Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 7917-7924.

Classificação de bioma caatinga usando *Support Vector Machines* (SVM)

Beatriz Fernandes Simplicio Sousa[1]
Adunias dos Santos Teixeira[1]
Francisco de Assis Tavares Ferreira da Silva[2]

[1]Universidade Federal do Ceará - UFC
Caixa Postal 12.168 - 60021-970 - Fortaleza - CE, Brasil
beatrizsimplicio@gmail.com, adunias@ufc.br

[2] Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 21- 61760-000 - Eusébio - CE, Brasil
tavares@roen.inpe.br

Abstract: Support Vector Machines (SVMs) are a kind of neural network and a relatively new supervised classification technique to the land cover mapping community. This paper aims to analyze the performance of SVMs to classify caatinga biome and compare these results with the supervised classification method of maximum likelihood. The general procedure involved the classification from satellite image LANDSAT-TM, bands 2, 3 e 4. The classifications using SVM methods were made using commercial software (ENVI 4.3) and a toolbox in MATLAB 7.0 (SVM-KMToolbox). And the classification using the maximum likelihood were made using ENVI 4.3. The study area was Iguatu, Ceara, Brazil. Iguatu is located at Brazilian caatinga biome. Quantitative Validation has been done by comparison between 137 GPS collected points used as a ground truth and the image obtained of each classification proposed. Using these points, it was generated a confusion matrix to each classification and it was calculated the Kappa coefficient and the overall accuracy. Experimental results allow us to conclude through the Kappa coefficient values that SVM presented a superior performance comparing with the Maximum Likelihood algorithm. Same manner, it is also possible to conclude that the SVM-KMToolbox presented superior performance comparing with the SVM of ENVI 4.3.

Key-words: artificial intelligence, semi arid, satellite image classification, confusion matrix, inteligência artificial, semi-árido, classificação de imagens de satélite, matriz de confusão.

1. Introdução

O mapeamento da cobertura do solo é uma informação essencial em estudos de gestões ambientais, em avaliações de biodiversidade e no apoio a decisão de ações ambientais, sociais e políticas econômicas.

Ações de degradação à natureza, provocadas pelo homem, fazem com que esta tenha reações que prejudicam ao próprio agente causador. Atualmente, esta situação faz com que o homem tenha uma preocupação cada vez maior com a questão da degradação ambiental (OLIVEIRA et al., 2003) e das práticas não-sustentáveis de uso dos recursos naturais assim como tente utilizar os recursos naturais em níveis compatíveis com sua disponibilidade e/ou capacidade de renovação.

A caatinga, um dos mais ricos biomas brasileiros, encontra-se ameaçada. O que agrava a situação é que este bioma, que é exclusivamente brasileiro e ocupa 11% do território nacional, é um dos menos estudados e protegidos do Brasil (WANDSCHEER, 2008).

O trabalho de Ribeiro et al., (2002) afirma que após séculos de gestão desastrosa dos recursos naturais, atingimos estágios onde a manutenção do nosso padrão de vida tornou-se incompatível com o de muitas outras espécies e que uma possibilidade de reverter esta situação talvez esteja no conhecimento sobre a dinâmica dos ecossistemas. Segundo Wandscheer (2008), na caatinga a maior agressão decorre da extração de lenha. No entanto, o conhecimento mais aprofundado sobre este bioma pode tornar possível realizar esta atividade sem destruí-la, pela adoção de um planejamento da exploração sustentável, de forma que a caatinga possa se recompor.

Assim, a busca pelo conhecimento do que existe sobre a superfície do solo através de classificadores de imagens de satélite tem sido objeto de estudo e pesquisa pela comunidade científica (SOUSA, et al., 2007). E são através destas que na literatura é possível encontrar desde métodos clássicos para classificação de imagens de satélite (máxima verossimilhança, mínima distância) como também métodos mais avançados de classificação como, por exemplo, os que utilizam redes neurais artificiais (CARVALHO et al., 2004). Outros métodos de reconhecimento de padrões baseados em inteligência artificial, como o que utiliza máquinas de vetores de suporte (SVM), tem atraído atenção da comunidade de sensoriamento remoto (GIGANDET et al., 2005). As máquinas de vetores de suporte estão presentes na nova geração de sistemas de aprendizado supervisionados baseados na teoria de aprendizagem estatística. Vapnik e colaboradores lideraram os estudos neste novo método e vêm obtendo sucesso em diversos problemas de classificação (GONÇALVES et al., 2006).

No presente trabalho, objetivou-se comparar o desempenho de uma SVM na classificação de bioma caatinga utilizando a SVM-KMToolbox desenvolvida por Canu et al. (2005), para ambiente MATLAB®, e também pelo software comercial ENVI 4.3®. Assim como, comparar estes resultados ao classificador estatístico MAXVER.

2. Metodologia de trabalho

2.1 Área de estudo

A área de estudo foi extraída de uma cena do satélite Landsat-5 (217/65) adquirida em 30/07/2007 às 9 horas. Optou-se por utilizar a combinação das bandas 2 (0,50 - 0,60 µm), 3 (0,63 - 0,69 µm) e 4 (0,76 - 0,90 µm) do espaço RGB. Nesta cena está localizado o município de Iguatu, no Estado do Ceará – Brasil, entre os paralelos 6º 17’ 42,33’’S e 6º 33’ 41,04’’S e os meridianos 39º 3’ 55,64’W e 39º 29’ 28,52’’W como pode ser observado na Figura 1. Esta é uma área ainda carente de estudos mais aprofundados, portanto, pesquisas que busquem melhorias na classificação de áreas que englobam o bioma caatinga são extremamente importantes para vários estudos que dependem da informação do uso e ocupação do solo.

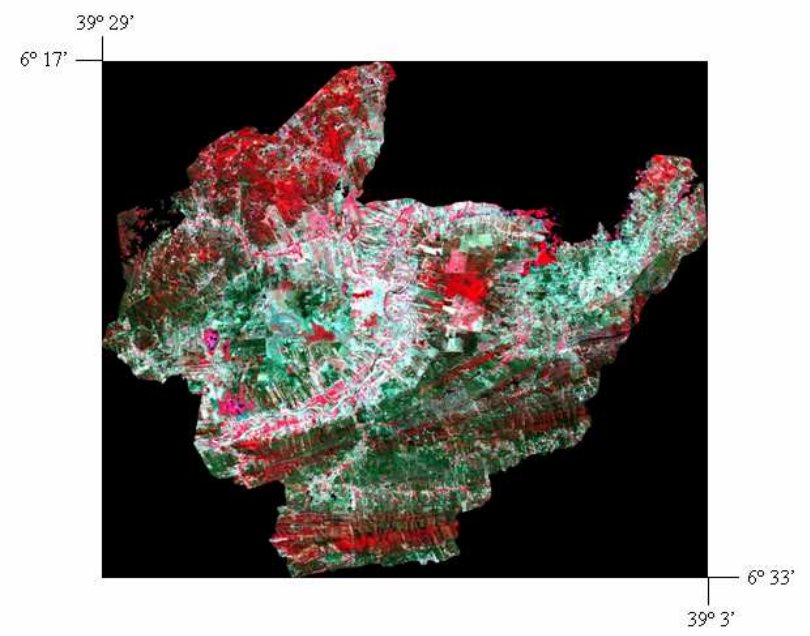

Figura 1. Município de Iguatu-CE.

2.2 Classes de interesse e pontos de validação

A cena foi dividida em cinco classes genéricas: **Classe 1** – Água: rios, açudes e lagoas; **Classe 2** – Antropizada: áreas que sofreram algum tipo de degradação ou áreas descobertas; **Classe 3** – Caatinga Arbórea Densa (CAD): Engloba a vegetação arbórea densa, de porte mais elevado. Nas regiões de serra observa-se uma vegetação com característica mais exuberante, onde as condições climáticas fornecem maior vigor na vegetação. Também está presente nas regiões do interior mais planas e mais secas apresentando uma leve diferença de tonalidade; **Classe 4** – Caatinga Herbácea Arbustiva (CHA): Segundo Fernades e Bezerra (1990) esta vegetação é do tipo xerófila surgindo em áreas com características de semi-aridez. Engloba a vegetação herbácea arbustiva (porte baixo a médio) aberta à densa; **Classe 5** – Agricultura.

Realizou-se o processo de amostragem no software ENVI 4.3® e posteriormente os dados foram exportados para um formato legível do MATLAB 7.0 (*.txt), isto evita tendenciosidade nas comparações. É importante salientar que as amostras foram obtidas na mesma cena mas fora da área de estudo, o que possibilita a generalização pelos classificadores utilizados.

Para a obtenção dos pontos representativos da verdade terrestre, foi realizada uma missão à área de estudo. Utilizou-se aparelho GPS Garmim para coleta de 137 pontos representativos às cinco classes para, através destes, construir a matriz de confusão. A partir desta é possível obter dois parâmetros que qualificam a classificação: a exatidão global e o índice de Kappa, expostos nas Equações 1 e 2, respectivamente. Valores de exatidão específica de cada classe também são obtidos tal qual os de exatidão global (Equação 1), a diferença é que o número de acertos e de pontos amostrais correspondem aos valores da classe.

$$EG = \frac{A}{N} * 100 \quad (1)$$

Onde, EG= exatidão global; A é o número de acertos e N o número de pontos amostrais.

$$K = \frac{N\sum_{i=1}^{\gamma} x_{ii} - \sum_{i=1}^{\gamma} (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{\gamma} (x_{i+} * x_{+i})} \quad (2)$$

Onde,
K = coeficiente Kappa de concordância; **N** = número de observações (pontos amostrais); r = número de linhas da matriz de erro; x_{ii} = observações na linha i e coluna i, respectivamente; x_{i+} = total marginal da linha i; x_{+i} = total marginal da coluna i.

2.3 Classificador SVM

Uma *Support Vector Machine* (SVM) é basicamente uma máquina linear, cuja idéia principal é construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima (HAYKIN, 2001).

Neste método, a partir de um espaço de entrada de padrões não-linearmente separáveis é formado um novo espaço de características, em dimensão outra, onde os padrões serão linearmente separáveis. Assim, um hiperplano de separação ótimo entre os exemplos é construído (VAPNIK, 1995). Isto permite a classificação de dados do sensoriamento remoto que geralmente são não linearmente separáveis no espaço de entrada (GIDUDU, 2007).

As funções usadas para projetar os dados do espaço de entrada para o espaço de alta dimensão são chamadas de kernels. Diferentes kernels têm sido propostos na literatura, são eles: lineares, polinomiais, gaussianas (mais comumente chamadas de funções de bases radial) e sigmóides. Diferentes definições da função Kernel e seus respectivos parâmetros provocam alterações nos resultados fornecidos por uma SVM.

A classificação em um espaço de alta dimensão resulta em *overfitting* nos dados de entrada, porém, nas SVMs isso é controlado através do princípio da minimização estrutural de risco (VAPNIK, 1995).

O hiperplano ótimo, que diminui classificações errôneas, pode ser obtido pela otimização do problema quadrático através da teoria Lagrangiana. Assim, é preciso encontrar os multiplicadores Lagrangianos que maximizem a seguinte função objeto (Equação 3):

$$\sum_{i=1}^{n} \lambda i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \lambda_i \lambda_j y_i y_j \ker nel(x_i x_j) \quad (3)$$

Sujeito a:

$$\sum_{i=1}^{n} \lambda i y_i = 0 \text{ e } 0 \le \lambda_i \le C$$

Onde λi e λj são os multiplicadores de Lagrange e C é um termo de regularização que penaliza classificações errôneas.

Classificadores SVM possuem essencialmente técnicas de classificação binárias. Assim, quando o problema de classificação possui mais de duas classes pode-se utilizar estratégias que permitam esta classificação como: *one-against-one*(1A1) e *one-against-all*(1AA). A estratégia 1A1 constrói uma máquina para cada par de classes. O número de máquinas é obtido através da fórmula N(N-1)/2, onde N é o número de classes (GIDUDU, 2007). Esta estratégia foi escolhida para realizar a classificação.

O entendimento matemático deste método pode ser aprofundado nos trabalhos de Vapnik (1995) e Haykin (2001).

No software MATLAB 7.0 foi utilizada a SVM-KMToolbox (*SVM and Kernel Methods MATLAB Toolbox)* desenvolvida por Canu et al. (2005) com C=1000 e λ=0.0000001. A função Kernel utilizada foi a Gaussiana. Um Kernel $K(x_i,x_j)$ é o produto interno em algum espaço de características, $K(x_i, x_j)= \Phi(x_i)\ \Phi\ (x_j)$. A representação desta função matemática está apresentada da Equação 4:

$$K(x, x_i) = \exp\left(\frac{-\left\| x_i - x_j \right\|^2}{\sigma^2} \right) \quad (4)$$

O software ENVI 4.3 dispõe ao usuário de quatro tipos de funções Kernels: linear, polinomial, sigmóide e função de base radial (RBF). Optou-se por utilizar esta última. Ao fazer uso desta função, o software permite ao usuário que este tenha controle sob os parâmetros Gamma e C, aos quais foram atribuídos os valores 0,33 (valor atribuído através das instruções do manual do ENVI que é o valor correspondente a 1/n, onde n é o número de bandas utilizadas) e 100, respectivamente.

3. Resultados e Discussão

O índice é de Kappa amplamente utilizado na literatura para qualificar uma classificação (NISHIDA,1998; QUEIROZ, 2004). Um fator importante a ser observado é que os pontos utilizados na construção deste índice devem estar fora dos pontos que serviram de amostras para o treinamento, caso contrário, pode-se obter valores de Kappa superestimados. As Tabelas 1 e 2 correspondem às matrizes de confusão geradas a partir dos pontos coletados com aparelho GPS e comparados às classificações pelo método SVM no ENVI 4.3 e na SVM-KMToolbox, respectivamente.

Tabela 1. Matriz de confusão obtida a partir de classificação SVM no ENVI 4.3

		Verdade de Campo						
		Água	Agricultura	Antropizada	CHA	CAD	Total Linhas	Exatidão Específica (%)
Classificação	Água	3					3	100
	Agricultura		19	1	1	1	22	86,4
	Antropizada			39		3	42	92,9
	CHA		4	7	34	11	56	60,7
	CAD		4			10	14	71,4
	Total Colunas	3	27	47	35	25	137	

Tabela 2. Matriz de confusão obtida a partir de classificação na SVM-KMToolbox

		Verdade de Campo						
		Água	Agricultura	Antropizada	CHA	CAD	Total Linhas	Exatidão Específica (%)
Classificação	Água	3					3	100
	Agricultura		19	2	1		22	86,4
	Antropizada			42			42	100
	CHA		4	12	39	1	56	69,6
	CAD		5			9	14	64,3
	Total Colunas	3	28	56	40	10	137	

Os valores de EG obtidos através da Equação 1 foram de 76 e 81% para a classificação usando SVM no ENVI e na SVM-KMToolbox, respectivamente. Desta forma, pode-se concluir que o último método obteve maior índice de pixels classificados corretamente do que o primeiro.

A classificação SVM pelo ENVI obteve índice de Kappa de 0,68. Este mesmo índice para a classificação utilizando a SVM-KMToolbox foi de 0,74. Os valores do coeficiente de Kappa são considerados, de acordo com a classificação proposta por Landis e Koch (1977), como boa para o ENVI e muito boa para a SVM-KMToolbox.

Os resultados apresentados pelo software ENVI foram obtidos ao utilizar valor de C e Lambda que o software tem como padrão. Ao testar valores de C e lambda iguais ao da SVM-KMToolbox o coeficiente de Kappa reduziu para 0,65.

A classificação supervisionada pelo método da máxima verossimilhança é um método amplamente utilizado no meio científico (QUEIROZ, 2005; SOUSA, 2007). Desta forma, este pode ser utilizado como referência às novas classificações testadas. A Tabela 3 expõe o resultado da matriz de confusão obtida pela MaxVer.

Tabela 3. Matriz de confusão obtida a partir do método de classificação MaxVer.

		Verdade de Campo						
		Água	Agricultura	Antropizada	CHA	CAD	Total Linhas	Exatidão Específica (%)
Classificação	Água	3					3	100
	Agricultura		17	2		3	22	77,3
	Antropizada		4	37	1		42	88,1
	CHA		6	13	37		56	66,1
	CAD		3		2	9	14	64,3
	Total Colunas	3	30	52	40	12	137	

Este método apresentou K=0,65 o que, segundo a classificação Landis e Koch (1977), qualifica este resultado como bom. Observando os valores de Kappa obtidos a partir da classificação SVM é possível concluir que o método da máxima verossimilhança apresentou resultado inferior a estes. O valor de EG encontrado foi de 75%, sendo este também inferior aos obtidos no método SVM.

Pela observação dos valores de exatidão específica obtidos para a classe CHA (Tabelas 1, 2 e 3), nota-se que a SVM-KMToolbox, quando comparada aos demais métodos, apresentou valor superior com 69,6% dos pixels classificados corretamente. Já para a classe caatinga herbácea arbustiva (CAD) o método SVM aplicado pelo software ENVI apresentou resultado superior. Porém, observando a Tabela 1 também é possível notar que a classe CAD possui vários pixels atribuídos a outras classes, o que não ocorre quando é utilizada a ferramenta SVM-KMToolbox.

4. Conclusões e Perspectivas

Os valores de exatidão global (EG) e coeficiente de Kappa (K) obtidos revelam o alto potencial das SVMs na classificação de dados do sensoriamento remoto.

Para este experimento de classificação de bioma caatinga recomenda-se o uso da SVM-KMToolbox, já que esta apresentou índices de EG e K superiores às demais classificações estudadas.

Softwares comerciais devem apresentar capacidade de testes de generalização, ou seja, devem ser capazes de treinar um algoritmo para determinada imagem e aplicar o conhecimento obtido em outras imagens. No software comercial testado, esta opção ainda não é oferecida ao usuário, o qual fica limitado a produzir uma classificação com amostras de interesse contidas na imagem.

O trabalho apresenta parte do desenvolvimento de uma metodologia de pesquisa em busca por melhores dados de validação, principalmente no que concerne aos novos paradigmas de classificação. No entanto, o trabalho também objetiva consolidar referências de medições, às quais permitam montar uma estação piloto de aferição de sensores de solo, orbitais e aerotransportados, utilizando-se sensores certificados e georeferenciados, numa

forma de auxiliar também a calibração da resposta espectral dos atuais e futuros satélites, principalmente, quando aplicados ao sensoriamento remoto do semi-árido.

Espera-se também que estas ferramentas permitam automação do monitoramento e quantificação da presença deste bioma e, por conseqüência, maior agilidade na produção de informações que subsidiem a adoção de práticas conservacionistas.

A proposta deverá evoluir para o desenvolvimento de sistemas automáticos de classificação e validação, através da coleta de dados de sensores certificados, de forma que, a partir de referências de dados cada vez mais confiáveis, seja possível indicar a potencialidade de novas ferramentas computacionais no processo de classificação de imagens de satélite quem contemplem às condições semi-áridas. Pretende-se também desenvolver novos classificadores específicos e adequados à realização da classificação de diferentes fitofisionomias de caatinga.

5. Agradecimentos

Às instituições que apoiaram esta pesquisa: Universidade Federal do Ceará e Instituo Nacional de Pesquisas Espaciais (INPE) e ao CNPq pelo apoio financiamento desta. Ao Dr. Rakotomamonjy por disponibilizar a SVM-KMToolbox para este estudo. Aos Engenheiros Agrônomos Clênio Jario, Alexandre Gomes Costa e Amaury pela colaboração na exaustiva coleta dos pontos no campo.

6. Referências bibliográficas

Canu, S.; Grandvalet, Y.; Guigue, V.; Rakotomamonjy, A. **SVM and Kernel Methods MATLAB Toolbox**. Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005.

Carvalho, L.M.T.; Clevers, J.G.P.W.; Skidmore, A. K.; Jong, S.M. Selection of imagery data and classifiers for mapping Brazilian semideciduous Atlantic forests. **International Journal of Applied Earth Observation and Geoinformation**, v. 5, n. 3, p. 173-186, 2004.

Fernandes, A.; Bezerra, P. Esquema Fitogeográfico (províncias): província nordestina ou das caatingas. **Estudo Fitogeográfico do Brasil.** Fortaleza: Stylus comunicações, 1990. 184 p.

Gidudu A.; Hulley G.; Marwala T. Image Classification Using SVMs: One-against-One Vs One-against-All. In: Asian Conference on Remote Sensing (ACSR), 28, 2007, Kuala Lumpur. **Proccedings**... Singapore: ACSR, 2007. Disponível em: < http://arxiv.org/abs/0711.2914v1>. Acesso em: 05 nov. 2008.

Gigandet, X.; Cuadra, M. B.; Pointet, A.; Cammoun, L.; Caloz, R.; Thiran, J. Region-based satellite image classification: method and validation. **Image Processing,** v. 3. p. 832-835, 2005.

Gonçalves, P.; Carrão, H.; Pinheiro, A; Caetano, M. Land cover classification with Support Vector Machine Applied to MODIS imagery. In **Global Developments in Environmental Earth Observation from Space**, A. Marçal (Ed.), pp. 517-526 (Rotterdam: Millpress), 2006.

Haykin, S. **Redes Neurais: princípios e prática**. Porto Alegre: Bookman, 2001. 900 p.

Landis, J.R.; Koch, G.C. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159-14, 1977.

Nishida, W.; Bastos, L. C. Classificação de Imagens de Sensoriamento Remoto Utilizando uma Rede Neural Artificial com Função de Base Radial. In: Simpósio Brasileiro de Sensoriamento Remoto (SBSR), 9., 1998, Santos. **Anais**... José dos Campos: INPE, 1998. Artigos, p. 991-1001. CD-ROM, On-line. ISBN 85-17-00015-3. Disponível em: <http://marte.dpi.inpe.br/col/sid.inpe.br/deise/1999/02.11.11.58/doc/8_122p.pdf>. Acesso em: 05 nov. 2008.

Oliveira, S. B. P.; Souza, M. J. N.; Leite, F. R. B.; Costa, R. N. S. Contribuição ao estudo da degradação ambiental no município de Caridade-CE. In: Simpósio Brasileiro de Sensoriamento Remoto (SBSR), 11, 2003,

Belo Horizonte. **Anais**... São José dos Campos: INPE, 2003. Artigos, p. 1391 - 1398. CD-ROM, On-line. ISBN 85-17-00017-X. Disponível em: <http://urlib.net/ltid.inpe.br/sbsr/2002/11.11.09.39>. Acesso em: 02 nov. 2008.

Queiroz, R. B.; Severino, P. A. R.; Rodrigues, A.G.; Gómez, A.T.U. Redes Neurais: Um comparativo com Máxima Verossimilhança Gaussiana na Classificação de Imagens CBERS. In: Workshop de Tecnologia da Informação aplicada ao Meio Ambient (WIWA), 2, 2004, Itajaí. **Anais...** Universidade Vale do Itajaí: UVI, 2004. Artigos, p. 746 a 749. CD-ROM, On-line. ISBN 1677-2822. Disponível em: < http://www.niee.ufrgs.br/cbcomp/cbcomp2004/html/pdf/Workshop_Ambiente/Intelig%eancia_Artificial/t170100151_3.pdf>. Acesso em: 05 nov. 2008.

Ribeiro, C. A. A. S.; Varella, C. A. A.; Sena Júnior, D. G.; Soares, V. P. Sistemas de Informações Geográficas. In: Borém, A.; M. P. del Giúdice; Daniel Marçal de Queiroz; Evandro Chartuni Mantovani; Lino Roberto Sousa, B. F. S.; Teixeira, A. dos S.; Leão, R. A. de O.; Filho, A. B. C. Uso do solo da bacia hidrográfica do Alto Piauí através de imagens do satélite CBERS. **Revista Centro de Ciências Agrárias**, v. 38, p. 327-334, 2007.

Vapnik, V.; Cortes, C. Support-Vector Networks. **Machine Learning**, v. 20, p. 273-297, 1995.

Wandscheer, L. Caatinga está sendo destruída mais rápido do que a Amazônia**. Agência Brasil,** Brasília, 26 out 2008. Disponível em: <http://www.agenciabrasil.gov.br/noticias/2008/10/29/materia.2008-10-29.7369241684/view>. Acesso em: 29 out.2008.