CLASSIFICAÇÃO DE ÁREAS QUEIMADAS POR *MACHINE LEARNING* USANDO DADOS DE SENSORIAMENTO REMOTO

Cícero Alves dos Santos Júnior [1], *Olga Oliveira Bittencourt*[1], *Fabiano Morelli*[1] *e Rafael Santos*[1]

[1] INPE – National Institute for Space Research
Av. dos Astronautas, 1758 - 12227-010 - São José dos Campos - SP, Brazil
{cicero.alves; olga.bittencourt; fabiano.morelli; rafael.santos}@inpe.br

RESUMO

Apresentamos um estudo para melhorar a automação do processo de classificação de áreas queimadas usando dados de sensoriamento remoto. Mostramos os atributos mais relevantes para enriquecer a base de conhecimento e o resultado da aplicação deles em uma comparação de modelos de classificação de *machine learning*. Validamos nosso estudo com dados de queimadas do Cerrado feitos por especialistas. Os melhores resultados foram obtidos com os modelos *Random Forest* e *Neural Networks* e indicam a viabilidade de utilização da abordagem no processo de classificação de áreas queimadas.

Palavras-chave – áreas queimadas, classificação, machine learning, dados de sensoriamento remoto.

ABSTRACT

We present an study to improve automation on 'Woody savannah burned areas' classification process in a continuous and periodical way through the use of machine learning classification models. We propose some relevant features to enrich a burns' knowledge database in some classification models. The developed approach is validated over a study area in the Brazilian Cerrado against reference data derived from classifications done by experts. Best results were obtained by Random Forest and Neural Network models and indicate enhancement on the methods used so far.

Key words – *burned areas, classification, machine learning, remote sensing data.*

1. INTRODUÇÃO

O Cerrado é uma região de savana rica em biodiversidade e um dos biomas mais ameaçados do país. Sua área ocupa em torno 204 milhões de hectares, 24% do território brasileiro, e já perdeu quase metade de sua cobertura vegetal original. Desmatamentos e queimadas são os maiores responsáveis por esse processo, tendo sido registrados mais de 30.000 focos de incêndio por ano neste bioma nos últimos 15 anos. Juntamente com o bioma da Amazônia, o Cerrado é responsável por quase 20% do total das emissões brasileiras de gases do efeito estufa. Para entender esse processo é importante compreender os aspectos relacionados à ocorrência do fogo e aos impactos econômicos, sociais e ambientais gerados [1–3].

O Instituto Nacional de Pesquisas Espaciais (INPE) mantém um programa de Monitoramento de Queimadas e Incêndios Florestais [4]. São analisados continuamente aspectos relacionados à ocorrência de fogo em áreas de vegetação tanto no bioma Cerrado quanto no restante do território brasileiro e parte da América Latina. O monitoramento é realizado por sensoriamento remoto de duas formas independentes, dependendo da resolução das imagens do satélite. Imagens de baixa resolução espacial (pixels maiores que 300m) são usadas para gerar produtos de dados diariamente, focos de incêndio e previsão de risco de fogo. Imagens de média resolução espacial (pixels em torno de 30m) são usados para análises mais precisas e menos frequentes como as estimativas periódicas de emissão de poluentes e, mais recentemente, estimativas de superfícies queimadas. Seus resultados são utilizados, por exemplo, como subsídios de políticas públicas como o Código Florestal Brasileiro e para contribuir para que as metas de redução das emissões de gases assumidas pelo Governo brasileiro na Convenção do Clima [5] possam ser atingidas.

Um desafio nesse monitoramento é combinar eficiência e rapidez na identificação de áreas queimadas. A realidade do uso da cultura do fogo no Cerrado [2] se alia a áreas de difícil acesso, escassez de recursos e um grande número de focos de queimada ativos. É necessário um processo rápido de análise e com baixa taxa de erro na identificação.

Esse trabalho está inserido no esforço de desenvolver uma abordagem genérica e automática para classificar áreas queimadas. Buscamos responder a pergunta: *Como combinar um conjunto de dados relevantes em uma abordagem de machine learning para classificar áreas queimadas?* Para isso, mostramos os atributos mais relevantes estudados e o resultado comparativo do uso deles em diversos modelos de classificador. Analisamos modelos robustos o suficiente para manipular dados não classificados previamente.

O restante desse trabalho está organizado da seguinte forma: Seção 2 mostra o uso do sensoriamento remoto para monitorar áreas queimadas. Seção 3 apresenta os experimentos, atributos relevantes, modelos de aprendizagem de máquina e discute os resultados. Seção 4 apresenta as conclusões e trabalhos futuros.

2. MAPEAMENTO DE ÁREAS QUEIMADAS POR SENSORIAMENTO REMOTO

O uso de imagens de sensoriamento remoto é uma abordagem periódica e eficiente para o monitoramento de áreas queimadas, principalmente em locais de difícil acesso e áreas de grande extensão territorial, como é o caso do Brazil.

Algumas linhas de pesquisa apresentam abordagens para calcular estimativa de áreas queimadas em eventos específicos [6] e a utilização de características regionais nas análises [7, 8]. Pereira et al. [9] apresentam uma abordagem para detectar automaticamente áreas queimadas utilizando focos

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

ativos de queimadas. Estudos recentes empregam os avanços de imagens de média resolução para mapear áreas queimadas em uma forma mais automática. Liu et al. [3] desenvolveram um algoritmo para monitoramento anual contínuo através de um modelo harmônico em séries temporais do LandSat ao invés da tradicional comparação de mudanças em imagens consecutivas.

Nos últimos anos, foi desenvolvida uma nova geração de satélites capazes de fornecer imagens com resoluções melhores e georeferenciamento mais preciso, como CBERS-4, Landsat 8 e Sentinel-2. Li & Roy [10] apresentam alguns avanços dessa nova geração. Entre os satélites de observação da Terra que disponibilizam imagens orbitais de média resolução destaca-se o Programa Landsat, que, desde a década de 80, gera imagens da mesma área a cada 16 dias. Seu mais novo satélite, o Landsat 8, disponibiliza imagens com resolução espacial na faixa de 30m para cada pixel.

2.1. Mapeamento de áreas queimadas no INPE

O processo de mapeamento de áreas queimadas do INPE se baseia na comparação de imagens LandSat da mesma área adquiridas em momentos distintos. Inicialmente são detectadas as áreas que sofreram mudanças, que podem ter sido causadas por vários fatores, entre eles, o fogo. Em seguida, dentro dessas áreas, são classificadas quais mudanças foram originadas por queimadas.

O processo de detecção [11] é automático e começa com a avaliação dos pixels da imagem e seus índices de espectrais vegetação NDVI (Índice de Vegetação Normalizada) e NBRL (Índice de Queimada Normalizada). Se esses valores estiverem dentro de limites pré-estabelecidos por especialistas, eles indicam que os pixels sofreram algum tipo de alteração a partir da comparação da imagem anterior. Após esse passo, realiza-se a segmentação através da verificação vizinhos mais semelhantes com a posterior delimitação de um polígono que indica a área de abrangência daquela alteração.

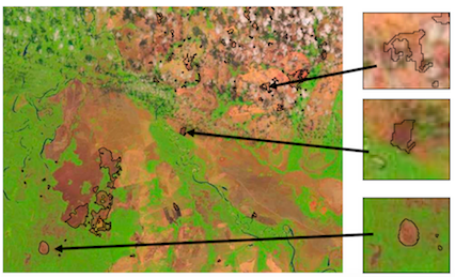

Figura 1: Fragmentos de uma composição RGB.

As imagens da Figura 1 ilustram o fragmento de uma composição RGB e detalham três mudanças identificadas. O primeiro fragmento à direita foi classificado como *não queimada* pois a mudança indicada foi causada pela presença de nuvens na imagem anterior. As outras duas indicações são *queimadas confirmadas*.

Atualmente, o processo de classficação se baseia na combinação de um conjunto de evidências gerados por: redes neurais, uma clusterização supervisionada, a ocorrência de focos de queimadas e áreas queimadas próximas em um período anterior. Após a classificação, é realizada uma avaliação manual por especialistas antes da publicação dos dados oficiais. Inicialmente essas verificações eram realizadas manualmente, um processo exaustivo e que consumia muito tempo. Atualmente os dados duvidosos são verificados e as demais dados são reavalidos por amostragem.

Em [12] apresentamos um modelo de rede neural que usava 1 neurônio para comparar os índices de vegetação NDVI e NBRL de um período de 3 meses de um ano com os dados dos mesmos meses no ano seguinte. A abordagem se mostrou promissora e estendemos a análise.

Nosso objetivo é automatizar o processo de avaliação, tornando-o mais rápido eficiente, e consequentemente, diminuindo o trabalho de reavaliação dos especialistas. A idéia é a construção de um modelo que analise a base histórica do período de um ano e a cada novo conjunto de dados gerado por uma imagem de satélite possa se autoajustar para classificar os novos dados com a mesma acurácia.

3. EXPERIMENTOS

Exploramos atributos, classificadores e suas combinações para resolver o problema de classificar dados de queimadas de monitoramento contínuo. Para isso, usamos a base de conhecimento adquirida pelo INPE ao longo das últimas décadas de monitoramento. Os experimentos foram realizados no ambiente Orange [13], uma plataforma de código aberto para análises de *machine learning* e visualização de dados.

3.1. Dados

A área de estudo desse trabalho compreeende a órbita-ponto 223/067, majoritariamente pertencente ao Bioma Cerrado. Ela contém regiões não protegidas e dois grandes parques de proteção ambiental: O Parque Florestal do Cantão e o Parque Florestal do Araguaia. Apesar da existência de grandes áreas de proteção, foram identificados aproximadamente 56.000 focos de incêndio e 45.000 supostas queimadas no ano de 2017. Assim, o conjunto contém um grande número de ocorrências que nos permite construir um conjunto de exemplos mais completo.

O conjunto de dados, descrito na Tabela 1 foi recuperado no início de Novembro e é composto por 10 imagens do período de 16/03/17 a 10/10/17. O monitoramento é um processo em operação contínua, avanços são ocasionalmente introduzidos no sistema e algumas vezes os resultados oficiais são reprocessados e atualizados.

Tabela 1: Descrição dos dados com evidências de queimadas

Queimadas		*Não queimadas*	
polígonos	area (ha)	polígonos	area (ha)
6.322	565.658,55	39.571	503.099,40

Essa base de dados é formada por objetos que foram selecionados por terem sofrido algum tipo de alteração com evidência de queimadas. Eles estão armazenados em formato shapefile. Além da geometria espacial, no caso polígonos, cada objeto possui um conjunto de atributos associados a ele gerados por tratamentos estatísticos aplicados aos valores dos pixels que pertencem a cada objeto. Cada imagem LandSat

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

original contém os espectros de reflectância de cada pixel e o conjunto de atributos é composto pela mediana de cada banda e outros índices de vegetação. Utilizaremos bn como anagrama para os dados de cada banda n. O conjunto inicial de atributos é composto por: b2, b3, b4, b5, b6, b7, area(ha), NDVI, NBR, dif_NDVI e dif_NBR.

3.2. Atributos relevantes na base de conhecimento

Analisamos diferentes índices e métricas relacionadas para sugerir os mais relevantes ao problema e como utilizá-los nos modelos. Esses experimentos não estão descritos neste trabalho e serão apresentados em um relatório interno completo. A Tabela 2 resume os atributos.

Tabela 2: Descrição dos atributos incluidos

atributo	*descrição*	*exemplo*
Anteriores	contagem de queimadas	3
Focos	contagem de focos ativos	12
MIRBI	(10 x b7) - (9.8 x b6) + 2	2,287
NDWI	(b3 - b6) / (b3 + b6)	0,498

3.2.1. Índices Espectrais de Vegetação:

Os índices *Mid-Infrared Burn Index* (MIRBI) e *Normalized Difference Water Index* (NDWI) são medianas estimadas a partir dos valores das bandas.

3.2.2. Queimadas próximas anteriores (Anteriores):

Investigamos a hipótese de *uma queimada espacialmente próxima em uma data anterior ser um indicativo relevante na classificação de uma nova área*. Ela pode, por exemplo, ser a continuação de uma queimada que estava acontecendo no momento da passagem do satélite. Para extrair a relevância do atributo e minimizar as limitações da resolução espacial, usamos um *buffer* em torno da área que estava sendo analisada. Experimentos mostraram que a distância de proximidade de 220m dentro um período de 64 dias anteriores influenciam de forma relevante a classificação. O resultado é a contagem do total de áreas queimadas próximas à area identificada.

3.2.3. Focos ativos (Focos):

Investigamos a hipótese de *um foco ativo próximo em uma data anterior ser um indicativo relevante na classificação.* Dados de focos de incêndio ativos são produzidos pelo INPE com baixa resolução espacial (variando de 375m a 4000m) por diferentes satélites. Nossa abordagem considera dados dos satélites Terra, Aqua e NPP cuja resolução temporal é de 4 imagens por dia. Os sensores MODIS do TERRA e AQUA tem resolução espacial de 1.000m e o sensor VIIRS do NPP-Suomi tem 375m e 750m de resolução espacial. A base de conhecimento é formada por focos ativos no período de 14/02/2017 a 10/10/2017 contém 56.451 objetos armazenados em formato shapefile com geometria espacial composta por pontos. Os dados mais recentes e atualizados podem ser obtidos em [4].

Nesse experimento, combinamos conjuntos com diferentes resoluções espaciais e temporais. As melhores caracterizações de queimadas usando focos de incêndio foi obtida considerando os focos ativos dentro de um *buffer* de 660m em torno da área do polígono em um período de até 32 dias anteriores à identificação. O resultado é uma contagem do número de focos dentro da área indicada.

3.3. Modelos de Classificador

Exploramos a performance de vários modelos clássicos de *machine learning* por não existir um modelo abrangente para descrever as queimadas a partir dos dados. Uma breve descrição dos parâmetros é mostrada abaixo:

K- Nearest Neighbors (kNN): Utiliza a influência da maioria dos 5 vizinhos mais próximos com distância Eucliana.

Decision Trees (DT): Aplica um conjunto de regras de decisão baseada nos melhores separadores de classe para realizar a classificação com uma profundidade máxima de 12 nós.

Random Forests (RF): Combina o resultado de um conjunto de 24 árvores de decisão e o mínimo de 5 instâncias em cada folha.

Neural Networks (NN): Baseado em uma rede de múltiplas camadas (MLP) com retropropagação para separar as classes. Foram utilizados os seguintes parâmetros: 100 neurônios, máximo de 300 ativações e penalidade de 0,00001.

Support Vector Machines (SVM): Define o melhor hiperplano para separar o conjunto em duas classes no espaço. Utilizamos os seguintes parâmetros: Cost: 100,00; Regression loss epsilon: 0,5; kernel Linear; Numerical tolerance: 0,01 and Iteration limit: 1000.

3.3.1. Treinamento dos modelos

O conjunto inicial de dados foi dividido em duas partes: 40.726 polígonos (dados de março a setembro) foram usados para o treinamento e os outros 5.167 (dados de outubro) para a validação. Destes 40.726, 4.716 correspondem a queimadas, 36.010 correspondem a não queimadas e é o mesmo para todos os modelos. Nosso conjunto não é balanceado e é importante verificar diferentes métricas para entendermos melhor a resposta. O resultado é a média dos valores obtidos em 10 testes de validação cruzada e indicam a resposta de cada modelo gerado.

A tabela 3 apresenta as métricas de acurácia, precisão, revocação e F1 calculadas para cada modelo. A acurácia é a proporção de exemplos corretamente classificados em relação ao número total de exemplos. A precisão é a proporção de positivos verdadeiros dentro do conjunto de positivos. No caso, queimadas confirmadas entre todas as indicações de queimadas. Revocação é a proporção das positivas verdadeiras entre todas as instâncias verdadeiras. No caso, o número de queimadas recuperadas dentro de todo o conjunto de queimadas existentes. F1 é a média harmônica ponderada entre precisão e revocação.

O resultado das acurácias foi maior que 0.95, exceto para o SVM e DT, que apresentam os maiores conjuntos de erros, tanto para os falso positivos (não queimadas classificadas

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

Tabela 3: Resultados da validação cruzada no treinamento

Modelo	*Acurácia*	*F1*	*Precisão*	*Revocação*
kNN	0.964	0.947	0.948	0.946
DT	0.879	0.956	0.956	0.957
RF	0.987	0.970	0.970	0.969
NN	0.991	0.970	0.970	0.970
SVM	0.784	0.839	0.833	0.881

como queimadas) quanto para os falso negativos (queimadas classificadas como não queimadas). RF e NN apresentaram os maiores resultados de acurácia, como valores em torno de 0.98, e os melhores resultados de precisão e revocação, com valores em torno de 0.97.

3.4. Validação

Para verificar se os modelos gerados eram capazes de classificar dados não conhecidos, simulamos um processo real em que usamos um conjunto de dados conhecidos para classificar a data mais recente. Essa validação foi realizada com o conjunto de áreas identificadas em 10/10/2017 que contém 5.167 polígonos, sendo 1.606 queimadas e 3.561 não queimadas.

A Tabela 4 apresenta os resultados. Eles seguem a tendência do conjunto anterior, em que o SVM obteve os valores mais baixos. Os modelos RF e NN obtiveram os melhores resultados, com todas as métricas acima de 90% na classificação de dados de queimadas não previamente conhecidos pelo modelo.

Tabela 4: Resultados de validação

Modelo	*Acurácia*	*F1*	*Precisão*	*Revocação*
kNN	0.874	0.807	0.817	0.800
DT	0.829	0.858	0.881	0.868
RF	0.962	0.901	0.907	0.904
NN	0.967	0.901	0.904	0.907
SVM	0.524	0.587	0.598	0.678

Os experimentos mostram a viabilidade da combinação de um conjunto de dados relevantes de aproximadamente 1 ano com modelos de *machine learning* na classificação de dados de queimadas.

4. CONCLUSÕES E TRABALHOS FUTUROS

Resultados mostraram que é possível usar a base de conhecimento proposta em uma abordagem automática de classificação. Definimos um conjunto de atributos relevantes para construir uma base de conhecimento de aproximadamente 1 ano capaz de classificar áreas queimadas e não queimadas. Aplicamos essa base em modelos de classificadores e os melhores resultados foram obtidos com os modelos *Random Forest* e *Neural Networks*.

Uma desafio encontrado foi a caracterização da classe *não queimadas*, que pode ser originada de eventos distintos como desmatamento, colheita ou preparação do solo. Assim, vimos que a classificação binária (queimadas e não queimadas) de todas as mudanças ocorridas com evidências de queimadas pode não caracterizar adequadamente esse conjunto.

Além desse desafio, nossa base de dados não é livre de erros. Algumas áreas são ambíguas e mesmo os especialistas tem dúvidas em certos locais. Nesses casos, se os especialistas não possuem certeza em determinadas áreas, elas são classificados como não queimadas, aumentando ainda mais a variabilidade dentro da classe.

O trabalho continua sendo aprimorado. Estamos testando a estratégia mais adequada para tratar os tratar os casos duvidosos e trabalhando na adaptabilidade do modelo para todas as áreas do projeto.

Para trabalhos futuros, sugerimos incorporar outros produtos de dados relacionados ao fogo como risco de fogo e tipo de solo para gerar modelos cada vez mais precisos na caracterização de queimadas.

5. AGRADECIMENTOS

This study was supported by National Council for Scientific and Technological Development (CNPq)/Coordination of Associated Laboratories (CLA/INPE) (no.300587/2017-1).

6. REFERÊNCIAS

[1] BOWMAN, D. et al. Fire in the earth system. v. 324, p. 481–4, 05 2009.

[2] PIVELLO, V. The use of fire in the cerrado and amazonian rainforests of brazil: Past and present. v. 7, p. 24–39, 04 2011.

[3] LIU, J. et al. Burned area detection based on landsat time series in savannas of southern burkina faso. *International Journal of Applied Earth Observation and Geoinformation*, v. 64, p. 210 – 220, 2018. ISSN 0303-2434.

[4] Instituto Nacional de Pesquisas Espaciais (INPE). *Programa de Monitoramento de Queimadas*. <http://www.inpe.br/queimadas/portal>. Accessed: 2018-01-28.

[5] Ministério do Planejamento, Orçamento e Gestão (MPOG). *Plano Plurianual 2016-2019: Desenvolvimento, produtividade e inclusão social*. <http://www.planej.gov.br/assuntos/planeja/plano-plurianual/relatorio-objetivos.pdf>. Accessed:2017-09-12.

[6] KATAGIS, T. et al. Trend analysis of medium- and coarse-resolution time series image data for burned area mapping in a mediterranean ecosystem. p. –, 01 2014.

[7] LIBONATI, R. et al. An algorithm for burned area detection in the brazilian cerrado using 4μm modis imagery. *Remote Sensing*, v. 7, n. 12, p. 15782–15803, Nov 2015. ISSN 2072-4292.

[8] BOSCHETTI, L. et al. Modis–landsat fusion for large area 30 m burned area mapping. v. 161, 02 2015.

[9] PEREIRA, A. A. et al. Burned area mapping in the brazilian savanna using a one-class support vector machine trained by active fires. *Remote Sensing*, v. 9, n. 11, 2017.

[10] LI, J.; ROY, D. A global analysis of sentinel-2a, sentinel-2b and landsat-8 data revisit intervals and implications for terrestrial monitoring. v. 9, p. 902, 08 2017.

[11] MELCHIORI, A. E. et al. A landsat-tm/oli algorithm for burned areas in the brazilian cerrado: preliminary results. In: ____. [S.l.]: Imprensa da Universidade de Coimbra, 2014. p. 1302–1311. ISBN 978-989-26-0884-6 (PDF).

[12] ANDRADE, R. N. de et al. Classificação semiautomatica de areas queimadas com o uso de redes neurais. In: *XVIII Brazilian Symposium on Geoinformatics*. [S.l.: s.n.], 2017. p. 92–97.

[13] DEMsAR, J. et al. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, v. 14, p. 2349–2353, 2013.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings