

Um Estudo Experimental com Classificadores baseados em Regiões e Perfis EVI

David Pereira Barbosa¹
Alexandre Noma¹
Thales Sehn Korting²
Leila Maria Garcia Fonseca²

¹ Universidade Federal do ABC – UFABC
Rua Santa Adélia, 166 – 09210-170 – Santo André – SP, Brasil
davidpereirab@gmail.com

² Instituto Nacional de Pesquisas Espaciais – INPE
Caixa Postal 515 – 12245-970 – São José dos Campos - SP, Brasil
thales, leila@dpi.inpe.br

Abstract. INPE is responsible for several projects, including PRODES and TerraClass. Basically, PRODES provides annual maps corresponding to annual deforestation in Amazonia. Based on a deforestation map, TerraClass provides a classification map for the deforested areas: agriculture, pasture, forest, hydrography, urban, etc. In order to build a classification map, manual classification is a cumbersome and tedious work. In this sense, automatic or semi-automatic approaches are highly desirable for classification of the deforested areas. Previous work compared different approaches by using TerraClass data from 2008 for binary classification: agriculture and non-agriculture. The present paper extends the previous work in three aspects: (1) by including recent TerraClass data from 2010; (2) by treating the multiclass case by considering 3 or more classes; and (3) by considering an additional approach for classification. Specifically, SVM, OPF, Naive Bayes, Decision Tree, Nearest Neighbors and a combination of the methods are compared by following a k -fold cross validation.

Keywords: remote sensing, image processing, pattern recognition, sensoriamento remoto, processamento de imagens, reconhecimento de padrões

1. Introdução

O Instituto Nacional de Pesquisas Espaciais (INPE) é um dos principais responsáveis pelo monitoramento contra o desmatamento por corte raso na Amazônia Legal. Mapas anuais de desmatamento na região são fornecidos pelo projeto PRODES e podem ser obtidos livremente em <http://www.obt.inpe.br/prodes/>. Baseado nestes mapas de desmatamento, mapas de classificação de uso e cobertura de solo são disponibilizados pelo projeto TerraClass em http://www.inpe.br/cra/projetos_pesquisas/terraclass.php. Com estes mapas de classificação, é possível, por exemplo, identificar as regiões de pastagem, agricultura, floresta, hidrografia, área urbana, etc.

As florestas representam grandes reservas de carbono. Quando as florestas são derrubadas e/ou queimadas, uma enorme quantidade de carbono é liberado na atmosfera, acelerando as mudanças climáticas. Neste sentido, além de monitorar, é fundamental proporcionar informações que possam evitar o desflorestamento. Neste sentido, existem projetos que pesquisam métodos para identificar e caracterizar possíveis padrões e trajetórias de ocupação na região Amazônica, por exemplo, o projeto descrito em http://www.dpi.inpe.br/gilberto/papers/luca_iam_research_project.pdf. Através das possíveis trajetórias de ocupação, é possível desenvolver políticas públicas de prevenção.

Por um lado, a precisão de estimação das trajetórias de ocupação depende diretamente da qualidade dos mapas de classificação. Por outro, é crucial que os mapas temáticos de uso do solo sejam disponibilizados rapidamente para possibilitar desenvolver e implantar medidas preventivas que sejam eficazes contra o desmatamento. Nesta direção, o presente trabalho é uma extensão do estudo experimental apresentado por Noma, Korting e Fonseca (2013). As principais contribuições com relação ao trabalho anterior incluem: (1) dados do TerraClass 2010 para os experimentos, (2) tratamento de 3 ou mais classes e (3) um classificador adicional obtido através da combinação por votação dos cinco já existentes.

O texto está organizado da seguinte maneira. A Seção 2 descreve brevemente os classificadores envolvidos no estudo. A metodologia adotada para os experimentos é abordada na Seção 3. Os resultados experimentais são exibidos e interpretados na Seção 4. Comentários finais e futuras direções são exibidos na Seção 5.

2. Trabalhos relacionados

O foco deste trabalho é sobre classificação supervisionada. Neste caso, assume-se a existência de um conjunto de dados de treinamento, em que cada amostra tem a sua classe conhecida. Dado um conjunto de treinamento, é possível construir um classificador ou um modelo de classificação baseado nos atributos ou características das amostras. Desta forma, dado um conjunto de dados de entrada, em que a classe é desconhecida para cada amostra, a ideia é usar o modelo de classificação para determinar a classe de cada amostra de entrada.

Na literatura, existem diferentes técnicas de classificação supervisionada, conforme Theodoridis e Koutroumbas (2008), Theodoridis et al. (2010). Por exemplo, uma abordagem bastante comum é baseada no vizinho mais próximo (*Nearest-Neighbor*). Neste caso, a classificação é determinada por uma métrica definida no espaço das características. Baseado nesta métrica, o objetivo é classificar cada amostra de entrada de acordo com um protótipo do conjunto de treinamento associado à distância mínima.

Um outro método clássico na literatura é o método de Bayes. Este classificador é dito ótimo pois minimiza o erro de classificação no caso em que a distribuição dos dados e os parâmetros envolvidos são conhecidos. Na prática, raramente estas informações estão disponíveis, sendo comum assumir uma distribuição Gaussiana, cujos parâmetros devem ser estimados. Além disso, assumindo-se independência entre as variáveis resulta no método ingênuo (*Naive Bayes*), extremamente popular especialmente para dados com alta dimensionalidade.

Alternativamente, é comum encontrar ferramentas disponíveis baseadas em árvores de decisão (*Decision Trees*), por exemplo, Lang e Tiede (2007), ITT (2008), Costa et al. (2010), Korting (2012). As regras de classificação de uma amostra de entrada podem ser representadas por uma árvore de decisão. O processo de classificação inicia-se pela raiz e termina em uma folha. Cada folha da árvore determina a classe da amostra de entrada.

Outras ferramentas bastante populares e disponíveis são baseadas em máquina de vetores de suporte (*Support Vector Machine*), método descrito por Boser, Guyon e Vapnik (1992), Cortes e Vapnik (1995), Chang e Lin (2011). Originalmente desenhada para classificação binária, é possível lidar com classificação multiclasse através da combinação de várias máquinas ou classificadores. No caso mais simples, em que as duas classes são linearmente separáveis, a ideia é obter um hiperplano que maximize a margem de separação entre as duas classes para agregar robustez durante a classificação. Note que uma ideia similar pode ser usada para o caso geral, em que as classes não são linearmente separáveis.

Na literatura, também foram propostos métodos que são naturalmente projetados para o caso multiclasse. Por exemplo, com um único classificador baseado em caminhos ótimos em grafos (*Optimum Path Forest*), proposto por Papa (2008), Papa et al. (2012), é possível lidar

com múltiplas classes. Este método explora a conectividade no espaço de características através de um modelo de classificação representado por uma floresta de caminhos ótimos, em que cada árvore pode ter uma classe distinta. Neste caso, cada amostra de entrada é classificada de acordo com um protótipo de uma árvore que ofereça um caminho de menor custo.

Ao invés de se apoiar na classificação de um único método, existem problemas em que uma combinação de vários classificadores pode produzir resultados mais robustos. A técnica de combinar classificadores, conforme Valiant (1984), Kearns e Valiant (1994), se apoia em um conjunto de classificadores base, também chamados de classificadores fracos, para obter um classificador mais robusto. Existem diferentes maneiras de combinar classificadores. Neste trabalho, consideramos uma maneira simples de combinação através de votação: dada uma amostra de entrada e um conjunto de classificadores base, a sua classe será definida pelo maior número de votos.

Finalmente, para problemas envolvendo imagens, é possível obter uma classificação baseada em pixels ou em regiões representadas por grupos de pixels adjacentes. A segunda estratégia é inspirada na abordagem descrita por Blaschke, Johansen e Tiede (2011). Por um lado, a classificação baseada em pixels pode proporcionar um resultado de classificação com um maior nível de detalhes. Por outro, a classificação baseada em regiões pode agregar robustez, evitando a presença de pixels espúrios no resultado da classificação, promovendo uma presença maior de regiões homogêneas.

3. Metodologia

A metodologia segue o estudo anterior, descrito por Noma, Korting e Fonseca (2013). Dados multitemporais, extraídos de regiões no estado do Mato Grosso (Figura 1), são usados como características para classificar o tipo de uso dentro dos polígonos de desmatamento mapeados pelo PRODES. As classes são definidas pelos temas do TerraClass 2008 e 2010, utilizando-se imagens Landsat. As regiões selecionadas foram as mesmas do estudo anterior, com predomínio de agricultura, pastagem e floresta, conforme Huete et al. (2002).

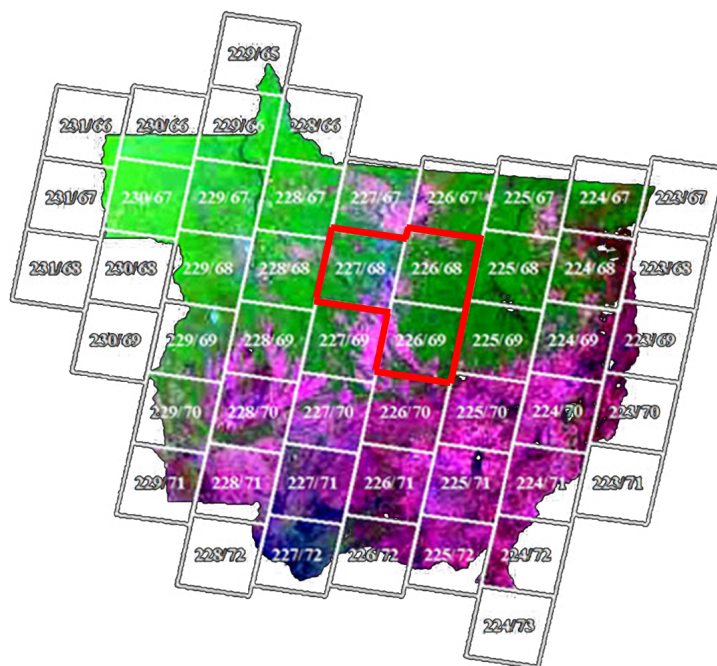


Figura 1: Regiões selecionadas para o experimento: 226/68, 226/69 e 227/68. (Adaptado de http://www.inpe.br/cra/projetos_pesquisas/terraclass2010.php)

Sequências de imagens do sensor MODIS, adquiridas em diferentes instantes no tempo, são usadas para formar perfis temporais EVI (*Enhanced Vegetation Index*). Cada pixel da sequência de imagens corresponde a um perfil anual da composição EVI de 8 dias do MODIS/Terra: uma sequência de 45 índices de vegetação ($365 / 8 = 45,625$).

Conforme apontado no trabalho anterior, a integração dos dados do TerraClass com os perfis temporais EVI envolveu resoluções espaciais distintas: imagens Landsat com resolução espacial de 30 metros e MODIS com resolução espacial de 250 metros. Para amenizar a imprecisão desta integração e para agregar robustez ao processo, a classificação foi baseada em regiões obtidas por segmentação usando o método de Baatz e Schape (2000), agrupando-se pixels adjacentes das imagens MODIS com perfis similares.

O experimento inclui cinco classificadores, Árvore de Decisão (DT), Bayes Ingênuo (NB), Vizinho Mais Próximo (NN), Máquina de Vetores de Suporte (SVM), Floresta de Caminhos Ótimos (OPF), e uma combinação deles por votação (VOT).

A avaliação é baseada em uma validação cruzada k -fold, com $k = 10$, seguindo por exemplo a descrição de Duda, Hart e Stork (2001). Neste caso, o conjunto de treinamento é particionado em 10 subconjuntos. Cada partição é usada como conjunto de teste e o seu complemento como conjunto de treinamento. Desta forma, é possível obter uma média de 10 medidas de desempenho para cada método. Seguindo a mesma medida de desempenho usada no trabalho anterior, a acurácia é definida pelo número de regiões classificadas corretamente dividido pelo total de regiões.

4. Resultados Experimentais

Para os experimentos, foram selecionadas as classes mais frequentes do TerraClass, com uma quantidade de amostras superior a 800. Cada amostra corresponde a uma região do processo de segmentação, usando-se os mesmos parâmetros anteriores: escala = 100, compacidade = 0,5, cor = 0,9 e mesmo peso para cada uma das 45 bandas. A Figura 2 exhibe exemplos de perfis temporais das classes selecionadas, ilustrando algumas dificuldades para classificação provenientes de similaridades entre perfis de classes distintas. Para cada região órbita/ponto, foram consideradas as seguintes classes:

- 226/68: (1) agricultura anual, (2) floresta e (3) pasto limpo, totalizando 3240 amostras para o ano de 2008 e a mesma quantidade para 2010;
- 226/69: (1) agricultura anual, (2) floresta e (3) não-floresta, totalizando 7380 amostras para 2008 e 6720 amostras para 2010;
- 227/68: (1) agricultura anual, (2) floresta, (3) não-floresta e (4) pasto limpo, totalizando 3480 amostras para 2008 e 5040 amostras para 2010.

Os resultados experimentais são exibidos nas Tabelas 1, 2 e 3, considerando-se a acurácia restrita a cada classe. Nas tabelas, os índices de acurácia foram multiplicados por 100 para facilitar a visualização.

A comparação envolve diferentes classificadores. Dentre os classificadores considerados, existem aqueles que são especificamente desenvolvidos para problemas com duas classes. Para uma comparação uniforme, adotou-se a estratégia “um contra todos”, possibilitando que classificadores binários tratem $n \geq 3$ classes. Especificamente, é possível combinar n classificadores, um classificador para identificar cada classe de interesse.

De maneira geral, os resultados foram compatíveis com o estudo anterior. Os melhores resultados foram obtidos pelo SVM e pelo método de votação (VOT).

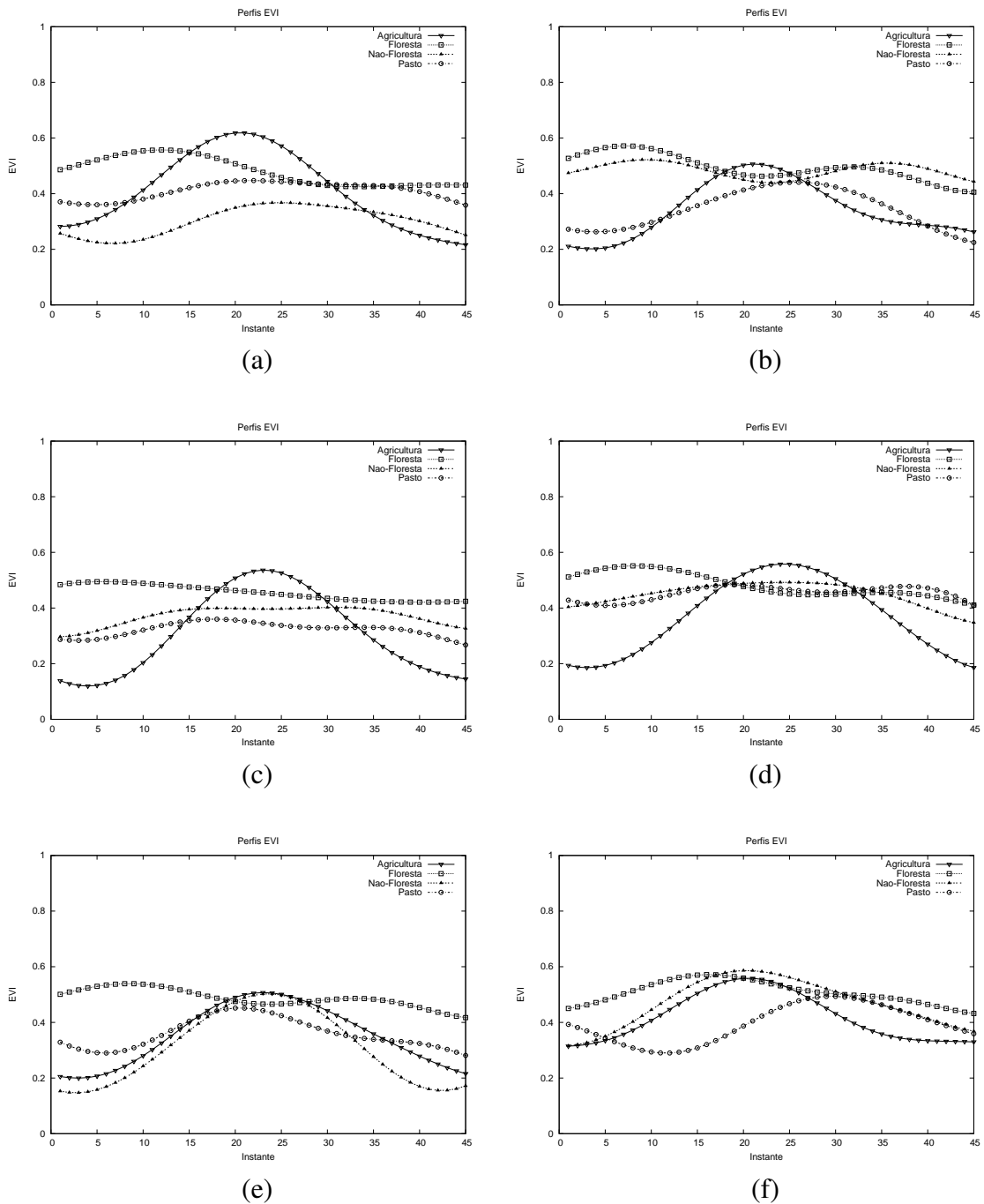


Figura 2: Exemplos de perfis EVI de amostras extraídas da região 227/68. (a) Perfis distintos. (b) Floresta e não-floresta similares, agricultura e pasto similares. (c) Pasto e não-floresta similares. (d) Perfil distinto de agricultura. (e) Perfil distinto de floresta. (f) Perfil distinto de pasto.

Tabela 1: Valores médios de acurácia com os respectivos desvios padrão para a região 226/68.

	Método	226/68		
		Classe 1	Classe 2	Classe 3
2008	NB	77,22 ± 4,43	89,26 ± 2,48	84,26 ± 2,62
	NN	73,70 ± 2,19	91,20 ± 1,87	78,61 ± 3,42
	DT	75,09 ± 3,25	91,48 ± 2,42	75,09 ± 4,44
	OPF	73,43 ± 2,69	91,20 ± 2,28	77,32 ± 3,68
	SVM	78,80 ± 3,64	93,43 ± 2,49	90,56 ± 3,38
	VOT	78,51 ± 3,32	92,59 ± 2,15	86,30 ± 2,15
2010	NB	77,04 ± 2,72	92,50 ± 2,91	83,61 ± 3,12
	NN	75,19 ± 4,11	91,48 ± 2,89	76,76 ± 4,26
	DT	77,32 ± 4,89	90,46 ± 3,24	75,19 ± 2,75
	OPF	74,54 ± 4,26	91,30 ± 2,90	75,83 ± 4,42
	SVM	80,19 ± 2,90	92,87 ± 1,91	90,83 ± 2,02
	VOT	79,35 ± 4,06	92,87 ± 1,99	85,37 ± 2,44

Tabela 2: Valores médios de acurácia com os respectivos desvios padrão para a região 226/69.

	Método	226/69		
		Classe 1	Classe 2	Classe 3
2008	NB	79,43 ± 3,59	89,51 ± 2,96	70,98 ± 3,54
	NN	74,31 ± 2,42	86,50 ± 1,75	67,68 ± 3,38
	DT	71,42 ± 1,79	85,49 ± 2,88	63,90 ± 3,49
	OPF	72,56 ± 3,06	86,26 ± 1,87	66,79 ± 3,59
	SVM	91,26 ± 1,81	92,81 ± 1,62	85,17 ± 2,45
	VOT	83,92 ± 1,97	90,61 ± 1,90	75,65 ± 1,69
2010	NB	80,22 ± 2,22	92,55 ± 2,08	70,18 ± 4,11
	NN	71,03 ± 3,82	85,27 ± 1,58	69,06 ± 3,97
	DT	71,38 ± 4,18	84,20 ± 2,49	64,24 ± 3,67
	OPF	69,91 ± 3,90	84,38 ± 1,62	67,90 ± 3,72
	SVM	89,91 ± 1,62	91,16 ± 2,09	82,37 ± 2,62
	VOT	81,30 ± 3,78	89,87 ± 1,13	76,25 ± 3,21

Tabela 3: Valores médios de acurácia com os respectivos desvios padrão para a região 227/68.

	Método	227/68			
		Classe 1	Classe 2	Classe 3	Classe 4
2008	NB	80,46 ± 5,05	87,36 ± 1,95	47,59 ± 4,31	86,10 ± 4,64
	NN	77,70 ± 3,64	79,77 ± 4,51	62,18 ± 4,32	84,14 ± 2,64
	DT	75,06 ± 2,77	79,43 ± 3,36	60,58 ± 5,79	83,79 ± 2,26
	OPF	74,83 ± 3,73	78,85 ± 4,76	62,41 ± 4,27	83,79 ± 3,13
	SVM	87,47 ± 3,57	93,45 ± 1,88	72,07 ± 6,11	91,15 ± 3,47
	VOT	84,88 ± 3,26	89,43 ± 1,98	64,94 ± 3,99	89,66 ± 2,57
2010	NB	82,22 ± 3,33	86,91 ± 3,48	49,60 ± 7,17	83,73 ± 2,93
	NN	75,32 ± 3,60	79,13 ± 3,28	61,19 ± 3,81	79,52 ± 5,03
	DT	74,76 ± 3,36	79,60 ± 4,60	59,52 ± 5,96	78,02 ± 3,74
	OPF	73,25 ± 3,51	78,10 ± 3,76	61,35 ± 3,78	78,57 ± 4,76
	SVM	84,68 ± 3,43	90,64 ± 1,97	65,56 ± 4,10	91,67 ± 3,29
	VOT	82,64 ± 3,77	87,14 ± 3,31	62,70 ± 2,88	87,30 ± 3,67

5. Conclusão

As principais contribuições deste trabalho foram: (1) inclusão de novos dados nos experimentos, provenientes do TerraClass 2010; (2) tratamento do caso multiclasse e (3) inclusão de uma estratégia baseada em votação para combinar diferentes classificadores. Os experimentos indicam um destaque para os métodos SVM e VOT.

No futuro, como uma possível extensão, é interessante comparar diferentes estratégias de combinação de classificadores para verificar um possível ganho nos resultados. Neste caso, ao invés de se apoiar na classificação de um único método, seria possível comparar diferentes maneiras de combinar vários classificadores fracos/base para produzir um classificador mais robusto.

Uma outra extensão interessante seria identificar a existência de um deslocamento temporal de perfis conforme a distância espacial aumenta entre dois pixels na sequência MODIS. Neste caso, uma alternativa seria transformar as características/perfis para o domínio das frequências para possivelmente evitar imprecisões devido ao deslocamento temporal entre perfis e melhorar os resultados de classificação.

Agradecimentos: Ao prof. Bernardo Rudorff pela disponibilização dos dados EVI usados nos experimentos. Ao prof. João Papa pela ajuda com o método OPF.

Referências

- BAATZ, M.; SCHAPE, A. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. In: ANGEWANDTE GEOGRAPHISCHE INFORMATIONSVARBEITUNG XII. [S.l.], 2000.
- BLASCHKE, T.; JOHANSEN, K.; TIEDE, D. Object-based image analysis for vegetation mapping and monitoring. In: ADVANCES IN ENVIRONMENTAL REMOTE SENSING: SENSORS, ALGORITHMS, AND APPLICATIONS. [S.l.]: CRC Press, Taylor and Francis, 2011.
- BOSER, B. E.; GUYON, I.; VAPNIK, V. A training algorithm for optimal margin classifiers. In: PROCEEDINGS OF THE FIFTH ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY. [S.l.]: ACM Press, 1992. p. 144–152.
- CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, v. 2, n. 3, p. 1–27, 2011.
- CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995.
- COSTA, G. et al. Knowledge-based interpretation of remote sensing data with the interimage system: major characteristics and recent developments. In: GEOBIA. [S.l.], 2010.
- DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. Second. [S.l.]: Wiley, 2001.
- HUETE, A. et al. Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote Sensing of Environment*, v. 83, n. 1–2, p. 195–213, 2002.
- ITT. Envi feature extraction module user's guide. In: GILCHING, GERMANY: EXELIS VISUAL INFORMATION SOLUTIONS. [S.l.], 2008.
- KEARNS, M.; VALIANT, L. G. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM*, v. 41, n. 1, p. 67–95, 1994.
- KORTING, T. S. *GeoDMA: a toolbox for data mining, object-based and multi-temporal analysis of remote sensing imagery*. Orientadores: L. M. G. Fonseca, G. Câmara. Tese (Doutorado) — INPE, 2012.
- LANG, S.; TIEDE, D. Definiens developer. *GIS Business* 9, p. 34–37, 2007.

NOMA, A.; KORTING, T. S.; FONSECA, L. M. G. Uma comparação entre classificadores usando regiões e perfis evi para agricultura. In: *Anais do XVI Simpósio Brasileiro de Sensoriamento Remoto*. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2013. p. 2250–2257.

PAPA, J. P. *Classificação supervisionada de padrões utilizando floresta de caminhos ótimos*. Orientador: A. X. Falcão. Tese (Doutorado) — Unicamp, 2008.

PAPA, J. P. et al. Efficient supervised optimum-path forest classification for large datasets. *Pattern Recognition*, v. 45, n. 1, p. 512–520, 2012.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition*. Fourth. [S.l.]: Academic Press, 2008.

THEODORIDIS, S. et al. *Introduction to Pattern Recognition: A Matlab Approach*. First. [S.l.]: Academic Press, 2010.

VALIANT, L. G. A theory of the learnable. *Communications of the ACM*, v. 27, n. 11, p. 1134–1142, 1984.