Seleção de variáveis em imagens hiperespectrais para classificadores SVM

Rafaela Andreola [1]
Victor Haertel [2]

[1]Faculdade da Serra Gaúcha – FSG
Rua Os Dezoito do Forte, 2366, Caxias do Sul, RS - CEP: 95020-472
rafaela.andreola@gmail.com

[2]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
victor.haertel@ufrgs.br

Abstract. Very recent articles report that Support Vector Machine (SVM) methods generally outperform traditional statistical and neural methods in classification problems involving hyperspectral images. In this paper we investigate the performance of the SVM classifier when applied to high dimensional image data, depicting natural scenes. Since the SVM classifier deals with a pair of classes at a time, a multi-stage classifier, structured as a binary tree is proposed, comparing classification approaches based on three different feature selection methods, i.e., selection of N features at regular intervals throughout the electromagnetic spectrum, Sequential Forward Selection (SFS) and the Recursive Feature Elimination technique (RFE). 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: feature selection, support vector machines, high-dimensional image data; seleção de variáveis, imagens hiperespectrais.

1. Introdução

Imagens de sensoriamento obtidas por sensores hiperespectrais fornecem, com uma resolução espectral muito alta, informações a respeito da superfície da Terra, geralmente resultando em dezenas de bandas espectrais. Comparado com dados de imagens multiespectrais, que envolvem um número reduzido de bandas espectrais, dados hiperespectrais oferecem uma capacidade muito mais alta para fins de discriminação das classes presentes na cena. Entretanto, do ponto de vista metodológico, classificação de dados em alta dimensionalidade não é uma tarefa trivial. No contexto da classificação supervisionada, possivelmente a maior dificuldade consiste na estimação do grande número de parâmetros resultantes da alta dimensionalidade dos dados. Em situações práticas, o número de amostras de treinamento é geralmente limitado, resultando em estimativas pouco confiáveis para os parâmetros do classificador. Esta condição limita a acurácia que pode ser obtida no processo de classificação, conforme ilustrado pelo conhecido Fenômeno de Hughes: iniciando o processo de classificação com um pequeno número de bandas espectrais, a acurácia obtida tende a aumentar na medida em que bandas adicionais, isto é, informação adicional é inserida no processo. Em um determinado ponto, a acurácia da classificação atinge um máximo para em seguida passar a declinar, na medida em que novas bandas continuam a ser adicionadas ao processo, refletindo a crescente incerteza nos valores estimados para os parâmetros do classificador (Hughes, 1968). Para reduzir esse problema, diferentes abordagens têm sido propostas na literatura. As abordagens mais conhecidas são: 1) Técnicas em Análise Discriminante Regularizada: regularização das matrizes de covariância (Tadjudin & Landgrebe, 1999); 2) Incremento no número de amostras de treinamento, com a introdução das amostras semi-rotuladas (Jackson & Landgrebe, 2001); 3) Redução na dimensionalidade dos dados: técnicas de seleção ou extração de variáveis (Jimenez & Landgrebe, 1999 e Serpico & Bruzzone, 2001). Uma outra possível alternativa consiste no uso de classificadores

não-paramétricos. Nesta alternativa, especial menção merece a técnica conhecida como Support Vector Machines (SVM) (Melgani & Bruzzone, 2004; Huang *et al*., 2002; Camps-Valls *et al*., 2006; e Bazi & Melgani, 2006).

Esta última abordagem tem mostrado ser particularmente promissora devido à sua baixa sensibilidade ao Efeito de Hughes em comparação com classificadores paramétricos. SVM está baseado no princípio de maximização da margem, não envolvendo portanto o conhecimento das distribuições estatísticas das classes no espaço hiperdimensional para realizar a classificação. Graças a esta propriedade, SVM é uma ferramenta de classificação de sucesso em diversas áreas de conhecimento.

Mais especificamente, em Huang *et al*. (2002), o classificador SVM é comparado com três outras técnicas de classificação mais tradicionais, como a Máxima Verossimilhança, Redes Neurais e classificadores em árvores de decisão, no mapeamento da cobertura do solo utilizando dados TM-Landsat5. O autor investiga também a influência do uso de dois tipos distintos de *kernels* e de seleção de amostras para treinamento. Melgani & Bruzzone (2004) fazem uma análise detalhada comparando diferentes estratégias de classificação SVM multiclasse com outros dois classificadores convencionais (*K-nearest Neighbors* e redes neurais RBF). Com base em dados AVIRIS, essa análise foi realizada no espaço característico original e em subespaços de várias dimensionalidades. Em Camps-Valls *et al*. (2006), os autores propõem uma família de classificadores baseados em SVM que permitem a fusão das informações espectrais e espaciais em imagens hiperespectrais.

Em geral, a maior parte destes trabalhos ressalta a superioridade de classificadores SVM sobre classificadores paramétricos e redes neurais mais tradicionais. Os resultados destes estudos sugerem que uma maior acurácia no processo de classificação pode ser obtida operando-se não no espaço característico original mas em subespaços deste, onde um menor numero de variáveis é considerado. Desta forma métodos de seleção de variáveis devem ser investigados.

A formulação matemática básica de SVM para resolver problemas de classificação binária e multiclasse é recordado na subseção 1.1. Estratégias para seleção de variáveis utilizados para comparação serão vistas na subseção 1.2. Materiais e métodos, resultados e discussões são apresentados nas seções 2 e 3, respectivamente. Finalmente, as conclusões são tiradas na Seção 4.

1.1 SVM

Por simplicidade, considerar-se-á primeiro o problema de classificação supervisionada para 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).

Seja $\mathbf{x}_i$ ($i=1, 2, ..., M$) 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 $\mathbf{x}_i \in \omega_1$, $y_i= -1$ se $\mathbf{x}_i \in \omega_2$. A forma geral da função de decisão linear é:

$$D(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b \qquad (1)$$

onde $\mathbf{w}$ 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.

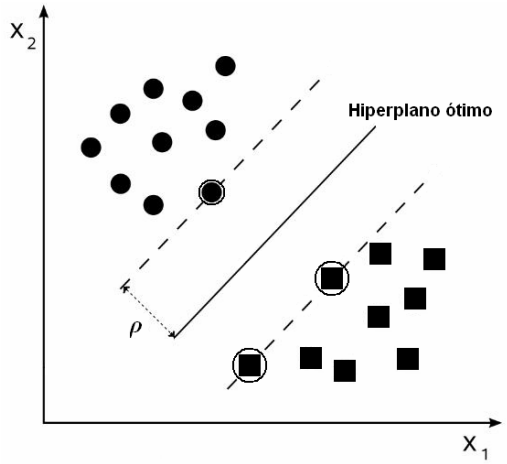


Figura 1: O hiperplano ótimo separando os dados com a máxima margem ρ. Os *support vectors* (amostras circuladas) e uma distribuição dos dados no R^2 (atributos x_1 e x_2). Fonte: Adaptado de ABE (2005).

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). Esse espaço é comumente denominado de espaço característico *(feature space)*.

Representando por $\mathbf{g(x)}=(g_1(\mathbf{x}), ..., g_l(\mathbf{x}))^t$ uma função de transformação que mapeia as amostras $\mathbf{x}_i$ 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:

$$D(\mathbf{x}) = \mathbf{w}^t g(\mathbf{x}) + b \tag{2}$$

onde $\mathbf{w}$ é um vetor l-dimensional e b é o termo independente (*bias*).

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

$$\sum_{i,j=1}^{M} h_i h_j \mathbf{H}(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \tag{3}$$

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

$$\mathbf{H}(\mathbf{x}, \mathbf{x}') = g^t(\mathbf{x}) g(\mathbf{x}') \tag{4}$$

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 $\mathbf{H}(\mathbf{x}, \mathbf{x}')$. A vantagem do uso de *kernels* é que não se precisa lidar com o espaço característico de alta-dimensão explicitamente: usa-se $\mathbf{H}(\mathbf{x}, \mathbf{x}')$ no treinamento e classificação ao invés de $g(\mathbf{x})$.

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

$$Q(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j \mathbf{H}(\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

sujeito às restrições:

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

onde a constante C representa o parâmetro que permite controlar a forma da função discriminante e, conseqüentemente, a fronteira de decisão quando os dados não são

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

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

sendo *S* o conjunto de índices correspondendo aos multiplicadores de Lagrange não-zeros (chamados de SVs) e *b* o coeficiente linear.

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

$$\mathbf{H}(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \| \mathbf{x} - \mathbf{x}' \|^2) \quad (8)$$

onde γ é um parâmetro positivo para controle.

A regra de classificação é, então:

$$D(\mathbf{x}) > 0 \quad \mathbf{x}_i \in \omega_1 \quad (9)$$
$$D(\mathbf{x}) < 0 \quad \mathbf{x}_i \in \omega_2$$

Se $D(\mathbf{x})=0$, então **x** está sobre o hiperplano separador e não é classificado. Quando as amostras de treinamento são linearmente separáveis, a região $\{\mathbf{x} \mid 1 > D(\mathbf{x}) > -1\}$ é a região de generalização. 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 como as presentes em Melgani & Bruzzone (2004). 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.

1.2 Seleção de variáveis

Segundo Landgrebe (2003), o emprego de todas as variáveis originais disponíveis (número total de bandas) ou de um número grande destas pode ser prejudicial à eficiência do processo de classificação. As bandas espectrais de cenas naturais frequentemente apresentam um alto grau de correlação, o que significa um volume alto de dados, com informações parcialmente repetidas. Em classificadores paramétricos ocorre ainda o problema da estimação de um número muito grande de parâmetros, geralmente a partir de um número limitado de amostras de treinamento. Deste modo, uma etapa importante que deve preceder o processo de classificação refere-se redução da dimensionalidade dos dados. Existem duas abordagens genéricas para esta finalidade: seleção de variáveis (*feature selection*) e extração de variáveis (*feature extraction*). A seguir, são brevemente revistas três abordagens entre as mais utilizadas para fins de seleção de variáveis:

A) SVUD (Seleção de Variáveis Uniformemente Distribuídas): as *N* bandas desejadas serão coletadas a intervalos regulares ao longo de todo o conjunto disponível. Esta abordagem baseia-se no fato de que bandas espectrais vizinhas são as que frequentemente apresentam os mais altos graus de correlação (repetição da mesma informação).

B) SFS (*Sequential Forward Selection*): trata-se aqui de um algoritmo iterativo, que busca selecionar o melhor subconjunto de variáveis. Parte-se do conjunto *X*, com dimensionalidade *M* o qual inicialmente contém as variáveis originais, e de um conjunto *S*, inicialmente vazio ($S=\varnothing$). A cada etapa do processo iterativo, uma variável em *X*-*S* é transferida de X para S, aquela que maximiza um determinado critério de separabilidade, (como por exemplo a distância de Bhattacharyya), é acrescentada a *S*. O processo continua até que o número de variáveis em *S* atinja o valor desejado *N* ($N<M$) (Serpico *et al*. 2003).

C) RFE (*Recursive Feature Elimination*): essa técnica realiza a seleção do subconjunto das *N* variáveis com maior poder de separação das classes, através de um método de eliminação seqüencial guiada pelo princípio de maximização da margem. O processo consiste em treinar o classificador SVM com o conjunto total de bandas *X* e remover a banda p $\in$ *X* que mais diminui a margem até que o número *N* de bandas seja atingido. Para cada banda p (p $\in$ *X*), é preciso calcular o valor de $W^2_{(-p)}(\alpha)$. Esta medida é inversamente proporcional à margem e é definida como

$$W^2_{(-p)}(\alpha) = \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{H}(\mathbf{x}_i^{(-p)} \mathbf{x}_j^{(-p)}) \tag{10}$$

onde $\mathbf{x}_i^{(-p)}$ representa a *i*-ésima amostra de treinamento definido por *S* sem a banda p. Em seguida, a variável com a menor diferença | $W^2(\alpha)$ - $W^2_{(-p)}(\alpha)$ | será removida de *S*. O processo continua até que o número de variáveis em *S* atinja o valor desejado *N* . Quando o espaço de entrada original é grande, para acelerar o processamento computacional, os autores em Guyon *et al*. (2002) sugerem que mais de uma banda deve ser removida ao mesmo tempo. Isso otimiza o processo, mas leva a soluções subótimas.

Para investigar melhor a questão de seleção de variáveis, testaremos os três métodos de seleção de variáveis citados para a classificação multiclasse, em um classificador SVM estruturado em forma de árvore binária.

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 utilizando diferentes métodos de seleção de variáveis. É empregado 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 (Johnson & Wichern, 1982):

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

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

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

Das classes disponíveis na cena foram selecionadas seis 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 – pastagens e árvores	747
ω_4 – soja cultivo convencional	614
ω_5 – soja cultivo mínimo	2468
ω_6 – soja plantio direto	968

Conforme mencionado na seção anterior, SVM como classificador considera um par de classes a cada vez. O classificador proposto por Andreola & Haertel (2010), desenvolvido em forma de árvore binária, emprega uma estrutura em estágios múltiplos, composta por um conjunto de classificadores binários (Figura 2). Para o treinamento do classificador, em cada nó da árvore, aplica-se o algoritmo que pode ser visto na Figura 2a. As bandas selecionadas por um dos três métodos já discutidos (etapa marcada em cinza no fluxograma) serão usadas para o cálculo dos coeficientes do classificador SVM. Na Figura 2b, pode-se observar o fluxograma do algoritmo de teste do classificador.

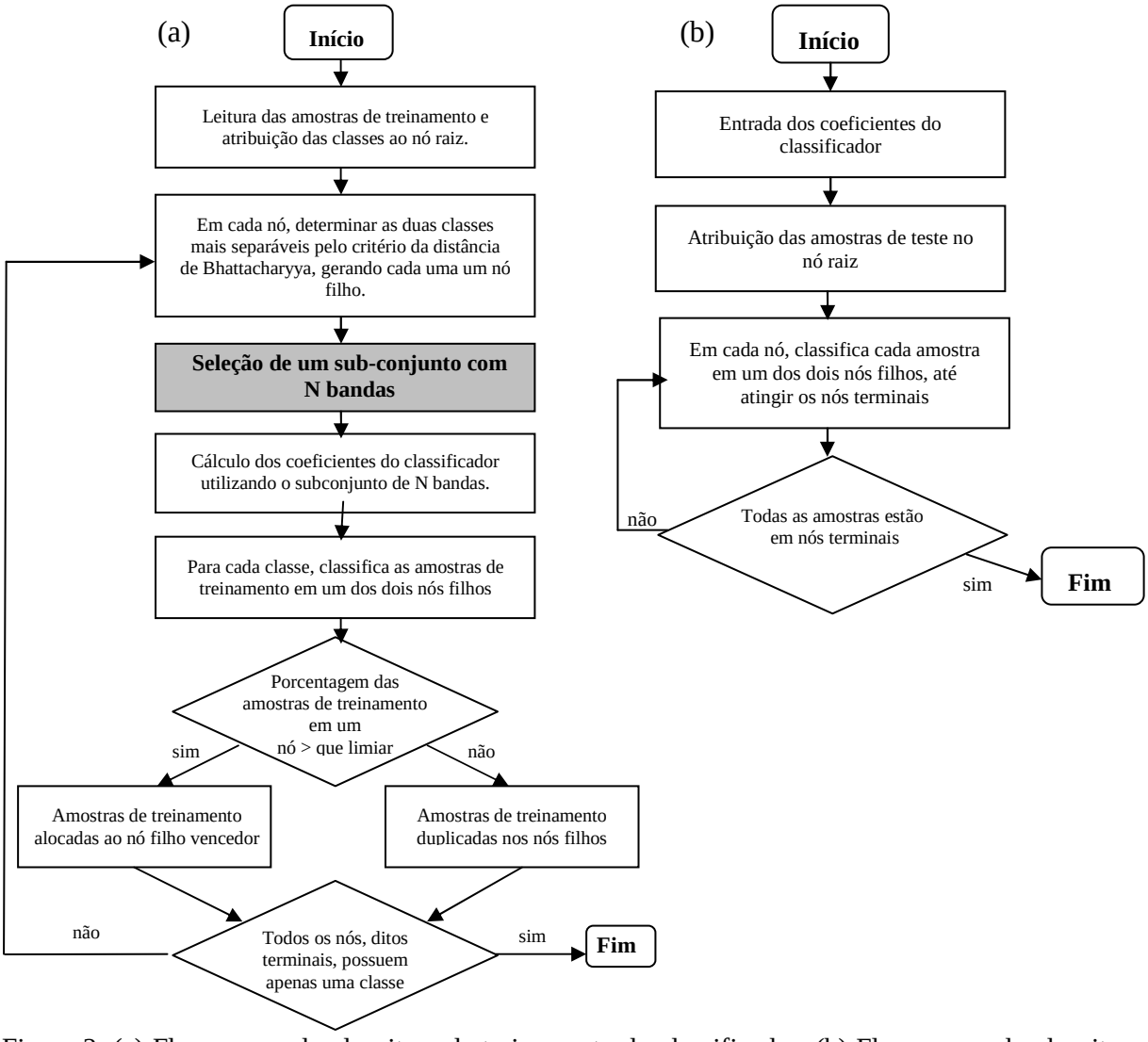
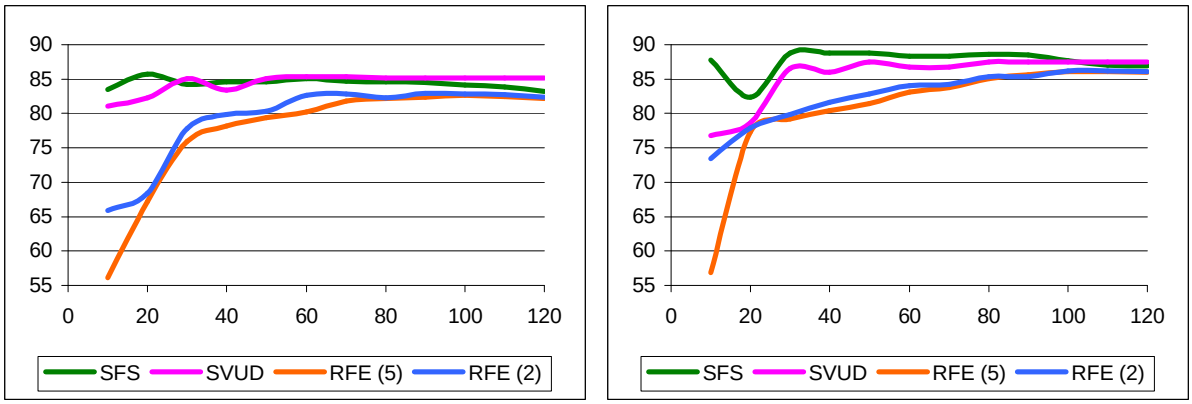


Figura 2: (a) Fluxograma do algoritmo de treinamento do classificador; (b) Fluxograma do algoritmo de teste do classificador.

3. Resultados e Discussões

Para exemplificar a diferença existente entre os resultados produzidos pelas três abordagens de seleção de variáveis (citadas anteriormente) no classificador SVM, elaborou-se dois estudos, que serão avaliados segundo a acurácia média. Primeiramente, dividiu-se o conjunto das amostras disponíveis em dois subconjuntos, um para fins de treinamento (50 amostras) e o outro para fins de teste (300 amostras). Nos experimentos a acurácia média produzida por cada uma das abordagens foi estimada variando a dimensionalidade dos dados de 10 à 120 bandas (com passo de 10). 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 2 e C (cujo valor influencia nas Equações 6 e 8) igual a 1 (Figura 3a). Os parâmetros do classificador SVM foram determinados de acordo com os melhores resultados obtidos por Andreola & Haertel (2010) para a metodologia citada. Em um segundo experimento, foi utilizado 100 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 3b). Para o método de seleção de variáveis RFE, decidiu-se eliminar 5 e 2 bandas por vez, conforme sugestão dos autores em Guyon *et al*. (2002), para melhorar o desempenho computacional. Desta forma, as soluções obtidas são consideradas subótimas. O tamanho das 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.



a) (b)

Figura 3: (a) Acurácia média para 50 amostras de treinamento e 300 amostras de teste. (b) Acurácia média para 100 amostras de treinamento e 300 amostras de teste.

A análise dos resultados dos dois experimentos (Figura 3) evidencia a melhor acurácia média obtida com os métodos SFS e SVUD. Para 50 amostras de treinamento (Figura 3a), o classificador atingiu 85.7% de acurácia média com 20 bandas para o método SFS e 85.3% com 60 bandas para SVUD. Para 100 amostras de treinamento, o classificador atingiu 88.8% de acurácia média com 30 bandas para o método SFS e 87.5% com 50 bandas para SVUD. Tais métodos já haviam sido investigados por Andreola & Haertel (2010). Os resultados obtidos com relação ao método RFE confirmam as conclusões obtidas por Bazi & Melgani (2006) que sugeriam novos estudos deste método com relação à imagens hiperespectrais de sensoriamento remoto. Devido às simplificações feitas na codificação e pelo fato de serem excluídas mais de uma variável por vez obteve-se soluções subótimas para o método RFE. Disto resulta uma acurácia média mais baixa que para os outros dois métodos.

4. Conclusões

Neste estudo apresentou-se três métodos de seleção de variáveis aplicados em um classificador SVM com o objetivo de compará-los no que se refere à acurácia média. Os resultados obtidos apontam para os métodos SFS e SVUD como os que obtiveram melhor eficácia. Apesar de apresentar os melhores resultados, o método SFS, por ter como critério de separabilidade uma distância estatística apresenta sinais do Efeito de Hughes, com a acurácia decrescendo com a inclusão de novas bandas. Sugere-se testes mais aprofundados com o método RFE, excluído apenas uma banda por vez, para que se obtenha resultados mais bem fundamentados.

Referências Bibliográficas

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

Andreola, R.; Haertel, V.; Classificação de Imagens Hiperespectrais Empregando Support Vector Machines. **Boletim de Ciências Geodésicas.** Vol. 16, no.2, p 210-231, 2010.

Bazi, Y.; Melgani, F.; Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images. **IEEE Trans. Geosci. Remote Sensing.** Vol. 44, no. 11, p. 3374–3385, nov. 2006

Camps-Valls, G.; Gomez-Chova, L.; Muñoz-Marí, J.; Vila-Francés, J.; Calpe-Maravilla, J.; Composite kernels for hyperspectral image classification. **IEEE Geosci. Remote Sens. Lett.** Vol. 3, no. 1, p. 93–97, jan. 2006.

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.

Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V.; Gene selection for cancer classification using support vector machines. **Mach. Learning**. Vol. 46, no. 1–3, p. 389–422, 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, no. 4, 2002.

Hughes G. F.; On the mean accuracy of statistical pattern recognizers. **IEEE Trans. Inf. Theory**. Vol. IT-14, no. 1, p. 55–63, jan. 1968.

Jackson, Q.; Landgrebe, D. A.; An adaptive classifier design for highdimensional data analysis with a limited training data set. **IEEE Trans. Geosci. Remote Sensing.** Vol. 39, no. 12, p. 2664–2679, dec. 2001.

Jimenez, L. O.; Landgrebe, D. A.; Hyperspectral data analysis and feature reduction via projection pursuit. **IEEE Trans. Geosci. Remote Sensing**. Vol. 37, no. 6, p. 2653–2667, nov. 1999.

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

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

Melgani, F.; Bruzzone, L.; Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. **IEEE Transactions on Geoscience and Remote Sensing**. Vol 42, no. 8, aug 2004.

Serpico, S. B.; Bruzzone, L.; A new search algorithm for feature selection in hyperspectral remote sensing images. **IEEE Trans. Geosci. Remote Sensing**. Vol. 39, no. 7, p. 1360–1367, jul. 2001.

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.

Tadjudin, S.; Landgrebe, D. A.; Covariance estimation with limited training samples. **IEEE Trans. Geosci. Remote Sensing.** Vol. 37, no. 4, p. 2113–2118, jul. 1999.