

# Classificação de imagens Landsat 7 ETM multiespectrais utilizando SVM

MAURO ALIXANDRINI<sup>1</sup>

<sup>1</sup>Universitat Karlsruhe - IPF (Bolsista CNPq)  
Kaiserstrae, 12, 76128 Karlsruhe/Alemanha  
Mauro.Alixandrini@ipf.uni-karlsruhe.de

**Abstract.** Recently a new learning algorithm family has a great attention from Remote Sensing Researchers. Dixon & Candade (2008) and Foody (2004) show the SVMs (Support Vector Machines) application have success with Remote Sensing, especially with high dimensional classifications problems. The SVMs are recognized for your high generalization grade, besides offer a mathematic solid basis based on probability statistic. Differently than ANN (Artificial Neural Network) that is based empiric risk minimization, the SVM is based on another concept, the structural risk minimization foresee no utilized data, realized a extrapolation from trainee data. Due this reason this technique evaluation is important in different remote sensing applications that need a high dimensionality grade. In the development about ML (Machine Learning) techniques, many researchers are inspired in biologics systems to development, for instance, learning symbolic field and cognitive process with decisions tree. Even now the Support Vector Machines are based on statistics theory. This paper has as objective an evaluation the SVMs performance in comparison with a classic classification with Maximum Likelihood (ML) and modern process as Neural Network for middle resolution satellites images as Landsat 7 ETM. An application going to be as class differentiation with high difficult qualification originated from agriculture and cattle breeding usage. The implementation results from proposed models can exemplify additional applications the SVMs, such as explore the concepts related.

**Keywords:** remote sensing, image processing, classification, SVM- Suport Vector Machienes.

**Palavras-chave:** processamento de imagens de sensoriamento remoto, classificaao, SVM – Maquinas de Suporte Vetorial.

## 1. Teoria do aprendizado estatístico

Uma ponto central do aprendizado de maquinas (ML) diz respeito  capacidade de generalizaao do classificador. O problema fica claro quando se assume a seguinte situaao: dado um classificador gerado por um algoritmo de aprendizado, este sera sujeito a um grupo chamado de treinamento, porem como determinar quao bom  seu desempenho em estimar saídas para entradas nao vistas durante a fase de treinamento? Este foi o problema exposto por Vapnik e Chernoventkis ainda na decada de 60, estes pesquisadores desenvolveram estudos que levaram  criaao da TAE - Teoria do Aprendizado Estatístico (Vapnik, 1995), que formaliza a relaao entre a capacidade de generalizaao do classificador, a complexidade de classe de funoes que o algoritmo de aprendizado pode gerar e a minimizaao dos erros no treinamento. A seguir, os pontos basicos da TAE serao abordados de maneira mais formal.

\_\_\_ Considere um conjunto S formado por n elementos tal que  $S = \{(x_1, y_1) \dots (x_n, y_n)\}$ , onde  $x_i \in R^m$  e  $y_i \in \{+1, -1\}$ . Os vetores possuem dimensao m e sao utilizados para representar objetos do mundo real, como, por exemplo, as informaoes espectrais dos pixels ou informaoes de contexto. Os  $y_i$  sao conhecidos como rotulos e correspondem as classes de saıda possıveis as quais os vetores estao associados. Os pares  $(x_i, y_i)$  sao os exemplos rotulados do conjunto de treinamento  $S^2$ . (Souza, 2005)

O princıpio basico surge da classificaao binaria (ou reconhecimento de padroes binario), ele  encontrado nas mais diversas reas de conhecimento: domınios medicos (Takeuchi e Collier, 2004), analise de textos (Joachims, 2002), problemas de visao computacional (Zheng et al., 2004), bioinformatica (Noble, 2004) e matematica computacional (Carvalho, 2005).

A solução consiste em descobrir uma função definida como:  $f : \mathbb{R}^m \rightarrow \{+1, -1\}$ , que é chamada de classificador com  $f \in F$ , onde  $F$  contém todos os possíveis classificadores que um algoritmo de aprendizado pode induzir. A geração de um classificador por um algoritmo e aprendizado aplicado a um conjunto de exemplos rotulados é realizada na etapa de treinamento.

Uma vez escolhida uma função de mapeamento  $f$ , pode-se classificar elementos não vistos durante a fase de treinamento, esta classificação só é possível sob a condição de que os elementos desconhecidos ao classificador (coletivamente referidos como conjunto de teste) sejam gerados independentemente pela mesma distribuição de probabilidade  $P$  que gerou os elementos do conjunto de treinamento. Define-se então o Risco esperado  $R(f)$ .

$$R(f) = \int \frac{1}{2} |f(\bar{x}) - y| dP(\bar{x}, y) \quad (1)$$

Porém o valor de  $dP(\bar{x}, y)$  nem sempre é conhecido, as informações disponíveis dizem respeito unicamente ao conjunto de treinamento. Com isso o natural é que se minimize o erro durante a fase de treinamento e esperasse que o mesmo aconteça com o restante do conjunto de dados. À essa quantidade de classificações incorretas com respeito as amostras de treinamento dá-se o nome de Risco empírico ( $R_{emp}(f)$ ).

Contudo essa estratégia, a minimização do Risco empírico, nem sempre é satisfatória incorrendo em um erro denominado por Vapnik (1995) como *overfitting*.

Embora não seja possível calcular  $R(f)$ , um valor muito pequeno de  $R_{emp}(f)$  não implica necessariamente em uma boa generalização, desta forma foi introduzida a seguinte inequação (Vapnik, 1995):

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h(\log(\frac{2 \cdot I}{h}) + 1) - \log(\frac{\eta}{4})}{I}} \quad (2)$$

com probabilidade de pelo menos  $1 - \eta$  para  $N > h$ . Onde  $h$  representa a dimensão VC do conjunto  $F$ ,  $\eta$  é um valor real entre 0 e 1 e  $N$  é a quantidade de exemplos do conjunto  $S$ . Ao termo da raiz quadrada dá-se o nome de confiança. Essa inequação fornece um limite superior para o Risco esperado, logo o objetivo é não só minimizar o risco empírico como determinar uma classe de funções que mantenham o parâmetro  $h$  pequeno. Para tanto  $F$  é dividida hierarquicamente em subconjuntos de forma que:

$$F_0 \subset F_1 \subset F_2 \subset \dots \subset F_n \subset F \quad (3)$$

em que  $VC(F_0) \leq VC(F_1) \leq VC(F_2) \leq \dots \leq VC(F_n) \leq VC(F)$

Observa-se que uma classe muito pequena como  $F_0$  pode reduzir o termo da raiz quadrada na equação, mas ao mesmo tempo o número de erros na fase de treinamento permanece alto. Enquanto uma classe muito grande, reduz os erros durante o treinamento, mas o valor de  $VC(F_k)$  leva a um grande termo na raiz quadrada. Por meio deste princípio se encontra uma função  $F$  que não minimiza somente o risco empírico, como também, o termo de Confidência. (Carvalho, 2005).

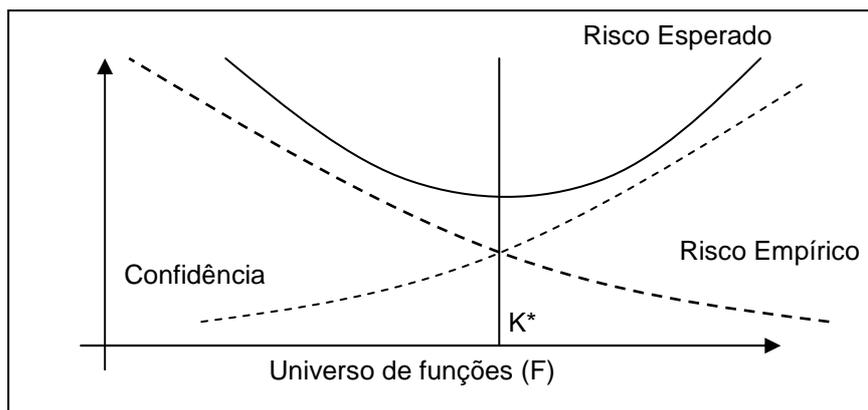


Figura 1 – Valor de equilíbrio  $K^*$  entre a confiança e o Risco Empírico. (Fonte: Carvalho,2005)

## 2. SVMs

A SVM possibilita a abordagem dos problemas de duas formas. Na primeira, não são considerados erros de classificação, ou seja, padrões que por ventura estejam fora dos valores típicos de sua classe irão desviar o hiperplano de separação de forma que este padrão (incorreto) seja classificado como correto. Na segunda, são estabelecidas variáveis de folga, de forma que padrões fora dos valores típicos da sua classe podem ser desconsiderados, dependendo da quantidade de folga que se use, diminuindo assim a probabilidade de erro de classificação.

Algumas das principais características das SVMs são (Smola,1999):

- Boa capacidade de generalização - os classificadores gerados por uma SVM em geral alcançam bons resultados em termo de generalização. Essa capacidade é medida por sua eficiência na classificação de dados que não pertençam ao conjunto utilizado em seu treinamento, portanto, é evitado o *overfitting* (memoriza 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 vistos durante o treinamento).
- Robustez em grandes dimensões – as SVMs são robustas diante de objetos de grandes dimensões, como por exemplo, imagens. Comumente há a ocorrência de *overfitting* nos classificadores gerados por outros métodos inteligentes sobre esses tipos de dados.
- Teoria bem definida – as SVMs possuem uma base teórica bem estabelecida dentro da Matemática e Estatística.

Apesar do bom desempenho em diversas aplicações, algumas ressalvas devem ser feitas com relação à complexidade computacional na busca por soluções, da ordem de  $n^2$  a  $n^3$ , em que  $n$  é o número de exemplos de treinamento. Apesar de polinomial, esse valor pode ser significativo em aplicações em que o volume de dados é grande. Porém, os tempos de treinamento são consideravelmente menores do que em outros algoritmos de aprendizagem. Outro problema é a forma como o conhecimento é adquirido e codificado pelos classificadores gerados por SVMs. Esse conhecimento não é facilmente interpretado, o que é uma desvantagem sobre outras técnicas de AM como as árvores de decisão, que organizam o conhecimento adquirido em uma estrutura diretamente interpretável (Cristianini e Shawe-Taylor, J. 2001).

### 2.1 Classificação Binária

Originalmente uma SVM constrói um classificador binário a partir de um conjunto de treinamento. A TAE fornece limites para a definição da função de classificação restringindo a função  $f$  a ser escolhida, a esse procedimento denominasse minimização do erro estrutural (SRM). A SVM constrói um conjunto de hiperplanos cujos limites da dimensão VC possam

ser computados e usa o princípio do SRM para identificar o hiperplano ótimo que maximize a margem entre as amostras de treinamento mais próximas.

No princípio do desenvolvimento as SVMs só eram capazes de separar padrões de treinamento linearmente separáveis. Com a adoção das chamadas “variáveis de folga” os padrões de treinamento não linearmente separáveis no espaço de característica puderam então ser tratados.(Baranoski,2005)

As SVMs não lineares projetam os dados de entrada em um espaço de dimensão maior onde os mesmos dados passam a ser linearmente separados. Para tanto, utilizasse o teorema de Cover que afirma que um espaço de entrada com padrões não linearmente separáveis pode ser transformado em um novo espaço de características em que os padrões são linearmente separáveis, desde que sejam satisfeitas as seguintes condições: a transformação seja não linear e a dimensão do espaço seja suficientemente grande. Dessa forma tornou-se possível construir um hiperplano ótimo separando as classes nesse espaço de características.

Uma segunda formulação para o problema das SVMs não lineares surge do aplicação do teorema de Mercer (Carvalho, 2005) para se encontrar uma função chamada *Kernel* que minimiza o custo computacional da avaliação das SVMs. A Função *Kernel* de maneira geral faria referencia a forma do hiperplano. Alguns *kernels* conhecidos são dados na Tabela 1.

Tabela 1 - Exemplos de funções Kernel.(Fonte:Chavez et al, 2006)

Kernel	Expressão	Parâmetros
Polinomial	$K(x_i, x_j) = ((x_i \cdot x_j) + a)^p$	$a, p$
RBF	$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2} \ x_i - x_j\ ^2\right)$	$\sigma^2$
Perceptron	$K(x_i, x_j) = \tanh(\beta_0 (x_i \cdot x_j) + \beta_1)$	$\beta_0, \beta_1$

Na máquina de vetores suporte RBF, o número de funções de base radial e seus centros são determinados automaticamente pelo número de vetores suporte e seus valores, respectivamente. Na máquina de vetores suporte do tipo *perceptron* de duas camadas, o número de neurônios ocultos e seus vetores de peso são também determinados automaticamente pelo número de vetores suporte e seus pesos, respectivamente (Chavez et al, 2006).

## 2.2 Classificação em Múltiplas classes

SVMs foram originalmente desenvolvidas para classificação binária. No caso de classificação em k classes,  $k > 2$ , existem duas abordagens básicas (Figura 2). A primeira é a redução do problema de múltiplas classes a um conjunto de problemas binários. Dois métodos usam essa abordagem: "decomposição um por classe" ("*one-against-all*"), separação das classes duas a duas ("*one-against-one*"). A segunda abordagem é a generalização de SVMs binárias para mais de duas classes. O método que utiliza essa abordagem é o método de Crammer e Singer.

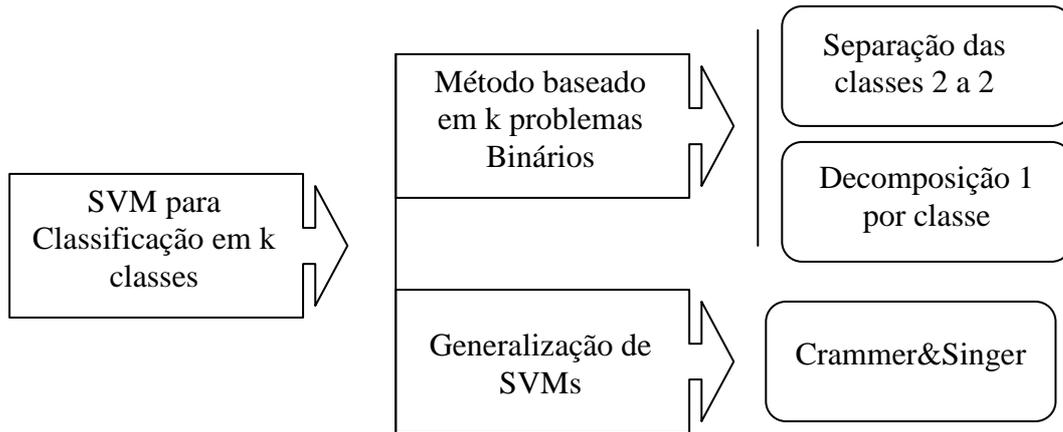


Figura 2 – Métodos Multiclasse.

### 2.2.1 Decomposição um por classe

O método de decomposição um por classe baseia-se na construção de  $k$  SVMs de classificação binária para separar uma classe de todas as outras. Em seguida, os resultados de todas as SVMs são agrupados, fazendo-se a classificação desejada nas  $k$  classes.

### 2.2.2 Decomposição duas a duas

O método de separação das classes duas a duas usa uma SVM binária para distinguir cada par de classes. Assim, são construídas  $k(k-1)/2$  SVMs. A classificação final é obtida a partir do resultado de todas as SVMs.

### 2.2.3 Método de Crammer e Singer

O método de Crammer e Singer usa uma maneira mais natural de resolver o problema de classificação em  $k$  classes,  $k > 2$ , que é construir uma função de decisão considerando todas as classes de uma vez. Nesse método, todos os exemplos de treinamento são usados ao mesmo tempo.

## 3. Classificação de imagens

Para a avaliação foram comparadas uma série de problemas binários com diferentes funções *Kernel* com classificações clássicas baseadas em pixel utilizando o classificador de Máxima verossimilhança e redes neurais artificiais.

Numa segunda etapa foram comparados problemas de múltiplas classes utilizando-se o método de Crammer e Singer adaptados para classificações multiespectrais a partir do algoritmo Libsvm (Lin, 2005). Também para este estudo utilizou-se o *Toolbox LS-SVMlab1.5* desenvolvido por Suykens et al (2002).

Foram utilizados recortes de imagens Landsat 7 TM, de cenas do oeste paranaense para os testes, utilizando-se diferentes combinações de bandas.

No primeiro teste utilizam-se três padrões de classificação binária de imagens, que se referem respectivamente a três níveis de separação espectral. Observa-se na figura 3 que mostra os gráficos bidimensionais das bandas 4 e 5 referentes aos 3 níveis abordados, classes bem separadas espectralmente(a), classes apresentando certa sobreposição de assinatura espectral (b) e classes misturadas (c).

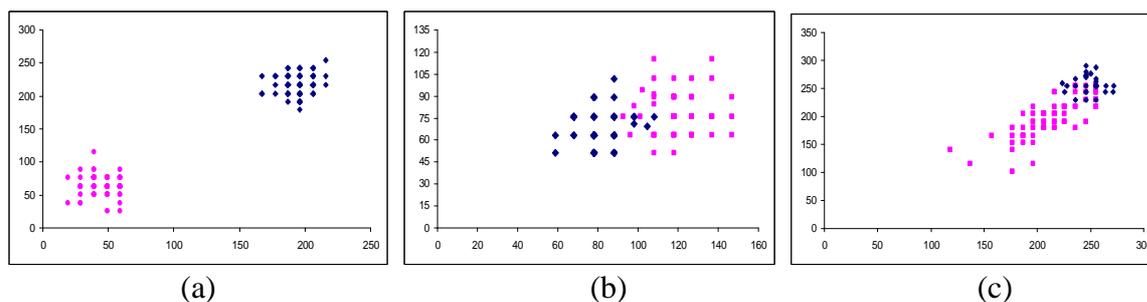


Figura 3 - Gráficos da distribuição das amostras nas bandas 4 e 5.

Num segundo teste utilizou-se um total de 5 classes, que eram formadas por padrões de culturas e diferentes estágios, solo exposto e arado para o plantio. O Recorte corresponde a uma imagem Landsat-7 de 2001 com as 3 bandas do espectro visível.

#### 4. Comparação de resultados

Os resultados para o primeiro teste revelaram que os três classificadores para os casos A e B encontram respostas semelhantes, pequenas discrepâncias decorreram da divergência dos valores de fronteira, onde o pixel toma um valor de mistura das classes.

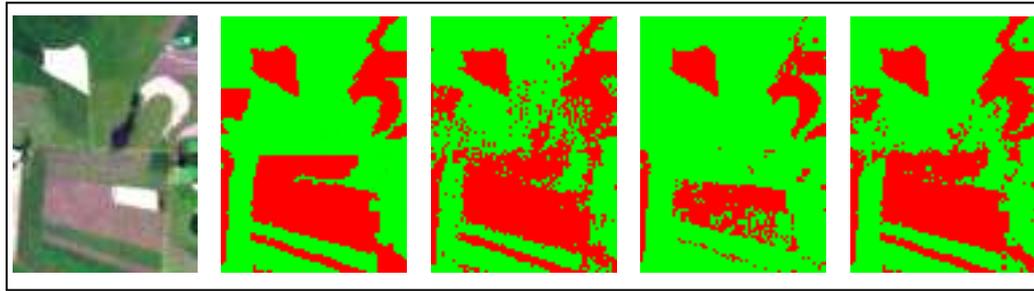
Como se observa na Tabela 2, os valores mais discrepantes apresentam-se para o caso do recorte C. Este se tratando de um problema que envolvia a mistura de classes e classes mais heterogêneas no espaço espectral.

Tabela 2 - Comparação dos resultados para o primeiro teste com 2 classes, Avaliação Global(AG) e relação de pixels classificados para as classes A/B.

	C1 AG	C1 Class. A/B	C2 AG	C2 Class. A/B	C3 AG	C3 Class. A/B
Maxver	97 %	1532/968	96 %	1697/1303	83 %	1029/771
Redes Neurais	95 %	1569/931	93 %	1621/1379	56 %	1353/447
SVM	96 %	1546/954	97 %	1682/1318	89 %	1105/695

No caso C onde existe uma mistura maior do valor das respostas espectrais, percebe-se um resultado mais acurado por parte da classificação por SVM. Essa melhor avaliação se deve a característica do problema em especial, existe uma mistura de assinaturas, uma das classes esta muito mais dispersa que a outro, fazendo com que a dispersão de probabilidade na classificação Maxver, confunda pixels que ficam na fronteira da segunda classe. Da mesma forma as classificação por redes neurais converge para uma solução menos abrangente para a classe mais dispersa, ocasionando uma maior perda na acurácia. A solução seria no caso das classificações Maxver e de Redes Neurais, a subdivisão das classes. Por se flexibilizar a SVM utilizada pode definir melhor as regiões das duas classes, porem a confusão se manteve sobre a região onde as classes sofrem sobreposição. Naturalmente a avaliação dessas regiões se torna ainda mais complexa.

A Figura 4 ilustra com detalhes o problema elaborado para o recorte C, onde se utilizam duas classes heterogêneas de cobertura do solo. A figura 4 ainda apresenta o resultado obtido a partir da classificação do recorte, com as mesmas amostras de treinamento para os três métodos de classificação propostos.



(a) (b) (c) (d) (e)

Figura 4 – Recorte C, classes de difícil definição – (a) composição RGB, (b) áreas de Controle, (c) classificação utilizando Máxima verossimilhança, (d) utilizando Redes Neurais, (e) utilizando SVM de kernel RBF.

As tabelas 3, 4 e 5 apresentam as Matrizes de Confusão referentes ao experimento com 5 classes feitos no 2 teste. Elas mostra uma avaliação feita com base no em áreas de controle, interpretadas com base no comportamento espectral das classes.

Tabela 3 – Matriz dos erros Classificação 2 Maxver- Multiclasse.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Total	<i>AU</i>
Classe 1	19033	0	6	0	0	19039	99,97%
Classe 2	34	2905	89	79	11	3118	93,17%
Classe 3	147	262	2374	53	0	2836	83,71%
Classe 4	0	5	33	1867	1	1906	97,95%
Classe 5	0	53	0	0	5273	5326	99,00%
Total	19214	3225	2502	1999	5285		
<i>AP</i>	99,06%	90,08%	94,88%	93,40%	99,77%		

Acurácia Global: 97,60% Coeficiente Kappa:0,96

Tabela 4 – Matriz dos erros Classificação 2 Redes Neurais- Multiclasse.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Total	<i>AU</i>
Classe 1	19078	0	6	0	0	19084	99,97%
Classe 2	14	2901	89	51	165	3220	90,09%
Classe 3	122	231	2374	102	0	2829	83,92%
Classe 4	0	6	33	1846	2	1887	97,83%
Classe 5	0	87	0	0	5118	5205	98,33%
Total	19214	3225	2502	1999	5285		
<i>AP</i>	99,29%	89,95%	94,88%	92,35%	96,84%		

Acurácia Global: 95,31% Coeficiente Kappa:0,95

Tabela 5 – Matriz dos erros Classificação 2 SVM- Multiclasse.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Total	<i>AU</i>
Classe 1	19097	0	2	0	0	19039	99,99%
Classe 2	0	3176	20	39	359	3118	88,37%
Classe 3	117	26	2479	0	0	2836	94,55%
Classe 4	0	23	1	1959	13	1906	98,15%
Classe 5	0	0	0	1	4913	4914	99,98%
Total	19214	3225	2502	1999	5285		
<i>AP</i>	99,39%	98,48%	99,08%	98,00%	92,96%		

Acurácia Global: 98,10% Coeficiente Kappa:0,97

Existe pouca discrepância nos resultados do último experimento, é verificado um melhor desempenho das SVM mas não significativo em relação a classificação Maxver.

## 5. Conclusão

Uma questão importante a ser avaliada é a questão do custo computacional, técnicas clássicas baseadas no princípio de Máxima verossimilhança por exemplo, apresentam um custo computacional muitas vezes menor que as SVMs não lineares, associado a imagens grandes esse custo de processamento pode ter papel decisivo na escolha do princípio de classificação.

As SVMs se revelam para o sensoriamento remoto uma alternativa interessante para problemas relacionados a confusão espectral, porém a maior dependência de um forte aumento da acurácia depende da amostragem, que deve independentemente do método refletir o valor semântico da classe.

A principal vantagem do uso das SVM na classificação de culturas esta na capacidade de definição de uma classe mais complexa de dados, que de forma clássica seriam divididas em sub-classes. Esse tipo de complexidade é muito frequente quando se trata da definição de classes referentes a diferentes estágios de cultura, devido a diferenças espectrais associadas a fatores como umidade do solo e época de plantio e variedade cultivada.

## 6. Referencias

### 6.1 Artigo em periódico

- B. Dixon; N. Candade **Multispectral landuse classification using neural networks and support vector machines: one or the other, or both?** International Journal of Remote Sensing 2008.
- Foody, G. M., 2004, **Supervised classification by MLP and RBF neural networks with and without an exhaustively defined set of classes**, *International Journal of Remote Sensing*, 25, 3091-3104.
- R.-E. Fan, P.-H. Chen, and C.-J. Lin. **Working set selection using the second order information for training SVM**. Journal of Machine Learning Research 6, 1889-1918, 2005.
- Takeuchi, K. e Collier, N. (2004). **Bio-medical entity extraction using support vector machines**. Artificial Intelligence in Medicine. Versão eletrônica disponível em <http://www.elsevierhealth.com/journals/aiim/>.
- Zheng, S., Liu, J., e Tian, J. W. (2004). **A new efficient svm-based edge detection method**. Pattern Recognition Letters, 25(10):1143-1154
- Chavez, G. Camara, M. Cord, F. Precioso, S. Philipp-Foliguet, and A. de A. Araujo, **"Video segmentation by supervised learning,"** in *Computer Graphics and Image Processing, 2006. SIBGRAPI '06*, 2006.

### 6.2 Livro

- Vapnik, V.N. **The Nature of Statistical Learning Theory**. Berlin: Springer-Verlag, 1995 .
- Cristianini, N & Shawe-Taylor, J. **An introduction to support vector machines and other kernel based learning methods**. Cambridge University Press. 2001.
- Suykens A.K., T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, **Least Squares Support Vector Machines**, World Scientific, Singapore, 2002 (ISBN 981-238-151-1)
- Joachims, T. (2002). **Learning to Classify Text using Support Vector Machines: Methods, Theory, e Algorithms**. KLUWER INTERNATIONAL SERIES IN ENGINEERING e COMPUTER SCIENCE. Kluwer Academic Publishers, Norwell, MA.
- Smola, A. J. e Schoolkopf, B. (2002). **Learning with Kernels**. MIT Press.
- Noble, W. S. (2004). Kernel Methods in Computational Biology, chapter **Support vector machine applications in computational biology**, pages 71-92. MIT Press, Cambridge, MA.

### 6.4 Dissertação/Tese

- Baranoski, F. L. **Verificação da autoria em documentos manuscritos usando svm** Dissertação PUC-PR, 2005.
- Souza, Bruno F. de **Seleção de Características em SVMs Aplicadas a Dados de Expressão Gênica**. Dissertação USP. 2005.
- Carvalho, B. P. R. **Novas Estratégias para Detecção Automática de Vetores de Suporte em Least Squares Support Vector Machines**, Dissertação UFMGano: 2005.
- Lin, Chih-Jen. **Libsvm**. support vector machines tool for classification (C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR). Version 1.0 released April 2000. Current Version: 2.88, October 2008.
- Suykens, J.A.K. Van Gestel, T. De Brabanter, J. De Moor, B. Vandewalle, J. **Least Squares Support Vector Machines**, LS-SVMLab. World Scientific, Singapore, 2002 (ISBN 981-238-151-1) 2002.