Avaliação da influência do número de amostras de treinamento no índice Kappa

Maola Monique Faria [1]
Ligia Tavares de Souza [1]
Elpidio Inácio Fernandes Filho [1]
Márcio Rocha Francelino [1]

[1]Universidade Federal de Viçosa – UFV
Avenida Peter Henry Rolfs, s/n, Campus Universitário, Viçosa – MG, Brasil
{maola.faria, ligia.souza, elpidio, marcio.francelino}@ufv.br

Abstract. This paper aims to assess the effect of the number of training variables on the Kappa index from four evaluators: Logit, Neural Network, SVM and Random Forest. The study used a Landsat 8 scene cropped to the municipal boundaries of Matas de Minas (MG). From the Landsat 8 bands it was generated the NDVI and EVI and the principal components of the image. It was collected a variable set composed by 83.347 points randomly distributed in the area of study, covering all the interested classes to the study. These extracted the band values from the Landsat 8 image, from the NDVI and EVI, and from the principal components. The tests were madde with 8 sizes of variables set: 20, 40, 60, 80, 100, 150, 200 e 500. A routine was implemented on the R environment that divided the set archive accordingly with the sizes of variable set evaluated and processed the classifications with each one of the algorithms evaluated: Logit Random Forest, Neural Network and SVM. The validation of the model used 150 variables to each class. The Random Forest algorithm presented less sensitivity to the size variation of the training variable set. On the other hand, the SVM presented higher sensitivity and the worst result with a fewer number of variables.

Palavras-chave: Logit, SVM, Random Forest, Neural Network**.** Logit, SVM, Random Forest, Rede Neural.

1. Introdução

O procedimento de classificar e agrupar os pixels de uma cena com base em suas características espectrais utilizando algoritmos em um programa computacional é denominado por Richards e Jia (1999) como classificação de imagens. Esse procedimento pode-se dar a partir da interferência do analista no treinamento do algoritmo ou não, caracterizando, respectivamente, a classificação supervisionada e a não supervisionada. A classificação supervisionada requer que o analista treine o algoritmo com base na coleta de amostras das diferentes classes de usos de interesse no estudo em áreas homogêneas, para que com base nessas o mesmo identifique os pixels espectralmente similares aos das amostras.

Na fase de treinamento dos classificadores, atenção especial deve ser dada à seleção e coleta de amostras de treinamento, tanto no nível de qualidade representativa das classes de uso e cobertura presente na cena, como que quantitativamente. Tso e Mather (2009) afirmam que o tamanho das amostras é importante para a determinação da acurácia dos parâmetros estatísticos que descrevem as classes a serem obtidas. Os mesmos afirmam também que o número de amostras de treinamento relaciona-se diretamente com o intervalo de confiança das estimativas de acurácia de uma classificação, e com os parâmetros estatísticos estimados utilizados pelos algoritmos de classificação. Porém, o processo de coleta de amostras é oneroso e caro, pois exige que o tamanho da amostra seja mantido a um mínimo que assegure uma boa exatidão do mapa produzido (Congalton, 2009).

Nos primeiros estudos em sensoriamento remoto aplicado à classificação de uso do solo, pesquisadores utilizavam uma equação com base na distribuição binominal para a definição do número de amostras por classe. A desvantagem dessa técnica é que ela não considera na definição do número de amostras a geração da matriz de erro (Congalton, 2009). Diante disso, é recomendável a utilização da distribuição multinominal para o cálculo do tamanho do conjunto de amostras (Tortora, 1978).

Dessa forma, o número de amostras passa a ser calculado com base no número de classes de uso do solo a serem mapeados, a exatidão desejada e a abrangência da classe de maior interesse (Congalton, 2009).

Diante do exposto, o presente trabalho tem por objetivo avaliar o efeito do número de amostras de treinamento no índice Kappa obtido a partir de quatro classificadores: Logit, Rede Neural, SVM e Random Forest.

2. Metodologia de Trabalho

1.1. Caracterização da área de estudo

A região das Matas de Minas localiza-se na porção sudeste do Estado de Minas Gerais (Figura 1), sendo composta por 61 municípios espalhados por uma área de 1.749.114 ha (3% de Minas Gerais, estado responsável por aproximadamente 50% da produção nacional de café) com uma população de 899.876 habitantes (que representa 5% da população do Estado) (IBGE, 2010). A região produz 5 milhões de sacas por ano (que representa 24% da produção do Estado), sendo que 80% das fazendas de café são de até 20 ha, ou seja, de pequenos produtores (IBGE, 2006).

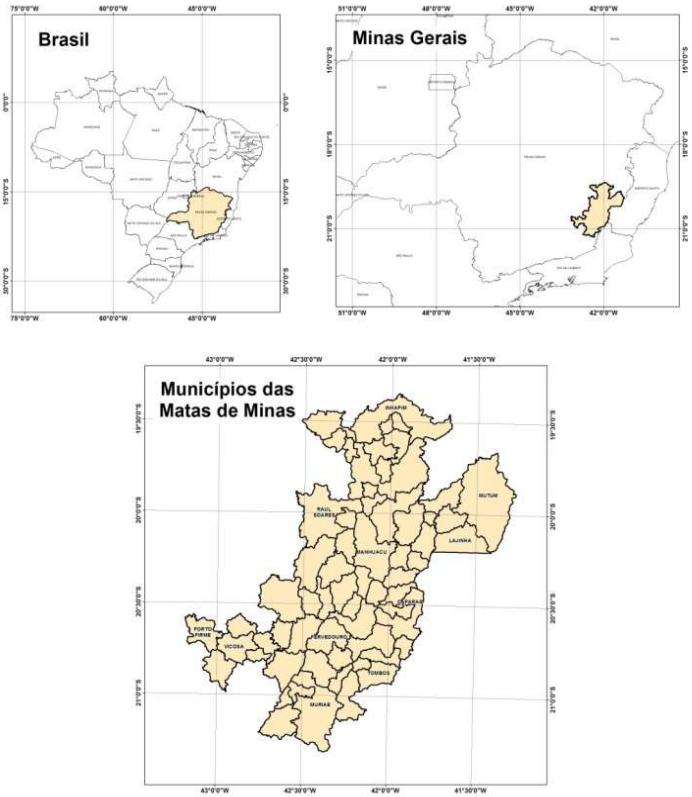

Figura 1. Mapa de localização dos municípios das Matas de Minas.

1.2. Base de dados

O estudo foi realizado utilizando-se uma cena Landsat 8 que foi recortada de acordo com os limites da área de estudo, com as devidas correções radiométricas e geométricas. Foi utilizado a interface dos softwares ArcGis 10.1 e R para o processamento das análises.

No ArGis foi realizada a coleta das amostras, totalizando um conjunto amostral formado por 83.347 pontos distribuídos de forma aleatória em toda a área de estudo e abarcando todas as oito classes de uso de interesse no estudo, a saber: café, mata, eucalipto, solo, água, pastagem, nuvem e sombra.

A partir das bandas da imagem Landsat 8 foram gerados os índices de vegetação NDVI e EVI e as componentes principais da imagem, sendo que o primeiro componente principal (PCAc1) e o segundo componente (PCAc2) respondem por 99,3% da variabilidade dos dados armazenados na imagem, por isso no escopo do presente trabalho utilizou-se somente esses dois componentes. Salienta-se que utilizou-se somente as bandas 2, 3, 4, 5, 6 e 7 da Landsat 8 para a execução do presente trabalho.

Ainda utilizando a interface do ArcGis foram extraídos os valores das bandas da imagem Landsat 8, dos índices de vegetação e dos componentes principais com base no arquivo de amostras no formato de pontos, utilizando-se a função *Extract Multi Values to Points* do *Spatial Analyst Tools*.

Posterior a extração dos valores, o arquivo de pontos em formato Dbase foi inserido no software R, onde foi realizado o particionamento do arquivo de amostras conforme os tamanhos dos conjuntos amostrais a serem testados e a classificação empregando quatro diferentes algoritmos: Logit, Random Forest, Redes Neurais e *Support Vector Machine* (SVM).

Para a avaliação da exatidão das classificações obtidas foi utilizado o índice Kappa (Congalton, 1991), cujos valores obtidos a partir dos diferentes algoritmos foram comparados entre si aplicando o teste estatístico z (α = 95), conforme Vieira (2001).

1.3. Cálculo do número de amostras com base na equação multinomial

Inicialmente, utilizou-se a equação multinominal proposta por Tortora (1978) e Congalton (1957) para calcular o número de amostras para a área de estudo, conforme a Equação 1.

$$n = \frac{B\Pi_i(1 - \Pi_i)}{b_i^2} \quad (1)$$

Os parâmetros pré-determinados para o processamento foram: existência de oito classes de uso do solo a serem mapeadas na área de estudo (k = 8), precisão desejada de 95%, e que a classe de especial interesse no estudo, no caso o café, abranja 30% da área do mapa (Πi = 30%).

O valor de B foi determinado a partir do qui-quadrado tabelado com 1 grau de liberdade e 1- α/k. Neste caso, o valor apropriado para B é $\chi 2 = 7.568$. Assim, o tamanho da amostra será:

$$n = 7{,}568(0{,}30)(1\text{-}0{,}30)/(0{,}05)2$$
$$n = 1{,}58928/0{,}0025$$
$$n = 636$$

Assim, foi necessário a coleta de 636 amostras divididas entre as oito classes de uso de interesse, cerca de 80 amostras por classe, para que haja o preenchimento adequado da matriz de erro. Com base nesse número, foram testados outros 7 tamanhos de conjunto de amostras por classe, sendo eles: 20,40,60, 100, 150, 200 e 500.

Foi implementada uma rotina em ambiente R que dividia o arquivo de amostras contendo 83.347 pontos de acordo com os tamanhos de conjunto de amostras por classe a serem testados. Sendo que para a validação dos modelos utilizou-se 150 amostras para cada classe.

As classificações empregando os diferentes números de amostras de treinamento foram processadas com os algoritmos Logit, Random Forest, Rede Neural e SVM utilizando script elaborado por Fernandes Filho (2014) no software R.

3. Resultados e Discussão

Na Figura 2 e 3 pode ser observado o efeito dos diferentes conjuntos de amostras no Kappa, respectivamente, médio e mínimo.

Ao adotar o número de amostras calculado a partir da equação multinominal de 80 amostras por classe, o valor do Kappa médio variou de 0,88 a 0,91 sendo que o menor valor foi o obtido pelo algoritmo SVM. Ao diminuir o número de amostras por classe verifica-se que ocorre também redução do valor do Kappa para todos os algoritmos, ocorrendo, porém, variação do valor médio e mínimo do mesmo.

Aumentando-se o número de amostras em relação ao calculado nota-se que ocorre uma melhora nos valores de Kappa mínimo e médio para todos os algoritmos, com estabilização do mesmo a partir do uso de 200 amostras de treinamento (Figuras 2 e 3).

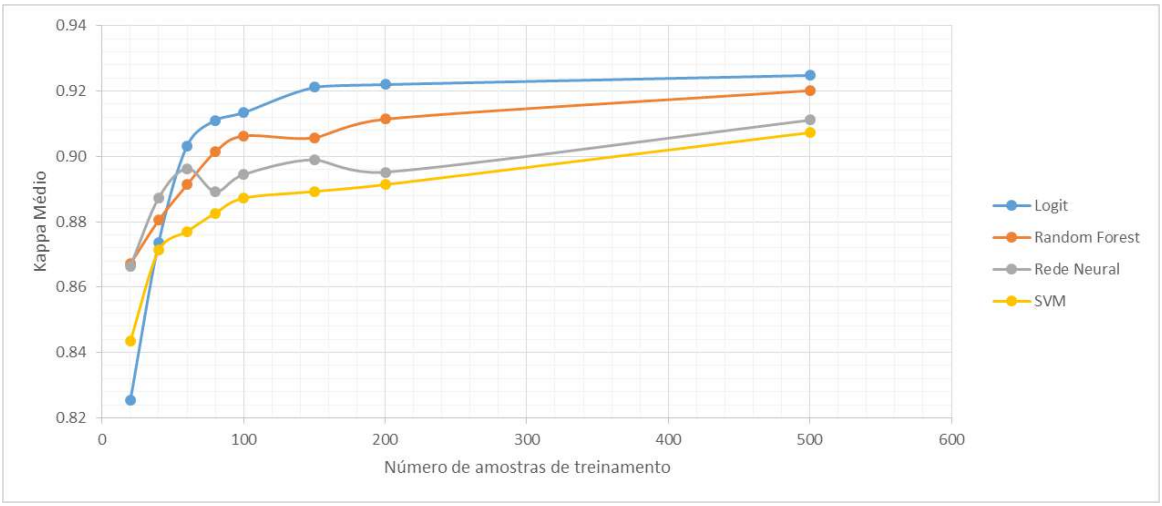

Figura 2. Efeito do número de amostras de treinamento no Kappa médio de classificações obtidas a partir de diferentes algoritmos.

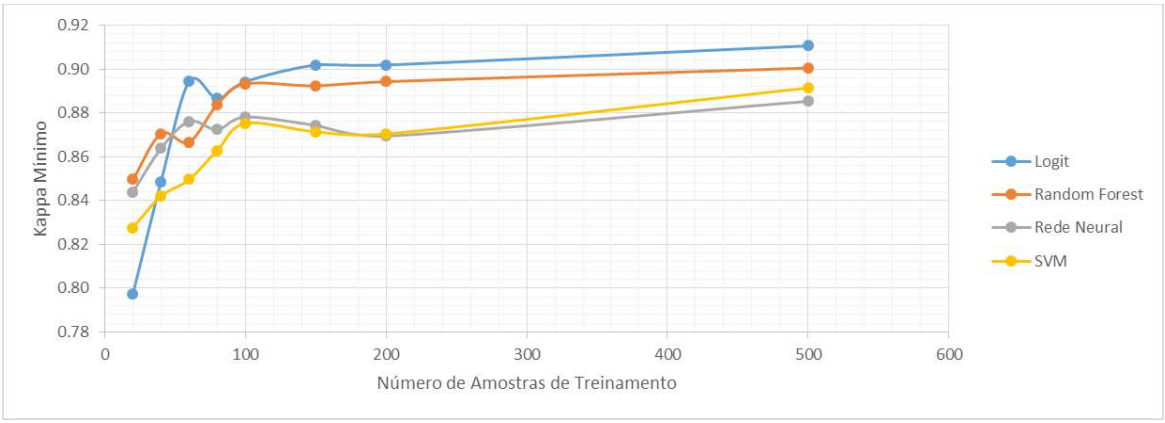

Figura 3. Efeito do número de amostras de treinamento no Kappa mínimo de classificações obtidas a partir de diferentes algoritmos.

Ao avaliar o efeito do número de amostras de treinamento separadamente para cada um dos algoritmos, verifica-se que o SVM é o algoritmo que apresentou maior sensibilidade à variação do tamanho do conjunto de amostras, sendo que os piores valores de Kappa são apresentados com conjunto de amostras menor que 200 amostras por classe. Fato este que contraria o afirmado pela literatura que aponta que o desempenho desse algoritmo é satisfatório para um pequeno conjunto de amostras devido a maximização dos limites de separabilidade das classes (Pal e Mather, 2004; Foody e Mathur, 2004).

Por outro lado, o algoritmo Random Forest foi o que apresentou menor sensibilidade à variação do tamanho do conjunto amostras, apresentando o melhor desempenho já com o menor conjunto de amostras.

Ao comparar os valores dos índices Kappas obtidos a partir dos diferentes algoritmos empregando os oito tamanhos de conjunto de amostras diferentes, foi observado que o emprego dos intervalos de 10, 20, 30, 40, 50, 80 e 100 amostras por classe, de modo geral, os Kappas obtidos pelos diferentes algoritmos não apresentam diferença estatística. Já a utilização de 150 e 200 amostras por classe, o algoritmo Logit apresentou melhor desempenho quando comparado ao SVM, porém não apresenta diferença estatística quando comparado aos demais algoritmos.

Com 500 amostras por classe o Logit apresentou melhor desempenho quando comparado à Rede Neural. Com esse tamanho de conjunto amostral por classe Logit e Random Forest apresentam melhor desempenho quando comparados ao SVM, porém quando comparados entre eles e à Rede Neural, estes não apresentam diferença estatística.

4. Conclusões

Quando considerado a equação multinominal para cálculo do número de amostras de treinamento por classe, os valores dos índices Kappas obtidos não são os melhores atingidos, visto que o patamar com menor variabilidade do valor deste ainda não foi atingido.

O algoritmo Random Forest foi o que apresentou menor sensibilidade à variação do tamanho do número de amostras de treinamento. Por outro lado, o SVM foi o que apresentou maior sensibilidade, tendo apresentado pior resultado com um número muito pequeno de amostras.

Quando comparados os valores dos índices Kappas obtidos pelos algoritmos Logit e SVM pelo teste z ($\alpha = 95\%$), estes somente se diferenciam estatisticamente empregando 150, 200 e 500 amostras por classe. Já o Random Forest se diferencia estatisticamente do SVM quando comparado pelo teste z ($\alpha = 95\%$) empregando 500 amostras por classe.

A metodologia apresentada no escopo do presente trabalho requer ser testada em outras áreas com outras classes de uso para sua validação. Por isso, a próxima etapa envolverá a avaliação do emprego dessa em cenas Landsat 8 para outros ambientes brasileiros.

Referências Bibliográficas

Campbell, J.B. Introduction to remote sensing. New York, The Guilford Press, 1987.

Congalton, R. G. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ., n.37, p. 35-46, 1991.

Congalton, R.G. Sample Design Considerations. IN: Congalton, R.G.Assessing the accuracy of remotely sensed data: principles and practices. Taylor & Francis Group, 2009. p.63-83.

Foody, M.G.; Mathur, A. A relative evaluation of multiclass image classification by Support Vector Machines. IEEE Transactions on Geoscience and Remote Sensing, n. 42, 2004, p. 1335 – 1343.

IBGE. Censo Agropecuário 2006. Rio de Janeiro, 2006.

Landis, J.R.; Koch,G.G. The measurement of observer agreement for categorical data. Biometrics, vol. 33, p.159-174, 1977.

Pal, M.; Mather, P. M. Assessment of the effectiveness of support vector machines for hyperspectral data. Future Generation Computer Systems, v. 20, n. 7, p. 1215–1225, 2004.

Richards, J. A.; Jia, X. Remote Sensing digital image analysis: An Introduction. 3. ed. Australia: Springer, 1999.
TORTORA, R. A note on sample size estimation for multinomial populations. The American Statistician. Vol. 32, n. 3. p. 100–102, 1978.

Vieira, C.A.O. Accuracy of remotely sensing classification of agricultural crops: a comparative study. 2001. 353p. Tese (Doutorado) - Universidade de Nottinghan.