resultados iniciais de um estudo que está sendo desenvolvido envolvendo métodos de classificação de dados em alta dimensionalidade (imagens hiperespectrais). A abordagem utilizada pelo método SVM apresenta um grande potencial para classificação destes dados nas condições freqüentemente encontradas na prática, quando se dispõe de um número apenas limitado de amostras de treinamento, geralmente insuficiente para uma estimativa confiável dos parâmetros em uma abordagem paramétrica. Um problema potencial que surge na utilização do classificador SVM reside no fato de que este processo é aplicável à somente duas classes a cada vez. Para resolver esta limitação, neste estudo é investigado a aplicação da função SVM em um classificador em estágio múltiplo com arquitetura em árvore binária. Os resultados preliminares mostram-se altamente promissores.

Testes mais aprofundados com outros *kernels* e outros parâmetros serão realizados para que se tenha resultados mais bem fundamentados.

Agradecimentos

Agradecimento especial ao pesquisador Elad Yom-Tov (IBM Haifa Research Laboratory, Israel) pela ajuda nos primeiros passos da implementação do algoritmo SVM.

Referências Bibliográficas

Abe, S.; **Support Vector Machines for Pattern Classifications.** Kobe, Japão: Ed. Springer, 2005.

Duda, O.R.; Hart, P.E.; Stork, D.G.; **Pattern Classification.** Sec. Edition. New York: Wiley-Interscience, 2000.

Fukunaga, K.; **Introduction to Statistical Pattern Recognition.** Sec. Edition. Academic Press: 1990.

Huang, C.; Davis, L.S.; Townshend, J.R.G; An Assessment of Support Vector Machines for Land Cover Classification. **International Journal of Remote Sensing,** vol. 23, nº 4, 2002.

Johnson, R. A.; Wichern, D. W. **Applied Multivariate Statistical Analysis**. New Jersey, USA: Prentice-Hall, 1982.

Melgani, F.; Bruzzone, L.; Classification of Hyperspectral Remote Sensing Images with Support Vector Machines, **IEEE Transactions on Geoscience and Remote Sensing**, Vol42, nº8, agosto de 2004.

Netto, A. V.; Sistema Baseado em Inteligência Computacional para Entendimento de Imagens Oftalmológicas. Revista eletrônica: **IEEE América Latina.** Vol. 3, Dezembro de 2005. Disponível em <http://www.ewh.ieee.org/reg/9/etrans/vol3issue5Dec.2005/3TLA5_3Netto.pdf>. Acesso em 19 de novembro de 2007.

Semolini, R.; **Support Vector Machines, Inferência Transdutiva e o Problema de Classificação**. Campinas, SP. 2002. Dissertação de Mestrado – Departamento de Engenharia de Computação e Automação Industrial – UNICAMP.

Serpico, S.B.; D'Inca, M.; Melgani, F.; Moser, G.; A comparison of feature reduction techniques for classification of hyperspectral remote-sensing data. Proceedings of SPIE, **Image and Signal Processing of Remote Sensing VIII**, Vol. 4885, 2003.