

Teste do Classificador “Support Vector Machine” (SVM) para a Avaliação da Cobertura das Terras em Imagens Capturadas do Software Google Earth®

EDUARDO DA SILVA⁽¹⁾, LUCIANO ACCIOLY⁽²⁾, ALEX ARAÚJO⁽³⁾, ADEMAR DA SILVA⁽²⁾
RAMON CANTALICE⁽⁴⁾, & JULIANA DA SILVA⁽⁵⁾

RESUMO - O uso e a cobertura do solo respondem pelo impacto das atividades humanas sobre o meio ambiente. Neste estudo, utilizou-se a classificação supervisionada pelos métodos “support vector machine” (SVM) e mínima distância (MD) para gerar mapas da cobertura das terras de uma área (117 ha), da bacia do rio Ipojuca (PE), a partir de imagens de alta resolução capturadas do software Google Earth®. Os índices de acurácia global e kappa foram 73% e 0,63 e 69% e 0,57 para os métodos SVM e MD, respectivamente, indicando uma melhor qualidade do mapa gerado pelo classificador SVM. No entanto, a acurácia por classe variou entre métodos e entre classes dentro de cada método, sendo esta variabilidade atribuída à tendência na amostragem dos dados de referência. As próximas etapas do trabalho envolverão a interpretação visual de toda imagem e checagem de campo mais intensa visando a melhoria dos mapas produzidos por estes classificadores.

Palavras-Chave: (classificação supervisionada; mínima distância; Quickbird)

Introdução

O uso e a cobertura do solo respondem pelo impacto das atividades humanas sobre o meio ambiente, constituindo informação indispensável para o planejamento de atividades relacionadas com o uso dos recursos naturais. Além desse aspecto, sua interpretação serve de base, também, para estudos socioeconômicos, avaliação de desvios entre aptidão das terras e o uso efetivo, avaliação de riscos de erosão, entre outros. Tradicionalmente, os mapas de uso e cobertura são realizados por meio da calibração de curvas espectrais extraídas de imagens de satélite com base em dados de campo. A partir dessa calibração são utilizados algoritmos de classificação através dos quais as imagens são transformadas em mapas onde para cada pixel é assinalada uma classe de uso. Recentemente, o algoritmo não paramétrico de classificação supervisionada conhecido por *Support*

Vector Machine (SVM) foi incorporado a vários softwares para o processamento, principalmente, de imagens complexas e com alto nível de ruído (HSU et al.[1]).

Embora algumas imagens de sensores multiespectrais para o mapeamento em escala regional (1:50.000 ou menor), como as dos sensores a bordo do satélite CBERS, estejam disponíveis de forma gratuita na Internet, as imagens para mapeamento em escalas maiores como as do satélite Quickbird, ainda representam um alto custo para a realização de qualquer projeto. No entanto, imagens de alta resolução podem ser visualizadas e capturadas a partir de softwares de domínio público como, por exemplo, o Google Earth®. Neste trabalho, um segmento de imagem do Quickbird cobrindo uma pequena parte do terço inferior da bacia do Rio Ipojuca (PE) foi capturado e classificado pelo método SVM. Os resultados dessa classificação foram comparados ao método paramétrico de classificação supervisionada da mínima distância (MD). Os resultados aqui apresentados são parciais, pois não envolvem comparação com a aplicação dos classificadores à imagem original.

Material e Métodos

A área de estudo, com 117,4 ha (Figura 1), tem seus limites dados pelas coordenadas 35°1'1,65"W 8°24'2,24"S e 35°0'15,03"W 8°24'29,02"S, localizando-se nas proximidades da foz do Rio Ipojuca. Esta área apresenta alguns dos usos e coberturas representativas da Zona da Mata e do Litoral Pernambucano como a cana-de-açúcar, em diferentes estágios de desenvolvimento, pequenas propriedades com fruteiras, remanescentes de mata Atlântica em diversos estágios sucessionais e manguezais. As atividades agrícolas e a vegetação nativa, desta área, encontram-se sob forte pressão devido aos recentes investimentos para ampliação do parque industrial do porto de Suape.

Neste estudo foram utilizadas imagens do satélite Quick Bird com data de passagem 26 de agosto de 2006, capturadas do software de domínio público Google Earth® em abril de 2009. As imagens foram salvas em três bandas (RGB) no formato tiff e importadas para o software ENVI

⁽¹⁾Estudante de graduação do Departamento de Eletrônica e Sistemas, Centro de Tecnologia e Geociências, Universidade Federal de Pernambuco. Av. Acadêmico Hélio Ramos, S/N, Cidade Universitária, Recife, PE, CEP 50740-530. E-mail: eduardo_alves_silva@yahoo.com.br.

⁽²⁾Pesquisador da EMBRAPA Solos UEP Nordeste. Rua Antônio Falcão, 402, Boa Viagem, Recife, PE, CEP 51020-240. ljoaccioly@hotmail.com

⁽³⁾Professor do Departamento de Engenharia Mecânica, Centro de Tecnologia e Geociências, Universidade Federal de Pernambuco.

⁽⁴⁾Professor do Departamento de Agronomia, Campus de Dois Irmãos, Universidade Federal Rural de Pernambuco. Rua Dom Manoel de Medeiros, S/N, Dois Irmãos, Recife, PE, CEP 52171-900. E-mail: cantalic@terra.com.br.

⁽⁵⁾Estudante de graduação do Centro de Ciências Biológicas, Universidade Federal de Pernambuco. Av. Professor Moraes Rego, S/N - Cidade Universitária, Recife, PE, CEP: 50670-420. E-mail: juli.alvess@gmail.com.

Apoio financeiro: CNPq (P. 479950/2007-5)

(SULSOFT. 2006)[2], onde foram georeferenciadas para projeção UTM. A imagem foi classificada utilizando-se a classificação supervisionada pelo método da mínima distância (MD), e pelo método “support vector machine” (SVM). Embora o método da máxima verossimilhança (MAXVER) seja, entre os métodos de classificação supervisionada paramétricos, o mais utilizado como padrão de comparação com outros classificadores, o reduzido número de amostras de treinamento para algumas classes justifica a adoção do método MD na comparação com o método SVM. O SVM é um sistema de classificação supervisionado, não paramétrico, que tem suas raízes na teoria de aprendizado estatístico (“statistical learning theory”) (HSU et al.[1]). No SVM as classes são separadas por hiperplano que maximiza a margem de decisão entre as mesmas. Os pontos mais próximos do hiperplano são conhecidos como vetores de suporte (“support vectors”) e são relevantes como elementos de treinamento do algoritmo (HSU et al.[1]). Embora conceitualmente desenvolvido como um classificador binário, no software ENVI ele foi implementado para multiclases por meio do algoritmo “pairwise classification”. Como produto de saída, o SVM produz valores de decisão para cada píxel em cada uma das classes. Estes valores são utilizados para estimativas de probabilidade que são armazenadas numa imagem utilizada como “rule image”. A classificação é realizada selecionando-se os valores mais altos de probabilidade. O método da mínima distância utiliza o vetor que representa a média de cada amostra de treinamento de cada classe para agrupar os demais pixels em torno destas médias com base, em geral, na distância euclidiana.

Os classificadores MD e SVM foram aplicados usando-se as classes: cana abandonada, cana adulta, manguezal e remanescente de mata Atlântica. Foram utilizados os parâmetros padrões do SVM presentes no ENVI, ou seja: o parâmetro de penalidade (“penalty parameter”), que permite certo grau de confusão entre classes de baixa separação, foi deixado em 100 (todos os pixels foram forçados a pertencer a uma das classes); o núcleo da função matemática (“kernel type”) utilizada na classificação foi do tipo “radial basis function” (RBF) com o termo gama desta função mantido em 0,333. A RBF tem sido sugerida como a primeira escolha ao se empregar o classificador SVM pelas seguintes razões [3]: a) o “kernel” mapeia, de forma não linear, as amostras em um espaço multidimensional; b) outros tipos de “kernel”, como por exemplo, o linear e o sigmóide são formas particulares do RBF; c) as dificuldades no processamento numérico dos dados são menores. O método MD foi aplicado, também, de forma que todos os pixels da imagem pertencessem a uma das classes citadas.

Amostras de cada uma das classes foram identificadas nas imagens, sendo parte delas utilizadas como amostras de treinamento, para a execução dos

métodos de classificação e parte usada como dados de referência para a medição da acurácia dos classificadores. Até o momento, poucos dados de campo foram coletados e a vetorização das classes, com base na interpretação visual, também, não foi concluída. A acurácia foi realizada por meio da geração da matriz de confusão. Por meio desta matriz foi possível avaliar a acurácia pelo método global e pelo índice Kappa. Um mapa final foi gerado para cada classificador, contendo as seguintes classes: cana-de-açúcar abandonada, cana-de-açúcar em estágio de desenvolvimento adulto, Árvores, mangues e solo exposto. A cana abandonada tem forte contribuição de diversas invasoras constituídas, principalmente, pelo capim brachiaria (*Brachiaria decumbens*). A classe Árvores inclui espécies de diversas naturezas dentre as quais remanescentes de mata Atlântica.

Resultados

A Figura 2 apresenta os mapas referentes à classificação pelos métodos MD e SVM, respectivamente, enquanto a Tabela 1 apresenta as áreas ocupadas por cada classe para cada classificador. As estimativas de acurácia global da classificação e do índice Kappa (K) foram 69% e 0,57 para o método MD e 73% e 0,63 para o método SVM, respectivamente. As Tabelas 2 e 3 apresentam a matriz de confusão obtida para o classificador MD e SVM. Os erros de omissão e inclusão, expressos em porcentagem, são apresentados na Tabela 4.

Discussão

Na Figura 2, observa-se que o delineamento do contorno das maiores áreas contínuas de cada classe tem, de uma forma geral, uma boa concordância com os alvos que representam estas classes assinaladas na Figura 1. No entanto, quando comparado ao classificador MD, o classificador SVM apresentou uma maior homogeneidade nessas áreas. Por exemplo, as áreas de mangue, que têm uma vegetação típica dos ambientes alagados por água salobra, onde as possibilidades de se encontrar outros tipos de vegetação não adaptadas a estas condições são pequenas, aparece mesclada com a classe árvore no mapa do classificador MD. O classificador MD não apresenta bom desempenho para classes que apresentam variância elevada e valores próximos para as médias (Lillesand et al. [3]) como é o caso das classes mangue e árvores. A mesma interpretação pode ser dada para as classes cana abandonada e cana adulta, cuja homogeneidade é maior no classificador SVM. Numa verificação visual mais detalhada, observa-se que a distribuição espacial das classes nas áreas menores apresenta diferenças significativas, principalmente, na parte norte dos mapas (Figura 2). Essas diferenças se refletem na área ocupada por cada classe quando se compara os dois classificadores (Tabela 1). Quando comparado ao método MD, a classificação SVM apresentou áreas menores para as classes cana adulta e cana abandonada e maiores para as demais classes. Em termos absolutos, as maiores diferenças foram para as áreas das classes mangue e

árvores. Em termos relativos, houve uma variação de 75% para a classe cana adulta e de 185% para a área da classe solo exposto (Tabela 1).

Ainda com base na interpretação visual da Figura 1, observa-se um melhor desempenho do classificador SVM na quantificação das áreas de cada classe. Tais resultados são devido aos distintos critérios utilizados pelos dois classificadores para a separação de classes. A acurácia global e o índice Kappa, que dão uma indicação da qualidade dos mapas gerados, foram maiores para o classificador SVM do que para o classificador MD. Estas duas medidas, no entanto, não representam o desempenho desses métodos com relação à acurácia de classes específicas (JENSEN, [4]). Para avaliar a acurácia do classificador com relação a cada classe é necessário interpretar os dados da matriz de confusão (Tabelas 2 e 3). Nesta matriz, a soma das colunas expressa o número de pixels (cada pixel representando uma área de 8,18 m²) que, teoricamente, representa a “verdadeira” área de cada classe, enquanto o total de cada linha representa o número de pixels assinalados a uma dada classe pelo classificador. A matriz de confusão é, portanto, uma maneira eficiente de avaliar o erro do classificador para cada classe, descrevendo com clareza os dois tipos de erro; o de inclusão e o de omissão que são apresentados em termos percentuais na Tabela 4. Os maiores erros encontrados foram para as classes cana abandonada e árvores. No caso da cana abandonada, do total de 1.850 pixel, apenas 536 e 641 foram classificados corretamente pelos classificadores MD e SVM, respectivamente. O que significa que 71% dos pixels que deveriam ter sido incluídos na classe cana abandonada, no classificador MD, foram assinalados em outras classes (71% de erro de omissão). Para o classificador SVM, o erro de omissão foi um pouco menor (65,3%). A classe árvore também apresentou baixa acurácia. Em imagens de alta resolução espacial, devido ao seu detalhamento, vários fatores podem afetar o valor digital de um pixel de uma dada classe. Estes fatores tornam as amostras de treinamento menos homogêneas, reduzindo a acurácia do classificador para estas classes. Assim, por exemplo, fatores como a geometria e proximidade das copas, altura e densidade podem ter afetado a homogeneidade da amostra de treinamento da classe árvore, reduzindo a sua acurácia (Treitz et al.[5]). O desempenho de ambos classificadores foi excelente para a classe solo exposto e bom para as classes mangue e cana adulta

quando são considerados tanto o erro de omissão quanto o de inclusão. Para a classe mangue, o classificador SVM teve um notável desempenho, omitindo desta classe menos de 2% dos pixels de referência. De um modo geral, no entanto, considerando os dados apresentados nas Tabelas 2, 3 e 5, pode-se inferir que a seleção das áreas para os dados de referência não foi adequada para as classes com maior erro. Tendências nos dados de referência podem responder pela baixa acurácia destas classes (JENSEN [4]). Desta forma, sugere-se, para as próximas etapas do trabalho, que os dados de referência sejam provenientes de um maior número de verdades de campo, auxiliando, portanto, para uma melhor interpretação visual, devendo esta ser seguida da vetorização em tela das classes. Deve-se, também, adequar o número de pixels por amostra de treinamento para possibilitar o emprego do classificador MAXVER nas comparações com o SVM.

Conclusões

O classificador SVM apresentou uma melhor correspondência com a interpretação visual da imagem capturada do software Google Earth®.

A qualidade geral dos mapas foi melhor para o classificador SVM.

Os dois métodos apresentaram alta variabilidade na acurácia entre classes.

Sugere-se, para as próximas etapas do trabalho, que os dados de referência sejam provenientes de um maior número de verdades de campo, evitando-se tendências nas estimativas de acurácia entre classes.

Referências

- [1] HSU, C.; CHANG, C. & LIN, C. 2009 [Online]. *A practical guide to support vector classifications*. Homepage: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- [2] SULSOFT. 2006. *Curso básico do software ENVI*. Apostila. 86p.
- [3] LILLESAND, T. M.; KIEFER, R. W. & CHIPMAN, J. W. *Remote Sensing and Image Interpretation*, New York: Wiley. 763p.
- [4] JENSEN, J. R. *Introductory digital image processing – A remote sensing perspective*, New Jersey: Prentice Hall. 318p.
- [5] TREITZ, P. M.; HOWARTH, P. J. & SUFFLING, R. C. 1992. Application of detailed ground information to vegetation mapping with high spatial resolution digital imagery. *Remote Sens. Environ.*, 42: 65-82.

Tabela 1. Área, em hectares, por classe para cada classificador.

Classe	Classificador MD	Classificador SVM	% em relação ao método MD
Cana Adulta	51,1	38,4	75
Cana Abandonada	24,7	20,1	81
Mangue	23,5	31,1	132
Árvores	15,4	23,0	149
Solo Exposto	2,6	4,8	185

Tabela 2. Matriz de erro do classificador MD.

Classificação	Dados de Referência					Total
	Mangue	Cana Adulta	Cana Abandonada	Árvores	Solo Exposto	
Mangue	1786	0	0	236	0	2061
Cana adulta	0	2741	1226	211	0	3080
Cana abandonada	0	339	536	34	38	717
Arvores	447	0	88	500	0	2286
Solo Exposto	0	0	0	0	307	345
Total	2233	3080	1850	981	345	8489

Tabela 3. Matriz de erro do classificador SVM.

Classificação	Dados de Referência					Total
	Mangue	Cana Adulta	Cana Abandonada	Árvores	Solo Exposto	
Mangue	2202	0	0	404	0	2606
Cana adulta	0	2462	900	24	0	3386
Cana abandonada	0	223	641	33	5	902
Arvores	31	395	309	520	0	1255
Solo exposto	0	0	0	0	340	340
Total	2233	3080	1850	981	345	8489

Tabela 4. Erros de inclusão e omissão, em porcentagem, de acordo com o classificador.

Classe	Inclusão		Omissão	
	DM	SVM	DM	SVM
Mangue	11,45	15,5	20,02	1,39
Cana adulta	46,66	27,29	11,01	20,06
Cana abandonada	57,32	28,94	71,03	65,35
Arvores	23,40	58,57	49,03	46,99
Solo exposto	0,00	0,00	11,01	1,45

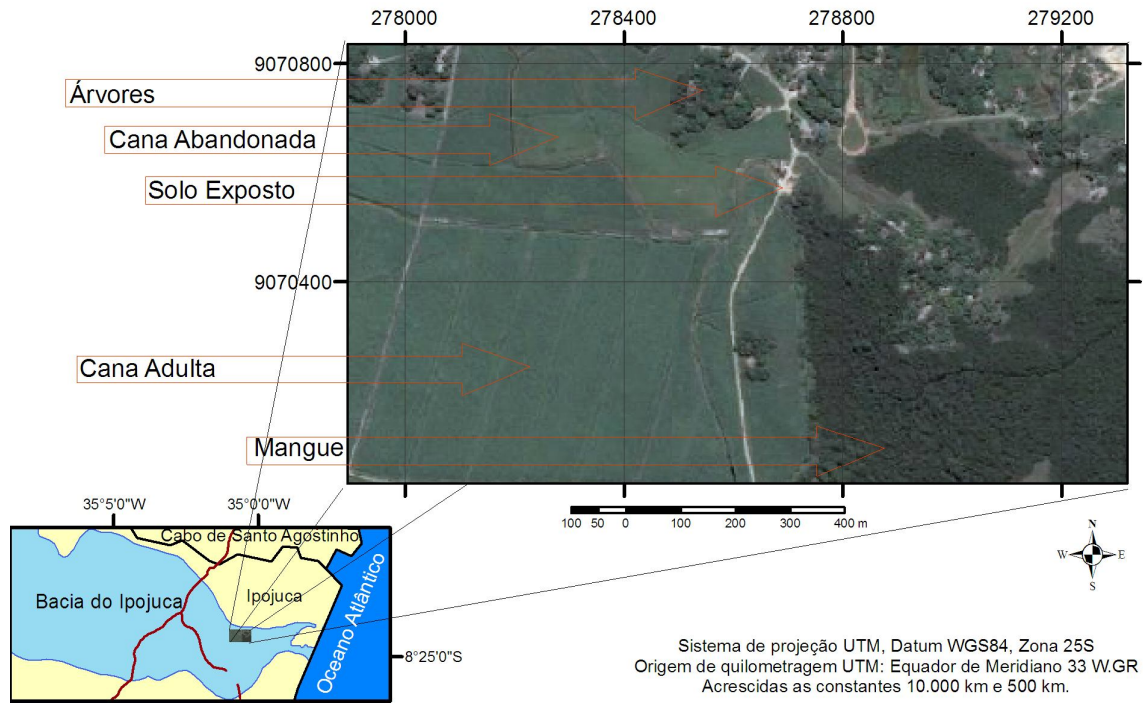


Figura 1. Localização e imagem de alta resolução, capturada do Google Earth®, mostrando as classes de cobertura da área de estudo.

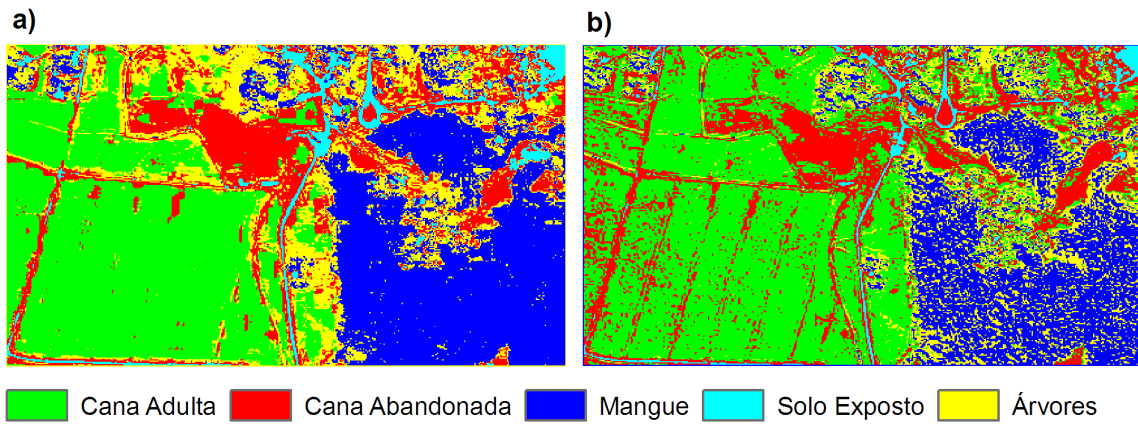


Figura 2. Imagens classificadas segundo os métodos SVM (a) e MD (b).