CLASSIFICAÇÃO DE IMAGENS HIPERESPECTRAIS EMPREGANDO SUPPORT VECTOR MACHINES

Classification of Hyperspectral Images with Support Vector Machines

RAFAELA ANDREOLA
VITOR HAERTEL

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.

RESUMO

Neste estudo é investigado o desempenho do classificador Support Vector Machines (SVM) na classificação de imagens em alta dimensionalidade. Como SVM opera em um par de classes a cada vez, propõe-se aqui a sua implementação em uma estrutura em forma de árvore binária, onde somente duas classes são tratadas em cada nó. A acurácia da imagem temática produzida por este esquema de classificação é avaliada para duas funções kernel distintas e em função do valor para dimensionalidade dos dados. Os testes foram realizados empregando imagens hiperespectrais adquiridas pelo sistema sensor AVIRIS. São aqui apresentados e discutidos os resultados obtidos.

Palavras-chave: Support Vector Machines; Classificador em Árvore Binária; Sensoriamento Remoto; Imagens Hiperespectrais.

ABSTRACT

In this study we investigate the performance of the Support Vector Machines (SVM) classifier when applied to the classification of high dimensional remotely sensed image data. As SVM deals with a pair of classes at a time, we propose its implementation in a binary tree approach where two classes only are dealt with at each node. The accuracy of the thematic image produced by this classification scheme was evaluated for two different kernel functions and different data dimensionality. Tests were performed using hiperspectral image data collected by the sensor system AVIRIS. Results are presented and discussed.

Keywords: Hyperspectral Image Data; Support Vector Machines; Binary Tree Classifier; Remote Sensing.

1. INTRODUÇÃO

Dados em alta dimensionalidade (hiperespectrais) podem oferecer um poder discriminante bem mais elevado do que dados tradicionais em baixa dimensionalidade. FUKUNAGA (1990) demonstra que classes espectralmente muito semelhantes entre si (classes que compartilham vetores de médias muito próximos) podem, frequentemente, ser separadas satisfatoriamente em espaços de dimensão mais alta. Esta é uma das motivações para o desenvolvimento de sistemas sensores com um número grande de bandas espectrais, conhecidos como sensores hiperespectrais. Entretanto, uma das principais dificuldades que surgem no processo de classificação de dados em alta dimensionalidade por meio de classificadores paramétricos como, por exemplo, o da Máxima Verossimilhança Gaussiana (MVG), diz respeito ao número geralmente limitado de amostras de treinamento disponíveis, em comparação com o número de parâmetros a serem estimados. Um número limitado de amostras de treinamento resulta em uma estimativa pouco confiável dos parâmetros e, conseqüentemente, em um valor reduzido na acurácia da imagem temática produzida. Este fato pode ser comprovado variando a dimensionalidade dos dados. Iniciando o processo de classificação com dados em dimensionalidade reduzida, a acurácia da imagem temática tende inicialmente a aumentar na medida em que informações adicionais (na forma de bandas espectrais) são incluídas. Em um determinado momento, a acurácia atinge um máximo para em seguida passar a diminuir, na medida em que a dimensionalidade dos dados continua a aumentar. Este problema é conhecido pela comunidade internacional como *fenômeno de Hughes*. Redução na dimensionalidade dos dados por meio de técnicas de extração ou seleção de variáveis – *feature selection/extraction* (WANG, 2008; ZHONG & WANG, 2008), introdução de mostras de treinamento semi-rotuladas (LICZBINSKI & HAERTEL, 2008; JACKSON & LANDGREBE, 2001; SHASHAHANI & LANDGREBE, 1994), técnicas de análise discriminante regularizada (FRIEDMAN, 1989; AEBERHARD *et al.,* 1994; KUO & CHANG, 2007; BERGE et al., 2007), são abordagens que vem sendo investigadas com o objetivo de minimizar as conseqüências de tal fenômeno.

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). O emprego de SVM na classificação de imagens hiperespectrais em sensoriamento remoto vem sendo investigado por alguns autores. MELGANI e BRUZZONE (2004) apresentam resultados obtidos com a aplicação de SVM em imagens hiperespectrais. Em seu estudo os dois autores comparam os resultados obtidos por SVM com os produzidos por outros dois classificadores também não-paramétricos (redes neurais RBF e K-vizinhos mais próximos). No presente estudo, a avaliação é feita comparando os

resultados produzidos por SVM com aqueles produzidos pelo classificador paramétrico mais frequentemente utilizado pela comunidade em sensoriamento remoto (MVG). SHAH *et al.* (2003) apresentam um sumário dos vários trabalhos desenvolvidos por aqueles autores com vistas à classificação de imagens hiperespectrais, incluindo o emprego de SVM com a etapa de otimização implementada via método Lagrangiano. WATANACHATURAPORN *et al.* (2004) relatam uma investigação inicial no emprego de SVM na classificação de imagens hiperespectrais em sensoriamento remoto. Em seu trabalho, aqueles autores investigam os efeitos na acurácia dos resultados causados pelos diferentes métodos de implementar SVM em problemas com múltiplas classes e pelo tipo de kernel utilizado. Outras aplicações empregando técnicas de SVM têm sido investigadas por outros autores como, por exemplo, BROWN *et al.* (2000) e BROWN *et al.* (1999). Nestes trabalhos os autores investigam a utilização de SVM em problemas de mistura espectral, comparando uma abordagem envolvendo SVM com o bem conhecido Modelo Linear de Mistura Espectral. A utilização do classificador SVM apresenta, entretanto, algumas dificuldades. Possivelmente a mais óbvia resida no fato de SVM ser aplicável diretamente a apenas um par de classes a cada vez (ABE, 2005). Na metodologia proposta, investiga-se a implementação de SVM em um classificador em estágio múltiplo estruturado na forma de árvore binária. Uma vantagem adicional desta estrutura reside na possibilidade de otimização na escolha das variáveis ou feições (*features*) que conferem um maior poder discriminante entre o par de classes a cada nó individual da árvore binária.

2. SUPPORT VECTOR MACHINES (SVM)

SVM é um classificador linear no qual busca-se minimizar o erro com relação ao conjunto das amostras de treinamento (risco empírico) e o erro com relação ao conjunto das amostras de teste (risco na generalização). O objetivo de SVM consiste em obter o equilíbrio entre esses erros, minimizando o excesso de ajustes com respeito às amostras de treinamento (*overfitting*) e aumentando, conseqüentemente, a capacidade de generalização do classificador (VAPNIK, 1999). O problema denominado de *overfitting* consiste em o classificador memorizar os padrões de treinamento, gravando suas peculiaridades e ruídos, ao invés de extrair as características gerais que permitirão a generalização ou reconhecimento de padrões não utilizados no treinamento do classificador (SMOLA et al., 2000).

A questão da generalização pode ser mais bem avaliada para o caso de duas classes. Assumindo que as amostras de treinamento das duas 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). Este princípio é implementado em SVM e a correspondente formulação matemática dada a seguir está baseada em ABE (2005).

Seja um conjunto com M amostras de treinamento $\mathbf{x}_i$ (i=1, ..., M) em um problema que consiste de duas classes linearmente separáveis (ω_1 *e* ω_2). Cada amostra fica associada a um rótulo: y_i=*1* se $\mathbf{x}_i \in \omega_1$, y_i = *-1* se $\mathbf{x}_i \in \omega_2$. A forma geral de uma função linear de decisão é dada por:

$$D(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{1}$$

onde $\mathbf{x}$ é um vetor m-dimensional representando o padrão a ser classificado, $\mathbf{w}$ também é um vetor m-dimensional (pesos) e b o termo independente. Como estamos supondo amostras linearmente separáveis, não ocorrerá a situação em que $\mathbf{w}\mathbf{x}_i + b = 0$. Desta forma, o critério para classificação pode ser escrito como:

$$\begin{aligned} \mathbf{w}\mathbf{x}_i + b > a \quad &\text{para } \mathbf{x}_i \in \omega_1 \ (y_i = 1) \\ \mathbf{w}\mathbf{x}_i + b < -a \quad &\text{para } \mathbf{x}_i \in \omega_2 \ (y_i = -1) \end{aligned} \tag{2}$$

para a>0. Dividindo ambos os membros da desigualdade por a, o critério para classificação fica:

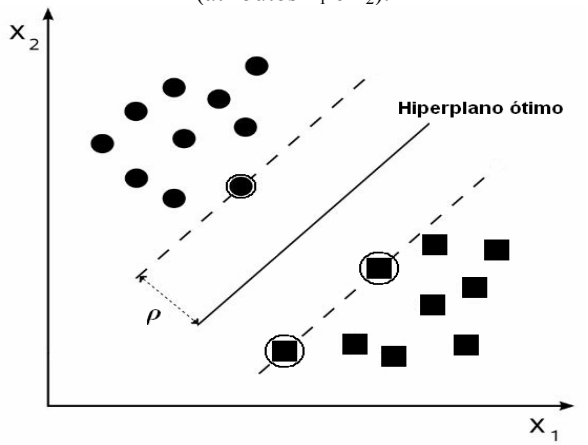
$$\mathbf{w}^T\mathbf{x}_i + b \begin{cases} \geq 1 & \text{para } y_i = 1 \\ \leq -1 & \text{para } y_i = -1 \end{cases} \tag{3}$$

Deste modo, ambas as condições podem ser combinadas em uma única:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \text{para } i=1, 2, ..., M \tag{4}$$

sendo M o número de amostras disponíveis.

Figura 1 - O hiperplano ótimo separando os dados com a máxima margem ρ. Os *support vectors* (amostras circuladas) e uma distribuição dos dados no R^2 (atributos x_1 e x_2).



Fonte: Adaptado de ABE (2005).

O hiperplano:

$$D(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = c \text{ para } -1 < c < 1 \quad (5)$$

forma, então, uma superfície de separação entre as duas classes. Para $c = 0$, a Equação (5) define um hiperplano situado à meia distância entre os dois hiperplanos extremos ($c=1$ e $c=-1$). A distância entre estes dois hiperplanos extremos é denominada de "margem", representada por ρ na Figura 1. Supondo a existência de pelo menos uma amostra $\mathbf{x}$ para a qual $D(\boldsymbol{x}) = 1$, e pelo menos uma outra amostra para a qual $D(\mathbf{x}) = -1$, então o hiperplano $D(\mathbf{x}) = 0$ representa a melhor superfície de separação entre estas amostras, no sentido de que maximiza o poder de generalização do classificador. A região entre os dois hiperplanos extremos ($-1 \leq D(\mathbf{x}) \leq 1$) pode ser entendida como a região de generalização. O hiperplano $D(\mathbf{x})=0$, ao maximizar o valor da margem, maximiza a região de generalização sendo, portanto, neste sentido ótimo (Figura 1). A distância $\mathrm{d}(\mathbf{x})$ de uma amostra $\mathbf{x}$ a um plano qualquer $D(\mathbf{x})$ é dada por:

$$\mathrm{d}(\mathbf{x}) = | D(\mathbf{x}) | / || \mathbf{w} || \quad (6)$$

O hiperplano ótimo será, portanto, aquele para o qual esta distância é máxima. Esta condição pode ser obtida minimizando-se $||\mathbf{w}||$, ou equivalentemente, minimizando:

$$Q(\mathbf{w}) = \frac{1}{2} || \mathbf{w} ||^2 = \frac{1}{2}\mathbf{w}^T\mathbf{w} \quad (7)$$

com respeito aos parâmetros da função $D(\mathbf{x})$, $\mathbf{w}$ e b. Para satisfazer a convenção adotada com relação ao rótulo de cada amostra (y_i), a restrição da Equação (4) deve ser imposta. Tal restrição é imposta de maneira a assegurar que não ocorram amostras de treinamento na região de separação entre as duas classes (entre as margens). A inclusão das restrições (4) no problema de minimização da Equação (7) pode ser resolvido por meio da técnica dos multiplicadores de Lagrange ($\boldsymbol{\alpha}$). Esta abordagem pode ser expressa por minimizar

$$Q(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{M} \alpha_i \left\{ y_i \left(\mathbf{w}^T\mathbf{x}_i + b \right) - 1 \right\} \quad (8)$$

com relação a $\mathbf{w}$, b e maximizar com relação a α_i, sendo $\boldsymbol{\alpha}=(\alpha_1, \ldots, \alpha_M)$ os multiplicadores de Lagrange, um vetor de dimensão M, com $\alpha_i \geq 0$. Deste modo, obtém-se a forma dual, expressa em termos de $\boldsymbol{\alpha}$ somente (HAMEL, 2009; ABE, 2005):

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \quad (9)$$

A solução deste problema permite expressar **w** em termos de **α** resultando em uma nova forma para a função de decisão (1)**:**

$$D(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b \tag{10}$$

onde *S* é o conjunto de índices dos *support vectors*, isto é, as amostras de treinamento para as quais $\alpha_i > 0$.

A formulação acima apresenta solução somente no caso de as amostras $\mathbf{x}_i$ pertencentes às duas classes serem linearmente separáveis. Em situações reais, entretanto, os dados frequentemente não são linearmente separáveis. Este fato ocorre com frequência em imagens multiespectrais cobrindo cenas naturais, como aquelas empregadas em sensoriamento remoto (Landsat-TM e SPOT, entre outros sistemas sensores), nas quais as diferentes classes de cobertura do solo muitas vezes não são linearmente separáveis. Para estender a formulação acima para conjuntos de dados não linearmente separáveis, permite-se que alguns dados possam violar a restrição da Equação (4), por meio da introdução do conceito de variável de folga (*slack variable*) representada por ξ_i ($\xi_i \geq 0$). Tais variáveis relaxam as restrições impostas ao problema de otimização. Neste caso, a restrição expressa na Equação (4) torna-se:

$$y_i (\mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b) \geq 1 - \xi_i \tag{11}$$

Esta abordagem é conhecida como SVMs com margens suaves (*soft-margin*) (HAMEL, 2009; LORENA & CARVALHO, 2007; ABE, 2005).

Para o caso de $0 < \xi_i < 1$ a correspondente amostra $\mathbf{x}_i$ não terá margem máxima, mas será rotulada corretamente. No caso de $\xi_i \geq 1$, a amostra $\mathbf{x}_i$ será rotulada erroneamente. Para levar em consideração o termo ξ_i, minimizando assim o erro sobre os dados de treinamento, a Equação (7) é reformulada como:

$$Q(\mathbf{w}, b, \boldsymbol{\xi}) = \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{M} \xi_i \tag{12}$$

A constante *C*, conhecida como "parâmetro de margem", estabelece o equilíbrio entre a maximização da margem e a minimização dos erros.

O procedimento, neste caso, é semelhante ao desenvolvido para o caso de *margens rígidas*, resultando em uma função de decisão semelhante à anterior (Equação 10), com a única diferença de $C \geq \alpha_i \geq 0$ para *i=1, ..., M* (HAMEL, 2009; ABE, 2005).

As SVMs lineares são eficazes na classificação de conjuntos de dados linearmente separáveis, contaminados com a presença de alguns ruídos e *outliers*. Entretanto, em situações reais ocorre com bastante freqüência classes não linearmente separáveis. A solução mais simples nestes casos consistiria na adoção

de polinômios de grau mais elevado. Entretanto, esta abordagem apresenta o risco de excesso de ajuste (*overfitting*), e a conseqüente redução no poder de generalização do classificador (DUDA et al., 2000). Uma opção mais eficiente consiste em mapear os dados para um espaço de dimensão mais alta, no qual os dados tornam-se linearmente separáveis (HAMEL, 2009; CRISTIANINI & SHAWE-TAYLOR, 2000).

Na abordagem apresentada a seguir, as M amostras no espaço original ($\mathbf{x}_1$, $\mathbf{x}_2$, . . . , $\mathbf{x}_M$), com dimensão m são mapeadas no novo espaço (espaço característico) por meio de uma função $\mathbf{g}$ de dimensão $n>m$: $\mathbf{g}=(g_1, g_2, \cdots g_n)$. Neste novo espaço as M amostras $\mathbf{x}_i$ (dimensão m) são mapeadas em M amostras com dimensão n:

$$\begin{bmatrix} g_1(\mathbf{x}_1) \\ g_2(\mathbf{x}_1) \\ \vdots \\ g_n(\mathbf{x}_1) \end{bmatrix}, \begin{bmatrix} g_1(\mathbf{x}_2) \\ g_2(\mathbf{x}_2) \\ \vdots \\ g_n(\mathbf{x}_2) \end{bmatrix}, \ldots, \begin{bmatrix} g_1(\mathbf{x}_M) \\ g_2(\mathbf{x}_M) \\ \vdots \\ g_n(\mathbf{x}_M) \end{bmatrix}$$

O processo consta então de dois passos:

1- uma função não linear $\mathbf{g}$ mapeia os dados do espaço original para um novo espaço de dimensão mais alta;

2- a classificação é feita neste novo espaço empregando uma função de decisão linear.

A forma geral da função de decisão no espaço original é dada pela Equação (1). Neste novo espaço, a função linear de decisão fica:

$$D(\mathbf{x}) = \mathbf{w} \,.\, \mathbf{g}(\mathbf{x}) + b \tag{13}$$

e a Equação (9) torna-se, portanto:

$$Q(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \, \alpha_j \, y_i \, y_j H(\mathbf{x}_i, \mathbf{x}_j) \tag{14}$$

onde $H(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{g}(\mathbf{x}_i)^{\mathrm{T}} . \mathbf{g}(\mathbf{x}_j)$, que recebe a denominação de kernel. A condição necessária para que uma função H seja um kernel é conhecida como condição de Mercer:

$$\sum_{i,j=1}^{M} h_i h_j H(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \tag{15}$$

para todo M, $\mathbf{x}_i$ e h_i, onde h_i é um número real (ABE, 2005).

Neste novo espaço, a função de decisão expressa em termos de $\boldsymbol{\alpha}$ - Equação (10) - torna-se:

$$D(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i \mathbf{g}(\mathbf{x}_i)\mathbf{g}(\mathbf{x}) + b \tag{16}$$

ou alternativamente expressa em termos de kernel:

$$D(\mathbf{x}) = \sum_{i \in S} \alpha_i \; y_i \; H(\mathbf{x}_i, \mathbf{x}) + b \tag{17}$$

com *b* dado por:

$$b = \frac{1}{U} \sum_{j \in U} \left(y_j - \sum_{i \in S} \alpha_i \; y_i \; H(\mathbf{x}_i \; \mathbf{x}_j) \right) \tag{18}$$

sendo *U* o conjunto dos *support vectors* denominados de *unbounded,* isto é, aqueles para os quais $0<\alpha_i<C$.

Existem, portanto, duas possíveis abordagens ao problema do mapeamento de dados em espaços de dimensão mais elevada para fins de classificação empregando funções de decisão lineares **g**(**x**) (HAMEL, 2009; HERBRICH, 2002):

1- Selecione explicitamente uma função **g** para mapeamento dos dados em um espaço de dimensão mais alta.

2- Selecione diretamente um kernel *H* que satisfaça as condições de Mercer. Este kernel vai definir de uma forma implícita a função de mapeamento **g.**

Do ponto de vista matemático, as duas possíveis abordagens citadas acima são equivalentes. A segunda abordagem (escolha direta de um kernel) apresenta, entretanto, a vantagem de ser mais fácil de implementar e de ser interpretada. Outra vantagem oferecida por esta abordagem consiste em não se necessitar operar diretamente no espaço em dimensão mais alta, no qual os dados estão sendo mapeados. Tanto a fase de treinamento do classificador quanto a fase de classificação dos dados utiliza-se diretamente $H(\mathbf{x}_i, \mathbf{x})$ em lugar da função de mapeamento **g**(**x**). Exemplos comuns de kernel são a *Radial Basis Function* (RBF) (Equação 19) e o kernel Polinomial (Equação 20):

$$H(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma \, || \, \mathbf{x} - \mathbf{x}_i \, ||^2) \tag{19}$$

onde γ é um parâmetro positivo para controle

$$H(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T \mathbf{x}_i + 1)^d \tag{20}$$

e *d* é um número natural e determina o grau do polinômio.

A regra de classificação é dada por:

$$\begin{aligned} D(\mathbf{x}_i) > 0 \quad & \mathbf{x}_i \in \omega_1 \\ D(\mathbf{x}_i) < 0 \quad & \mathbf{x}_i \in \omega_2 \end{aligned} \tag{21}$$

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

3. MATERIAIS E MÉTODOS

3.1 Materiais

Nestes experimentos são empregados dados em alta dimensionalidade (hiperespectrais) coletados pelo sistema sensor AVIRIS sobre uma área agrícola de testes, desenvolvida pela Purdue University, e denominada de *Indian Pines*, localizada no noroeste do Estado de Indiana, EUA, sob a denominação de 92AV220. Da cena 92av220, foi selecionado de um segmento de imagem de (435x435) um recorte de (145x118), num total de 17110 pixels. Esta área dispõe de dados de verdade terrestre.

O que torna a área atraente para os estudos que empregam dados em alta dimensionalidade é esta possuir classes com características espectrais muito semelhantes entre si e, portanto, difíceis de serem separados por meio de dados tradicionais em baixa dimensionalidade como, por exemplo, dados Landsat-TM. Do conjunto de 220 bandas que dispõe a cena AVIRIS (cobre a região $0.4\mu m$ à $2.4\mu m$ do espectro eletromagnético, com resolução espectral de $10nm$), foram removidas as bandas ruidosas causados por problemas atmosféricos (vapor de água, CO_2, O_3). A dimensionalidade final dos dados utilizados é de 190 bandas. A área selecionada apresenta 10 classes de cobertura do solo. Para realizar os experimentos foram selecionadas seis classes que apresentam a maior dificuldade de separação (Ver Tabela 1).

Tabela 1 - Relação das classes usadas nos experimentos.

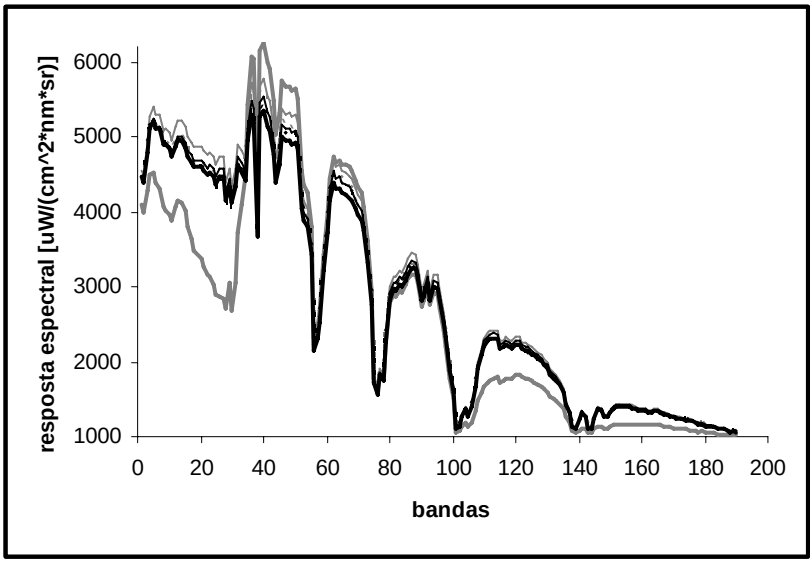
Classes	Descrição	Amostras Disponíveis
ω_1 – corn min	milho cultivo mínimo	834
ω_2 – corn notill	milho plantio direto	1434
ω_3 – grass trees	pastagens e árvores	747
ω_4 – soybean clean	soja cultivo convencional	614
ω_5 – soybean min	soja cultivo mínimo	2468
ω_6 – soybean notill	soja plantio direto	968

A imagem foi obtida no início da época de crescimento das culturas de soja e milho. Nesta etapa apenas aproximadamente 5% da área está efetivamente coberta pela vegetação, sendo os restantes 95% composto por solo exposto e resíduo de colheitas anteriores. Estas condições resultam em classes espectralmente muito

semelhantes (vetores de média muito semelhantes entre si), constituindo-se por esta razão em um desafio ao processo de classificação. A classe pastagens/árvores (*grass trees*) foi incluída por possuir características espectrais bem diferentes das demais sendo, portanto, facilmente separável das demais classes, servindo de referência no processo de classificação.

A Figura 2 ilustra o comportamento espectral médio das classes da Tabela 1, onde se verificam dois aspectos principais: a diferença espectral da classe pastagens/árvores com relação às demais classes, e a alta semelhança entre as outras cinco classes (variações das culturas de *milho* e *soja*).

Figura 2 - Curvas de resposta espectral média para cada uma das classes: milho cultivo mínimo (cinza pontilhada), milho plantio direto (cinza contínua fina), pastagens/árvores (cinza continua grossa), soja cultivo mínimo (preta contínua fina), soja plantio direto (preta contínua grossa), soja cultivo convencional (preta pontilhada).



Como nos dados utilizados o intervalo numérico de variação dos contadores digitais ao longo do conjunto das bandas espectrais é muito grande, decidiu-se padronizar estes dados (equações 22 e 23) para média igual a zero e desvio padrão igual a um (JOHNSON E WICHERN ,1982):

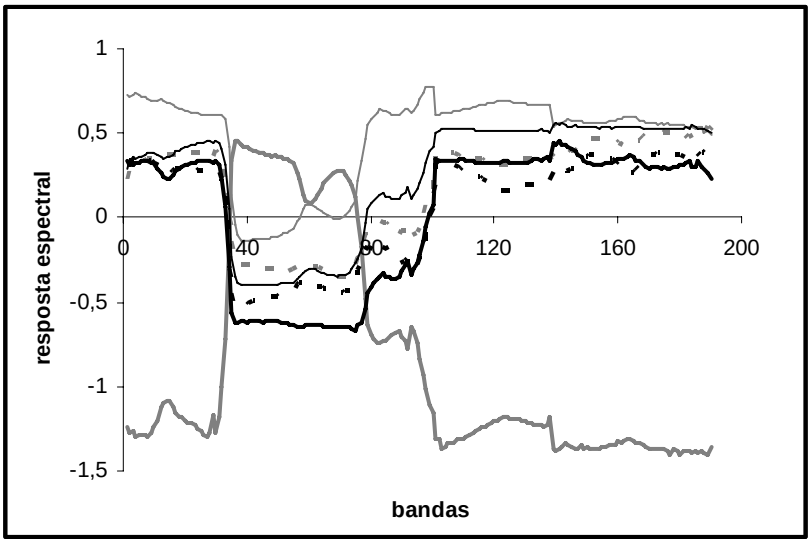
$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}) \tag{22}$$

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

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix} \tag{23}$$

O resultado deste processo de padronização está ilustrado na Figura 3.

Figura 3 - Curvas de resposta espectral média para as classes após a padronização: milho cultivo mínimo (cinza pontilhada), milho plantio direto (cinza contínua fina), pastagens/árvores (cinza continua grossa), soja cultivo mínimo (preta contínua fina), soja plantio direto (preta contínua grossa), soja cultivo convencional (preta pontilhada).



Do conjunto das amostras disponíveis para cada classe foram extraídos dois subconjuntos: um com amostras de treinamento e um segundo com amostras de teste (método *holdout*). Com a finalidade de capturar as variações naturais que ocorrem ao longo da área coberta pela imagem, as amostras em ambos os subconjuntos foram extraídos alternadamente do conjunto das amostras disponíveis nos dados de verdade terrestre.

Para tornar os resultados obtidos para as várias classes comparáveis entre si, os experimentos empregaram subconjuntos de treinamento e de teste de mesmo tamanho para todas as classes em estudo: 50, 100, 200 e 300 amostras por classe para treinamento e 300 amostras por classe para teste. As amostras de treinamento e teste foram tomadas a intervalos regulares no conjunto total de amostras para cada classe. Desta forma, as amostras de treinamento em um experimento não necessariamente constam no conjunto das amostras de treinamento do experimento seguinte.

3.2 Métodos

A metodologia adotada implementa SVM em uma árvore binária, do tipo *bottom-up,* a fim de possibilitar a utilização de SVM em problemas multi-classe. Os resultados produzidos por este classificador assim proposto foram comparados com aqueles resultantes do classificador MVG - largamente usado na comunidade científica em reconhecimento de padrões. Para o treinamento do classificador, em cada nó da árvore, aplica-se o algoritmo que pode ser visto na Figura 4a. As amostras de treinamento são inicialmente 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:

$$B = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \left(\frac{\left| \frac{(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)}{2} \right|}{|\boldsymbol{\Sigma}_1|^{1/2} |\boldsymbol{\Sigma}_2|^{1/2}} \right) \qquad (24)$$

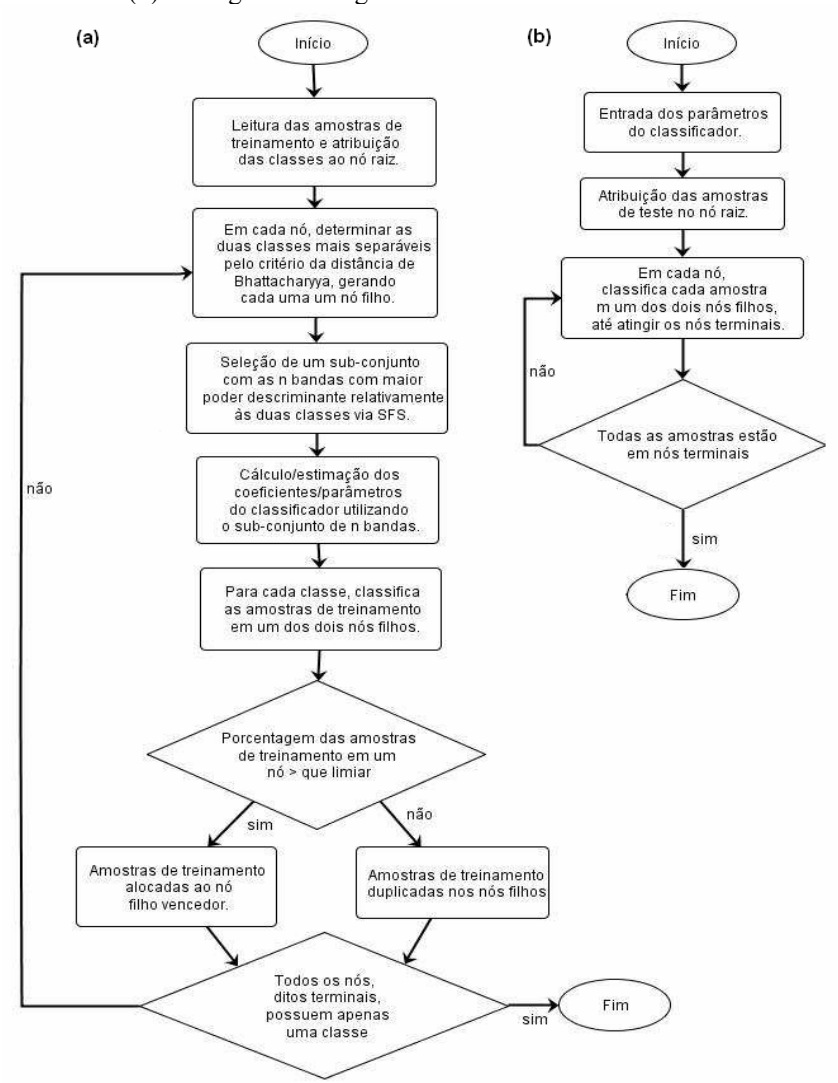
onde $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ são os vetores de médias das classes ω_1 e ω_2 respectivamente, e $\boldsymbol{\Sigma}_1$ e $\boldsymbol{\Sigma}_2$ as matrizes de covariância.

O uso do algoritmo SFS (*Sequential Forward Selection*) tem por objetivo selecionar, em cada nó, o subconjunto das N bandas com maior poder discriminante (SERPICO *et al., 2*003). 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 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 determinado limiar de verossimilhança (LV), todas as amostras serão atribuídas a este nó filho. Caso contrário, as amostras de treinamento desta classe são replicadas em ambos os nós filhos. Esse processo será repetido até que cada nó contenha apenas uma classe (nós terminais).

A Figura 4b ilustra o fluxograma do algoritmo para 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.

Figura 4 – (a) Fluxograma do algoritmo de treinamento do classificador.
(b) Fluxograma do algoritmo de teste do classificador.



Para fins de implementação da metodologia proposta neste estudo, foi desenvolvida uma ferramenta denominada de Classificador em Árvore Binária (CAB). O CAB, implementado em forma de árvore binária, possui duas versões, uma para o classificador MVG e outra para o classificador SVM. Desenvolvidos em ambiente MATLAB 6.1, o CAB-MVG e o CAB-SVM apresentam como resultado a Matriz de Confusão.

4. RESULTADOS E DISCUSSÕES

Os experimentos foram desenvolvidos com o objetivo de quantificar numericamente os resultados de desempenho da metodologia proposta na classificação de imagens digitais de alta dimensionalidade em sensoriamento remoto, utilizando diferentes kernels e parâmetros no classificador SVM implementados pela ferramenta CAB-SVM.

Foi realizada uma série de experimentos, tomando-se a dimensionalidade dos dados como variável independente e a resultante acurácia na classificação como variável dependente. O valor da dimensionalidade dos dados, isto é, o número de bandas espectrais empregadas, variou entre 20 e 180. Em um primeiro conjunto de experimentos as bandas espectrais foram selecionadas por meio do algoritmo SFS, a um intervalo de 20 bandas. Em um segundo conjunto de experimentos a seleção destas bandas foram feitas a intervalos regulares no espectro eletromagnético (sem o uso do SFS), com o intuito de verificar a eficácia do SFS em um classificador não paramétrico como o SVM. O objetivo dos experimentos é analisar o comportamento da acurácia produzida pelo classificador SVM em função da dimensionalidade dos dados e dos parâmetros escolhidos. Os resultados assim obtidos são comparados com aqueles obtidos nas mesmas condições, empregando-se um classificador paramétrico tradicional (MVG), implementado pela ferramenta CAB-MVG. Nota-se que o valor mínimo admissível para as amostras de treinamento no caso do CAB-MVG é igual à dimensionalidade dos dados mais um. Um valor inferior resultará em uma matriz de covariância singular e, portanto, não utilizável (LANDGREBE, 2003).

O número de amostras de treinamento foi escolhido deliberadamente pequeno com relação à dimensionalidade dos dados para desta forma melhor evidenciar os problemas que ocorrem em situações reais, ou seja, o pequeno número de amostras de treinamento normalmente disponíveis. Na realização dos experimentos foram usadas 80 bandas para o cálculo da distância de Bhattacharyya (no caso de 50 amostras de treinamento, todas elas são usadas para o cálculo da distância de Bhattacharyya) e LV de 99%. Decidiu-se fixar o LV em 99% para que fosse obtida sempre a maior estrutura possível, ou seja, o número máximo de nós terminais (MORAES, 2005). Segundo o autor, valores mais altos para o LV produzem, uma menor variabilidade no valor estimado da acurácia de cada classe individual, em função da dimensão dos dados.

Outras grandezas são requeridas pela ferramenta CAB-SVM. Os multiplicadores de Lagrange (Equação 8) foram calculados empregando a função quadprog.m disponível em MATLAB®, enquanto que o parâmetro de margem C (Equação 12) foi tomado igual a um (1). Nos vários experimentos realizados, envolvendo diferentes sub-conjuntos de amostras de treinamento, foram investigados distintos valores para o grau do polinômio (no caso do kernel polinomial) e para gamma (γ), no caso do kernel RBF. As Figuras 5-12 ilustram os resultados produzidos pelo classificador SVM (ferramenta CAB-SVM implementando os kernels polinomial e RBF) juntamente com aqueles produzidos pelo classificador mais tradicional MVG (ferramenta CAB-MVG), para 50, 100, 200 e 300 amostras de treinamento. Nas Figuras 5, 7, 9 e 11 estão ilustrados o resultados dos experimentos empregando bandas selecionadas via SFS e nas Figuras 6, 8, 10 e 12 os resultados com bandas selecionadas sem SFS. Deve-se observar aqui que os experimentos empregando a ferramenta CAB-SVM evidenciaram que a acurácia nos resultados depende dos valores adotados para o grau do polinômio no caso do kernel polinomial e para gamma (γ) no caso do kernel RBF. Para fins de comparação entre os dois classificadores, estas figuras ilustram os melhores resultados obtidos em cada caso.

Figura 5 - Acurácia Média para os classificadores MVG e SVM com kernel Polinomial grau 2 e RBF γ 1.5 para 50 amostras de treinamento com SFS.

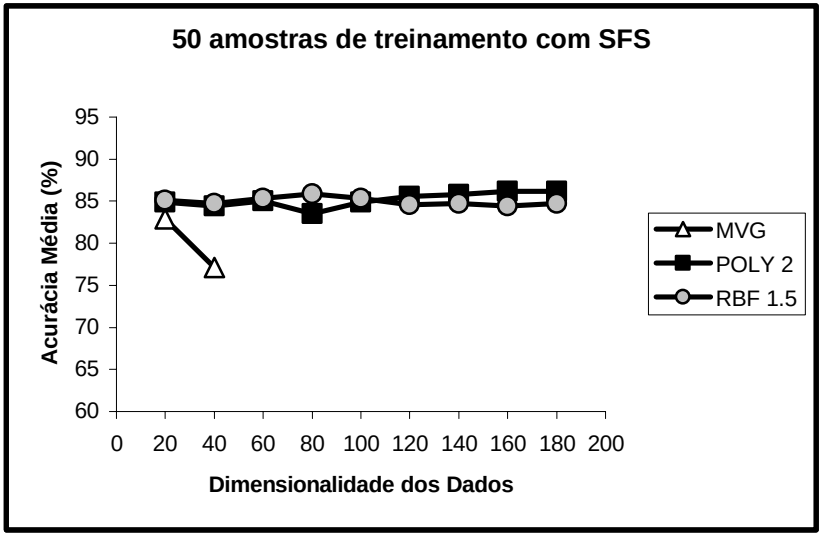


Figura 6 - Acurácia Média para os classificadores MVG e SVM com kernel Polinomial grau 2 e RBF γ 1.5 para 50 amostras de treinamento sem SFS.

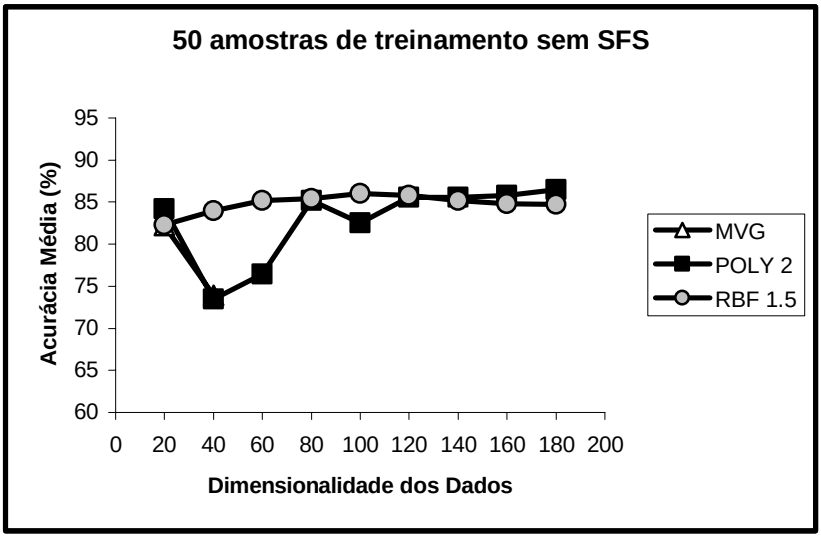


Figura 7 - Acurácia Média para os classificadores MVG e SVM com kernel Polinomial grau 3 e RBF γ 2 para 100 amostras de treinamento com SFS.

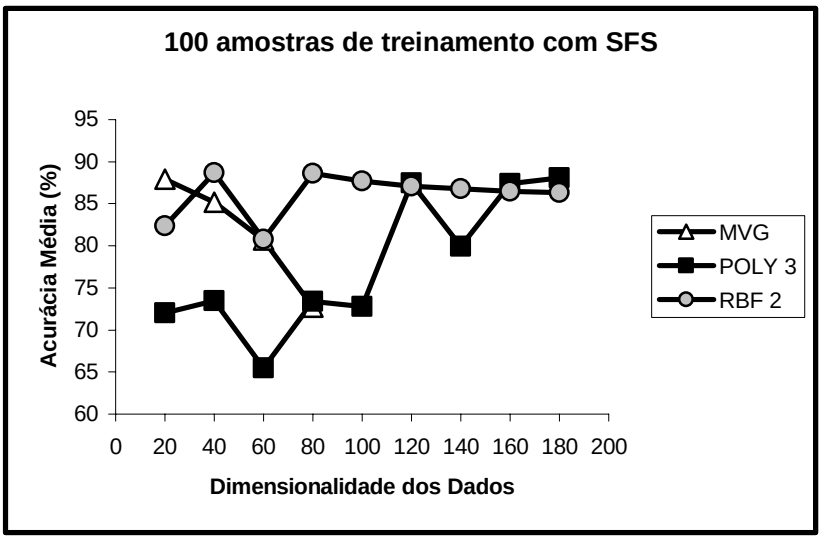
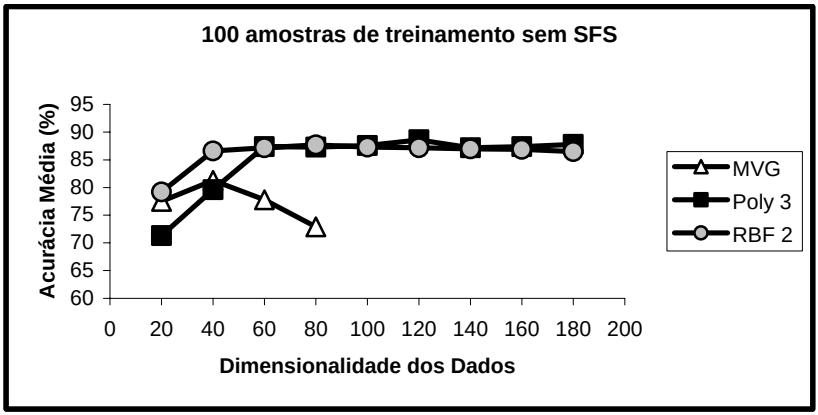


Figura 8 - Acurácia Média para os classificadores MVG e SVM com kernel Polinomial grau 3 e RBF γ 2 para 100 amostras de treinamento sem SFS.



Como se pode perceber pela análise das Figuras 5 à 8, os resultados obtidos para 50 amostras de treinamento são praticamente iguais com e sem o uso de SFS, e para 100 amostras de treinamento, os resultados sem SFS são melhores do que os com uso de SFS nos experimentos realizados com a ferramenta CAB-SVM. O mesmo não acontece para os experimentos realizados com a ferramenta CAB-MVG, onde os resultados se mostram claramente melhores com o uso do SFS, apesar de apresentarem, em ambos os casos, os efeitos do fenômeno de Hughes. Em todos os casos as acurácias médias para o classificador SVM são superiores às acurácias médias utilizando o classificador MVG.

Figura 9 - Acurácia Média para os classificadores MVG e SVM com kernel Polinomial grau 3 e RBF γ 0.5 para 200 amostras de treinamento com SFS.

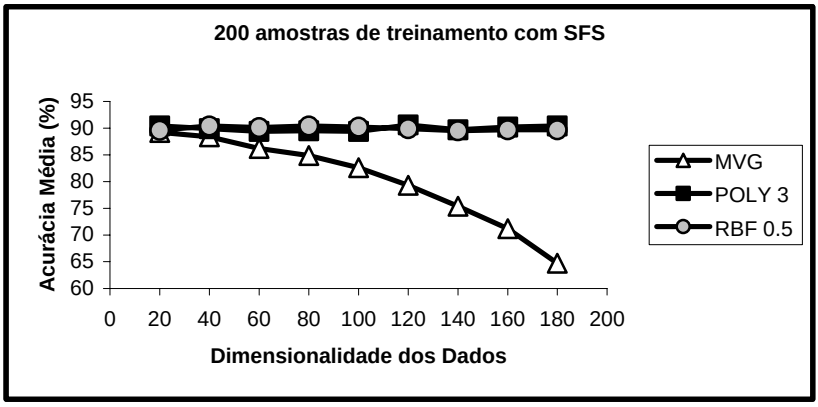


Figura 10 - Acurácia Média para os classificadores MVG e SVM com kernel Polinomial grau 3 e RBF γ 0.5 para 200 amostras de treinamento sem SFS.

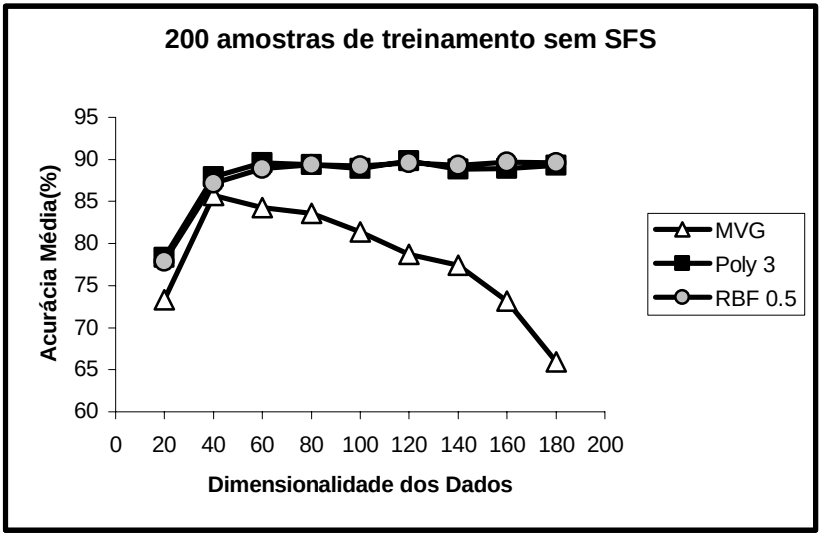


Figura 11 - Acurácia Média para os classificadores MVG e SVM com kernel Polinomial grau 3 e RBF γ 1.5 para 300 amostras de treinamento com SFS.

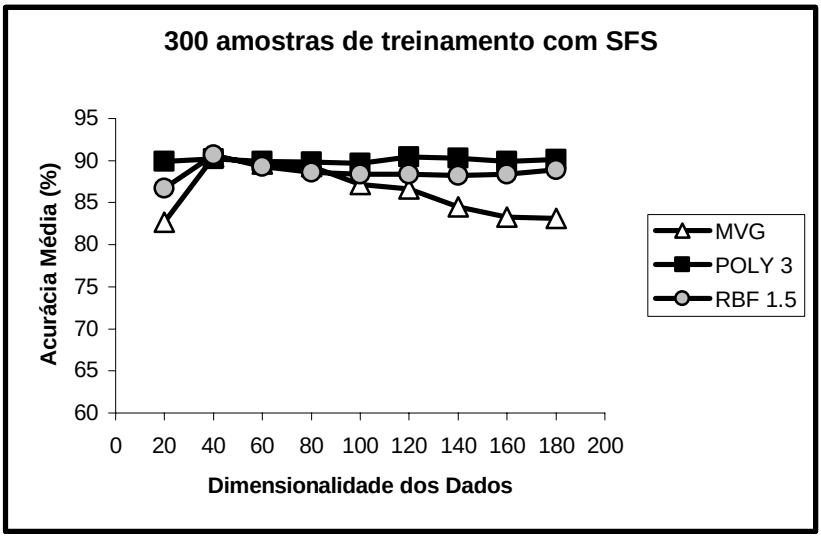
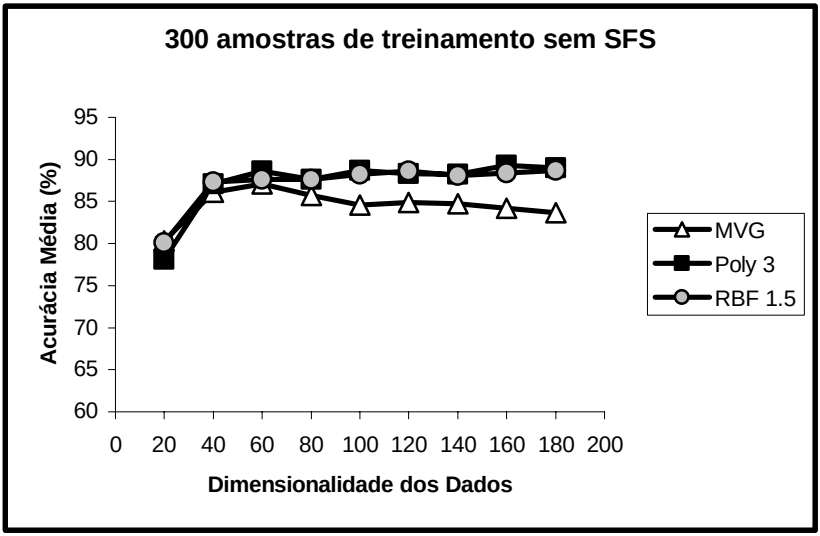


Figura 12 - Acurácia Média para os classificadores MVG e SVM com kernel Polinomial grau 3 e RBF γ 1.5 para 300 amostras de treinamento sem SFS.



Pode-se perceber que, mesmo com o aumento do número de amostras de treinamento de 200 para 300 (Figuras 9 à 12), a acurácia média não se eleva substancialmente (com o uso da ferramenta CAB-SVM – SVM implementado em forma de árvore binária); pelo contrário, na maioria dos casos a acurácia média para 300 amostras de treinamento é levemente menor ou igual que aquela para 200. Isso acontece porque o poder máximo de generalização do classificador é atingido com 200 amostras de treinamento, ou seja, o conjunto de 200 amostras representa bem as características de cada classe, e o incremento para 300 apresenta o risco de aumento no número de amostras ruidosas, para um reduzido acréscimo de informação.

Comparando-se os resultados ilustrados nas Figuras 5 à 12 pode-se perceber que o maior ganho obtido empregando o método SFS baseado em distâncias estatísticas ocorre em classificadores paramétricos como a MVG. Neste caso, os experimentos mostraram um ganho significativo no valor da acurácia média com o pico passando de 81.3% para 87.9% no experimento empregando 100 amostras de treinamento e de 85.7% para 89.2% com 200 amostras de treinamento. O mesmo não ocorreu nos experimentos empregando o classificador SVM. Utilizando-se SFS com o critério Distância de Bhattacharyya para seleção de variáveis o ganho para o classificador SVM mostrou ser mínimo, resultando ainda em uma oscilação nos valores de acurácia média estimada para distintos valores de dimensionalidade dos dados. Os experimentos envolvendo 100 amostras de treinamento (Figura 7) servem

para ilustrar este fato. Quando a seleção de variáveis é feita empregando distâncias estatísticas em um algoritmo como o SFS, a seleção de bandas será ótima para classificadores que implementam critérios semelhantes para classificação, como é o caso de MVG. No caso de classificadores não paramétrios como o SVM, que implementam critérios geométricos no processo de classificação, a mesma abordagem para seleção de variáveis possivelmente não será adequada, como ilustram os experimentos.

De um modo geral, nota-se que classificadores paramétricos como o implementado no aplicativo CAB-MVG, sofre os efeitos do fenômeno de Hughes, ou seja, para um número limitado de amostras de treinamento, seus parâmetros se tornam pouco confiáveis com o acréscimo do número de bandas, causando degradação em sua performance; por outro lado, as acurácias obtidas pelo classificador proposto (CAB-SVM – não paramétrico), em geral, após atingirem um patamar sofrem pouca variação na acurácia média.

5. CONCLUSÕES

Os experimentos desenvolvidos tendem a confirmar a eficácia da metodologia proposta. Os quatro conjuntos de dados apresentam valores de acurácia média superiores com a ferramenta CAB-SVM (que implementa o classificador SVM) do que os obtidos com a ferramenta CAB-MVG (que faz uso do classificador MVG). Como seria de esperar, o maior ganho em acurácia média (2%) obtido pelo classificador SVM comparado com a produzida pelo classificador MVG, ocorreu no experimento que empregou o menor número de amostras de treinamento (50 amostras). Neste experimento, foi empregado um kernel polinomial de segundo grau. O valor mais alto para acurácia média (90.7%) obtido empregando o classificador SVM na árvore binária (ferramenta CAB-SVM), ocorreu no experimento empregando 300 amostras de treinamento e um kernel RBF com o parâmetro γ igual a 1.5. Além de proporcionar resultados melhores que os obtidos com o CAB-MVG, o CAB-SVM apresenta ainda a vantagem de não sofrer os efeitos dos fenômeno de Hughes: a partir do momento que atinge determinado patamar, a acurácia média se mantém relativamente constante, quando na ausência de amostras ruidosas em seus conjuntos de treinamento.

Apesar dos resultados promissores o classificador proposto tem limitações, principalmente no que diz respeito à seleção de variáveis. Além de tomar um longo tempo de processamento, a função SFS (com o critério de separabilidade Distância de Bhattacharyya) se mostrou inadequada para seleção de variáveis em classificadores não paramétricos. Os resultados obtidos com a ferramenta CAB-SVM (feita a seleção de variáveis via SFS em cada nó da árvore) são praticamente iguais aos resultados obtidos sem o uso de SFS. Tal limitação do classificador proposto ficou bem demonstrada nos experimentos, e acontece devido à sua alta sensibilidade à ruídos ou desvios maiores presentes em amostras de treinamento, causando degradação de sua performance em termos de acurácia média. Como

sugestão para futuros desenvolvimentos neste tópico, sugere-se que sejam investigadas abordagens alternativas para seleção de variáveis em cada nó da árvore binária, visando maximizar os resultados em termos de acurácia.

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.

AEBERHARD, S.; COOMANS, D.; DE VEL, O.; Comparative Analysis of Statistical Pattern Recognition Methods in High Dimensional Settings. *Pattern Recognition,* vol.27, no8, pp 1065-1077, 1994.

BERGE, A.; JENSEN, A.C.; SOLBERG, A.S.; Sparse inverse covariance estimates for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing***,** vol.45, no5, pp 1399–1407, maio de 2007.

BROWN, M.; LEWIS, H.G.; GUNN, S.R.; Linear Spectral Mixture Models and Support Vector Machines for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, vol.38, no5, pp 2346-2360, setembro de 2000.

BROWN, M.; GUNN, S.R.; LEWIS, H.G.; Support Vector Machines for Optimal Classification and Spectral Unmixing. *Ecological Modelling,* vol.120, no2-3, pp 167–179, 1999.

CRISTIANINI, N.; SHAWE-TAYLOR, J., *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, England, Cambridge University Press, 2000.

DUDA, O. R.; HART, P. E.; STORK, D. G.; *Pattern Classification.* Second Edition. A Wiley-Interscience Publication: 2000.

FRIEDMAN, J. H.; Regularized Discriminant Analysis. *Journal of the American Statistical Association*, vol.84, no405, pp 165-175, 1989.

FUKUNAGA, K.; *Introduction to Statistical Pattern Recognition.* Second Edition. Academic Press: 1990.

HAMEL, L.; *Knowledge Discovery with Support Vector Machines*. USA, Wiley, 2009.

HERBRICH, R.; *Learning Kernel Classifiers: Theory and Algorithms*. The MIT Press, 2002.

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, no4, pp 725–749, 2002.

JACKSON, Q.; LANDGREBE, D.A.; An adaptive classifier design for high-dimensional data analysis with a limited training data set. *IEEE Transactions on Geoscience and Remote Sensing*, vol.39, no12, pp. 2664 – 2679, dezembro de 2001.

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

KUO, B.C.; CHANG, K.Y.; Feature extractions for small sample size classification problem; *IEEE Trans. Geosci. Remote Sensing,* vol.45, no3, pp 756–764, março de 2007;

LANDGREBE, D. A.; *Signal Theory Methods In Multispectral Remote Sensing.* Wiley Interscience, 2003.

LICZBINSKI, C.; HAERTEL, V.; A new Approach to Estimate a Priori Probabilities in Remote Sensing Digital Image Classification. *Canadian Journal of Remote Sensing Journal Canadien De Télédétection,* vol.34, no2, pp 135-142, abril de 2008.

LORENA, A. C.; CARVALHO, A. C. P. L. F.; Uma Introdução às Support Vector Machines. Revista de Informática Teórica e Aplicada. *Revista de Informática Teórica e Aplicada*, vol.14, no2, pp 43-67, 2007.

MELGANI, F.; BRUZZONE, L.; Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. *IEEE Transactions on Geoscience and Remote Sensing***,** vol.42, no8, pp 1778-1790, agosto de 2004.

MORAES, D. A. O.; *Extração de Feições em Dados Imagem com Alta Dimensão por Otimização da Distância de Bhattacharyya em um Classificador de Decisão em Árvore.* Porto Alegre, RS. 2005. Dissertação de Mestrado – Departamento de Pós-Graduação em Sensoriamento Remoto – UFRGS.

SERPICO, S. B.; D'INCA, M.; MELGANI, F.; MOSER, G.; A Comparison of Feature Reduction Techniques for Classification of Hyperspectral Remote-Sensing Data. *Procedings of Spie, Image and Signal Processing of Remote Sensing VIII***.** Vol 4885, 2003.

(Recebido em setembro de 2009. Aceito em dezembro de 2009.)