Aprendizado Ativo com dados de Ciência Cidadã para o monitoramento de florestas tropicais

Fernanda B. J. R. Dallaqua[1], **Álvaro L. Fazenda**[1], **Fabio A. Faria**[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)
Avenida Cesare Mansueto Giulio Lattes, nº 1201 – Eugênio de Mello – SP – Brazil

{fernanda.dallaqua,alvaro.fazenda,ffaria}@unifesp.br

Abstract. *In April 2019 the Citizen Science project ForestEyes was launched, which uses non-specialized volunteers classifying remote sensing segments searching for deforested areas. In this work, these volunteers' contributions build a small but efficient training set through an Active Learning procedure. This training set is built iteratively based on different strategies that choose samples that will bring more representativeness. The results showed the importance of a correct initial training set and the balancing of the samples in the classification accuracy.*

Resumo. *Em Abril de 2019 foi lançado o projeto de Ciência Cidadã ForestEyes, o qual se utiliza de voluntários não especializados classificando segmentos de sensoriamento remoto em busca de áreas desmatadas. Neste trabalho usou-se as contribuições dos voluntários na criação de um pequeno, porém eficiente conjunto de treinamento, através de um procedimento de Aprendizado Ativo. Esse conjunto de treinamento é construído iterativamente através de diferentes estratégias que escolhem amostras que trarão maior representatividade. Os resultados mostraram como uma definição apropriada do conjunto de treinamento inicial e o balanceamento das amostras podem ser importantes na acurácia da tarefa de classificação.*

1. Introdução

As florestas tropicais desempenham um importante papel no ecossistema global, uma vez que regulam o clima e a chuva, absorvem grandes quantidades de dióxido de carbono e possuem uma grande biodiversidade. Infelizmente, as florestas tropicais sofrem de desmatamento progressivo e indiscriminado [Luz et al., 2014].

Um dos programas de monitoramento de regiões florestais bem sucedido e conhecido é o PRODES (Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite) do INPE (Instituto Nacional de Pesquisas Especiais) [Valeriano et al., 2004], que produz as taxas anuais de desmatamento na Amazônia Legal Brasileira [Luz et al., 2014].

Geralmente uma grande quantidade de dados é necessária para treinar um classificador e esses dados necessitam ser rotulados por especialistas em uma extensa análise manual, o que deixa o procedimento caro em termos financeiros e de tempo. Assim, é desejável usar um conjunto de treinamento pequeno que obtém alta acurácia na classificação. O método de Aprendizado Ativo (AA, ou ainda, *Active Learning*, em inglês) [Tuia et al., 2011], propõe superar este desafio construindo iterativamente o conjunto de treinamento, adicionando apenas amostras que poderão trazer melhor representatividade.

Neste trabalho, em substituição aos especialistas na rotulação do conjunto de treinamento, serão utilizados voluntários não especializados, que contribuirão através de um projeto de Ciência Cidadã (CC). Em CC voluntários coletam, analisam e classificam dados para resolver vários problemas técnicos e científicos [Grey, 2009].

Em Abril de 2019, foi lançado o projeto ForestEyes [Dallaqua et al., 2019], hospedado na famosa plataforma Zooniverse.org [Smith et al., 2013]. Este projeto tem como objetivo aliar CC com aprendizado de máquina, onde voluntários classificarão segmentos de imagens de sensoriamento remoto que serão utilizados como conjunto de treinamento aplicado ao método de AA.

Este trabalho descreve os resultados obtidos com o procedimento previamente descrito, nas classificações de áreas contidas no estado de Rondônia. As diferentes imagens para os conjuntos de treinamento e teste foram segmentadas a partir de imagens de sensoriamento remoto do satélite Landsat-8 [Lauer et al., 1997] obtidas em 2016.

2. Projeto ForestEyes

No projeto ForestEyes (https://www.zooniverse.org/projects/dallaqua/foresteyes) os voluntários analisam segmentos de imagens de sensoriamento remoto do Landsat-8 e os classificam em 3 diferentes classes: Floresta, Não-Floresta ou Indefinida, caso o segmento apresente uma mistura de *pixels* de Floresta e Não-Floresta, onde não seja possível definir uma clara maioria. Cada segmento de imagem é analisado por 15 voluntários, sendo computado o voto da maioria como a classificação final. Na primeira campanha do projeto, os voluntários classificaram 1022 segmentos de uma pequena área de Rondônia no ano de 2016 [Dallaqua et al., 2019].

O já citado programa PRODES forneceu o conjunto verdade utilizado na avaliação e validação do método proposto. Desta forma, a partir de uma classificação ”binarizada” do PRODES sobre cada segmento de imagem (que representam cada tarefa no projeto de CC), criou-se um novo conjunto verdade, onde a classe correta para um segmento é a classe da maioria dos *pixels* dentro do próprio. Considerando esse conjunto verdade gerado, os voluntários obtiveram uma acurácia de 86,5% na tarefa de classificação dos segmentos do conjunto de treinamento [Dallaqua et al., 2019].

3. Metodologia

As amostras rotuladas do conjunto de treinamento pelos voluntários que foram classificadas como Indefinidas, ou que apresentaram empate entre pelo menos duas classes foram eliminadas, o que o reduziu para 934 amostras perfeitamente balanceadas entre as classes. Com esta filtragem, a acurácia dos voluntários subiu para 94,6%.

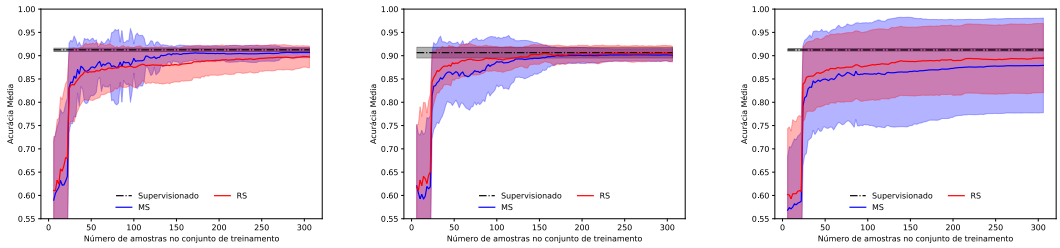
Para cada segmento de imagem, tanto para o conjunto de treinamento quanto para o de teste, extraiu-se descritores de textura pelo método de Haralick et al. [Haralick et al., 1973]. O conjunto de teste possui 543 segmentos Não-Floresta e 386 Floresta.

Para o procedimento de AA foi utilizado o classificador SVM (*Support Vector Machines*) com as estratégias *Margin Sampling* (MS) [Schohn and Cohn, 2000] e *Random Sampling* (RS). Para a primeira estratégia, as amostras a serem inseridas no treinamento são as que estão mais próximas ao hiperplano enquanto a segunda as escolhe aleatoriamente. O procedimento começou com 6 amostras aleatoriamente selecionadas,

e inserções de 2 novas amostras para cada iteração do AA, até todas as amostras serem incluídas. Foram realizados 30 procedimentos, cada um com diferentes conjuntos de treinamento iniciais. O refinamento dos parâmetros do SVM (popularmente conhecido em inglês por *grid search*) foi realizado apenas nas iterações 1 e 10 dos procedimentos.

4. Resultados

Como a estratégia MS utiliza informação para decidir as amostras a serem inseridas no treinamento, era esperado que ela obtivesse um melhor resultado que a escolha aleatória de amostras. Porém, realizando o que foi descrito na seção 3, não foi esse o resultado encontrado, como pode ser visto na Figura 1a. Uma possível hipótese para este fato considera que o balanceamento do conjunto das amostras ajuda o RS a conseguir bons resultados.

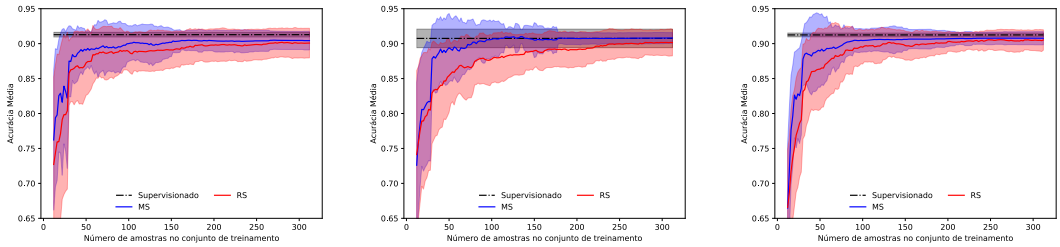


(a) Proporção 50-50. **(b) Proporção 60-40.** **(c) Proporção 70-30.**

Figura 1. Resultados iniciais para o método de AA.

Investigando essa possibilidade criou-se subconjuntos de amostras, variando a proporção entre Floresta (F) e Não-Floresta (NF) de duas formas: considerando 60% de amostras da classe F e 40% na classe NF, e na proporção 70% e 30%, respectivamente. Porém, como pode ser observado nas Figuras 1b e 1c, mesmo com subconjuntos desbalanceados, RS foi tão bom quanto o MS.

Posteriormente, percebeu-se que muitos conjuntos de treinamento iniciais eram fortemente ou totalmente desbalanceados, fazendo com que a estratégia MS não conseguisse diferenciar as amostras, gerando a mesma distância ao hiperplano para todas. Desta forma, as amostras inseridas inicialmente no treinamento eram simplesmente as primeiras da lista e não as que seriam mais representativas. Tal fato pode ter influenciado todo o procedimento, principalmente até 200 amostras.



(a) Proporção 50-50. **(b) Proporção 60-40.** **(c) Proporção 70-30.**

Figura 2. Resultados para o método de AA com agrupamento no conj. inicial.

Para testar essa nova hipótese, foram criados conjuntos de treinamento iniciais mais balanceados através de agrupamento de conjuntos (popularmente conhecido como "clusterização") com o método *k-means*, em conjunto com o método cotovelo (*elbow method*, em inglês) [Kodinariya and Makwana, 2013], que permite definir o melhor número de *clusters* a ser usado para o *k-means*. Os resultados podem ser vistos na Figura 2, onde se percebe que MS apresenta melhor acurácia em todos os casos. Na Figura 2c percebe-se ainda que os resultados obtidos apresentam comportamento mais estável em comparação com a Figura 1c.

5. Conclusão

Neste trabalho foi proposto o uso de dados classificados por voluntários não especializados para construir um conjunto de treinamento pequeno e eficiente aplicado ao método de AA. Foi mostrado que com apenas uma parte das amostras (200 amostras) é possível obter acurácia média próxima ao aprendizado supervisionado, que utiliza todas as 934 amostras. Tal resultado pode trazer melhorias no tempo de processamento, além de apresentar baixo custo no uso de especialistas, os quais foram substituídos por usuários comuns, sob um projeto de CC. Finalmente, foi mostrado que o balanceamento das amostras e a escolha do conjunto de treinamento inicial influenciam na acurácia dos resultados, sob diferentes heurísticas ou métodos de escolha de amostras do AA.

Referências

Fernanda B. J. R. Dallaqua, Álvaro L. Fazenda, and Fabio A. Faria. ForestEyes project: Can citizen scientists help rainforests? In *IEEE 15th International Conference on eScience*, pages 18–27. IEEE, 9 2019.

François Grey. Viewpoint: The age of citizen cyberscience. *Cern Courier*, 29, 2009.

Robert M. Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

Trupti M. Kodinariya and Prashant R. Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

Donald T. Lauer, Stanley A. Morain, and Vincent V. Salomonson. The landsat program: Its origins, evolution, and impacts. *Photogrammetric Engineering and Remote Sensing*, 63(7):831–838, 1997.

Eduardo F. P. Luz, Felipe R. S. Correa, Daniel L. González, François Grey, and Fernando M. Ramos. The forestwatchers: a citizen cyberscience project for deforestation monitoring in the tropics. *Human Computation*, 1(2):137–145, 2014.

Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, volume 2, page 6. Citeseer, 2000.

Arfon M. Smith, Stuart Lynn, and Chris J. Lintott. An introduction to the zooniverse. In *First AAAI conference on human computation and crowdsourcing*, 2013.

Devis Tuia, Michele Volpi, Loris Copa, Mikhail Kanevski, and Jordi Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011.

Dalton M. Valeriano, Eliana M. K. Mello, José Carlos Moreira, Yosio E. Shimabukuro, Valdete Duarte, I. M. Souza, J. R. Santos, Claudio C. F. Barbosa, and R. C. M. Souza. Monitoring tropical forest from space: the prodes digital project. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 35:272–274, 2004.