Proposta de Classificadores Semissupervisionados baseados em Rotulação de Agrupamentos via Distâncias Estocásticas

Gabriela Ribeiro Sapucci[1]
Rogério Galante Negri[1]

[1] Universidade Estadual Paulista – UNESP
Instituto de Ciência e Tecnologia – Departamento de Engenharia Ambiental
Rodovia Presidente Dutra, km 137,8 – 12247-004 – São José dos Campos – SP, Brasil
gabrielasapucci@gmail.com
rogerio.negri@ict.unesp.br

Abstract. Remote sensing image classification is one of the most important applications of Pattern Recognition in environmental studies. Image classification methods generally have supervised learning or unsupervised. As supervised learning methods perform sorting by means of a function or decision rule modeled through information provided in advance, the quality of the results is directly related to the quality of the set of training standards, which doesn't always guarantee quality results. Unsupervised learning, in turn, build your knowledge in function of analogies observed about the data, which can be a complex task. Alternatively, the semi-supervised learning aims to deal with the weaknesses of both paradigms, by combining concepts of learning with and without supervision. In this context, this research project proposes the formalization and implementation of two methods of semi-supervised classification, which combines classic tools in the area of pattern recognition: the Hierarchical Divisive Algorithms, K-Means and stochastic distances. From a set of groups, defined by the combination of Hierarchical Divisive Algorithm and K-Means and another defined only by K-Means, through unsupervised learning, stochastic distances are used for labeling of each of these groups. Through case studies on the use and classification of ground cover around the Tapajós National Forest, the quality of the results obtained according to the Kappa coefficient was analyzed and the proposed methods were compared with other classification methods already known in the literature.

Keywords: stochastic distances, semi-supervised learning, image classification distâncias estocásticas, aprendizado semissupervisionado, classificação de imagem

1. Introdução

A área da computação denominada Reconhecimento de Padrões se baseia na busca por padrões em bases de dados (Meneses e Almeida 2012). Dentre as diferentes aplicações de Reconhecimento de Padrões, encontra-se a classificação de imagens de Sensoriamento Remoto com aplicações, por exemplo, no monitoramento de florestas, rios e em estudo de áreas afetadas por desastres naturais. O desenvolvimento de diferentes técnicas de classificação digital de imagem contribuiu para a automatização do processo de extração de informações de imagens, eliminando a subjetividade da interpretação humana, além de reduzir esforços de trabalho do analista (Meneses e Almeida 2012).

Os métodos de classificação de imagens distinguem-se através de paradigmas de aprendizagem, dentre os quais, os mais comumente utilizados são os aprendizados supervisionado e não supervisionado. O aprendizado supervisionado, modelado através de informações fornecidas a $priori$, difere-se do não supervisionado, uma vez que este fundamenta-se em analogias observadas sobre os padrões de dados.

Dependendo do objetivo ou da área de estudo a ser classificada, determinados algoritmos apresentam limitações e não são capazes de gerar resultados satisfatórios. Tendo em vista as fragilidades encontradas nos processos de classificação de imagens, uma alternativa é o

aprendizado semissupervisionado, o qual combina conceitos dos aprendizados com e sem supervisão.

No aprendizado semissupervisionado é necessário uma pequena quantidade de dados rotulados para a construção dos modelos, o que diminui consideravelmente os custos com a classificação de imagens (Santos 2012).

Diante do contexto apresentado, a presente pesquisa tem como objetivo implementar dois métodos fundamentados em conceitos de aprendizado semissupervisionado indireto, para rotulação de agupamentos através de distâncias estocásticas, sendo tais agrupamentos determinados com o emprego de técnicas de classificação não supervisionada. Os algoritmos de agrupamento utilizados serão o K-Médias, no primeiro método, e o Algoritmo Hierárquico Divisivo associado ao K-Médias, no segundo método. Estes métodos serão comparados com outros presentes na literatura, quanto ao seu desempenho na classificação da imagem da área de estudo.

Para análise da qualidade dos classificadores será realizado um estudo de caso no entorno da Floresta Nacional do Tapajós. As imagens classificadas geradas serão comparadas com classificações provenientes dos métodos Máxima Verossimilhança (MV) e Mínima Distância Euclidiana (MDE).

2. Algoritmos de Agrupamento

Os algoritmos de agrupamento particionam um conjunto de padrões em grupos, de acordo com alguma relação de similaridade (Manning e Schutze 1999). As técnicas de agrupamento possibilitam a construção de importantes ferramentas para a análise exploratória de dados, sobretudo para os casos em que existe pouco ou nenhum conhecimento prévio (Jain e Dubes 1988). O Algoritmo Hierárquico e o K-Médias são exemplos de algoritmos de agrupamento.

Os Algoritmos Hierárquicos particionam o conjunto de dados sucessivamente, gerando uma representação hierárquica dos agrupamentos, de acordo com o grau de semelhança. Estes algoritmos podem ser classificados de duas formas: aglomerativos e divisivos. Especialmente, o Algoritmo Hierárquico Divisivo (AHD) inicia-se com um único conjunto, o qual é particionado em agrupamentos menores, recursivamente e com o emprego de algum algoritmo de agrupamento, até atingir um critério de parada (Xavier 2012). Os critérios de parada do AHD são geralmente parâmetros pré-definidos, como o diâmetro dos agrupamentos e o número mínimo de elementos por agrupamento.

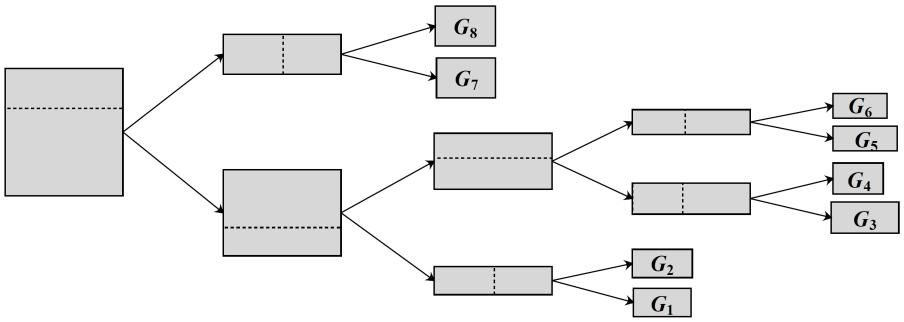


Figura 1: Formação dos agrupamentos segundo AHD. A dimensão dos retângulos expressa a noção de diâmetro e número de elementos dos agrupamentos.

Com relação ao algoritmo K-Médias, o conjunto de dados é particionado em K agrupamentos. De modo iterativo, cada elemento do conjunto de dados é associado com o centroide mais próximo. Os centroides, definidos como representantes dos grupos, são

atualizados a cada iteração de acordo com a média dos elementos associados a ele na etapa anterior, até que ocorra a convergência (Webb 2002).

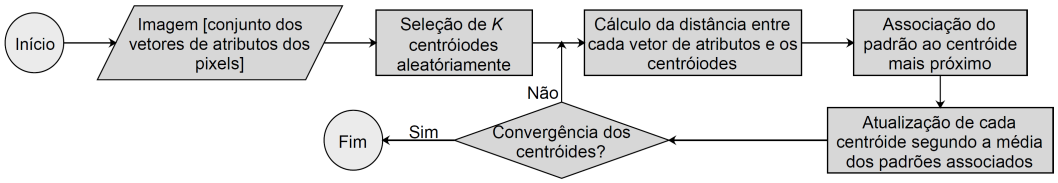


Figura 2: Fluxograma do algoritmo K-Médias.

Os agrupamentos gerados a partir destas técnicas não possuem uma classe temática associada. Deste modo, a classificação destes agrupamentos pode ser realizada através de um processo de classificação supervisionada por mínima distância estocástica, na qual uma determinada classe é atribuída a um agrupamento com base na menor distância estocástica entre as distribuições estatísticas que modelam o conjunto de padrões deste agrupamento e as que representam os padrões de cada classe considerada.

3. Rotulação de Agrupamentos através de Distâncias Estocásticas

Nesta seção são propostos dois classificadores de imagem fundamentados em conceitos de aprendizado semissupervisionado indireto. A classificação de imagens segundo as duas propostas sugere a utilização do Classificador de Mínima Distância Estocástica para a rotulação automática de agrupamentos. Na primeira proposta, estes agrupamentos são antes determinados através do Algoritmo Hierárquico Divisivo (AHD) e na segunda proposta, os agrupamentos são antes determinados através do K-Médias. As subdivisões efetuadas no AHD são realizadas pelo K-Médias com $K = 2$.

Uma vez definidos os agrupamentos pelos métodos mencionados, o processo de rotulação destes agrupamentos é conduzido com base no processo de classificação por Mínima Distância Estocástica. Este processo de classificação considera um conjunto de treinamento contendo exemplos de c classes distintas, representado por $\mathcal{D} = \{(\mathbf{x}_i, \omega_j) \in \mathcal{X} \times \Omega;\ i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, c\}$ e agrupamentos identificados a partir dos padrões presentes em $\mathcal{I}$, definidos por $G_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}\}$, com $i = 1, \ldots, k$ e n_i correspondendo ao i-ésimo agrupamento, $\mathbf{x}$ constituindo padrões representantes dos vetores de atributos dos pixels de uma imagem $\mathcal{I}$, $\mathcal{X}$ representando o espaço de atributos e Ω indicando o conjunto de classes. G_i é rotulada de acordo com o conjunto de treinamento segundo a regra de decisão:

$$(G_i, \omega_j) \Leftrightarrow j = \underset{j=1,\ldots,c}{\arg\min}\, \mathrm{B}(f_{G_i}, f_{\omega_j}), \tag{1}$$

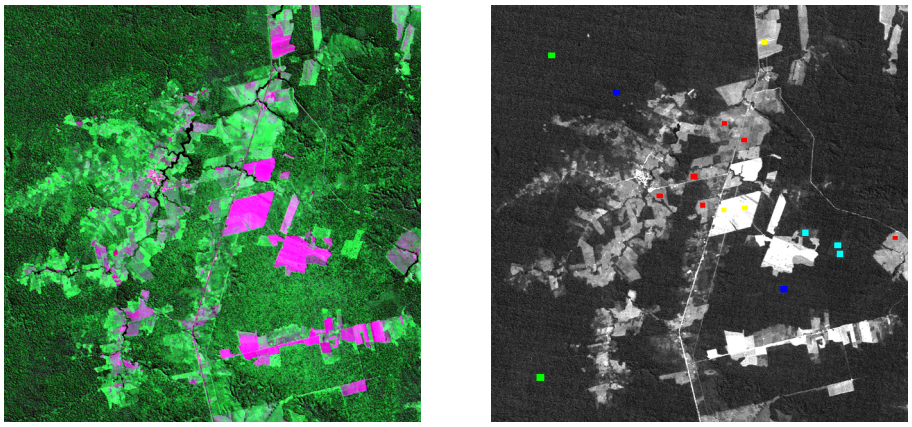
onde f_{G_i} e f_{ω_j} são funções densidade de probabilidade que modelam a distribuição dos padrões de G_i e dos padrões rotulados em $\mathcal{D}$ associados a ω_j, respectivamente. $B(\cdot, \cdot)$ é a distância estocástica de Bhattacharyya (Richards e Richards 1999).

Diante o exposto, os classificadores propostos baseiam-se na junção entre os paradigmas supervisionado e não supervisionado. Enquanto a determinação de agrupamentos é realizada de forma não supervisionada, na primeira proposta pelo AHD associado ao K-Médias e na segunda pelo K-Médias, a associação de uma classe a cada um destes agrupamentos é realizada através de um processo supervisionado, por Mínima Distância Estocástica.

Para fins de simplificação, o método proposto Algoritmo Hierárquico Divisivo com K-Médias associado à Distância Estocástica de Bhattacharyya será tratado como AHD+KM+DE e o K-Médias associado à Distância Estocástica de Bhattacharyya será chamado de KM+DE.

4. Experimentos e Resultados

Nesta seção é apresentado um estudo de caso a fim de verificar a performance dos métodos propostos. Para a relização dos experimentos foi empregada uma imagem referente ao entorno da Floresta Nacional do Tapajós, no estado do Pará, obtida através do satélite LANDSAT-5 TM (óptico - resolução espacial: 30 m). A área de estudo adotada abrange uma porção de 650×650 pixels, extraída da imagem original. Além da imagem, foram utilizadas informações de diferentes classes de uso e cobertura do solo distribuídas na área de estudo (Figura 3). A Tabela 1 mostra as classes empregadas na classificação e a quantidade de pixels por classe presentes nas amostras de treinamento e validação indicadas na Figura 3(b). Foram utilizados o software ENVI 4.7, empregado na coleta de amostras da imagem LANDSAT-5 TM da área de estudo, e a linguagem de programação IDL 7.1 (Interactive Data Language) para implantação dos métodos. Ainda, utilizou-se da biblioteca de funções SLIC (Negri 2013). Esta biblioteca é programada em IDL e é composta por funções de uso comum em diferentes etapas que envolvem os procesamentos de classificação de imagens.



(a) Área de estudo

(b) Distribuição espacial das amostras de treinamento e validação sobre a área de estudo

Figura 3: Imagem LANDSAT-5 TM da área de estudo, em composição colorida R(5)G(3)B(4) e em escala de cinza.

		Quantidade de pixels	
Cor	***Classe***	***Treinamento***	***Validação***
●	Floresta	211	262
●	Agricultura	177	252
●	Regeneração Antiga	180	262
●	Pastagem	356	147
●	Regeneração Jovem	270	216

Tabela 1: Amostras de treinamento e validação.

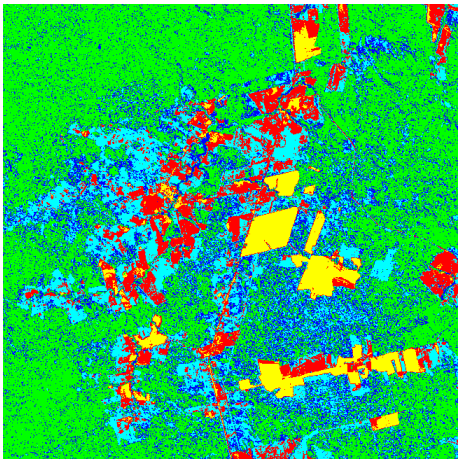
Na classificação da imagem empregada no estudo considerou-se um cenário com 4 conjuntos de treinamento de dimensões distintas, sendo compostos por 10, 15, 25 e 50 padrões rotulados por classe. Ainda, vale mencionar que foram testadas diferentes configurações de parâmetros para os métodos KM+DE e AHD+KM+DE, isto é, adotou-se diferentes valores

de K para o KM+DE e diferentes valores de diâmetro e número mínimo de elementos por agrupamento para o AHD+KM+DE.

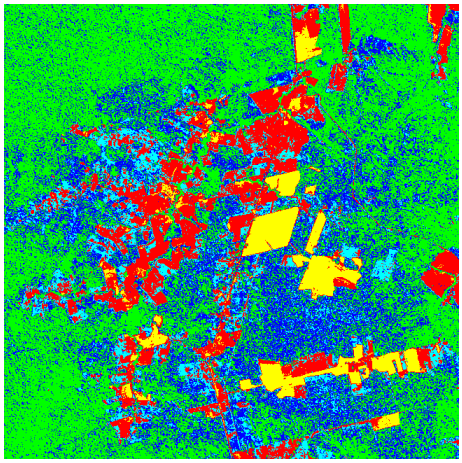
Os experimentos realizados visaram comparar a eficácia dos modelos de classificação semissupervisionada indireta propostos (AHD+KM+DE e KM+DE) com os classificadores de Máxima Verossimilhança (MV) e Mínima Distância Euclidiana (MDE) disponíveis na literatura. Detalhes a respeito destes classificadores são encontrados em (Webb 2002). Os resultados da classificação da imagem da área de estudo foram avaliados segundo o coeficiente Kappa (Congalton e Green 1999), a fim de verificar as acurácias apresentadas pelos métodos propostos.

A partir dos classificadores AHD+KM+DE, KM+DE, MV e MDE, foram geradas para cada um quatro imagens classificadas, sendo estas correspondentes aos conjuntos de treinamentos compostos por, 10, 15, 25 e 50 pixels por classe. Estes conjuntos foram definidos a partir do conjunto original através de um processo de seleção aleatória. Em especial, no método KM+DE foram obtidos 12 resultados de classificação, uma vez que cada um dos treinamentos foi testado considerando três quantidades diferentes de centroides: 10, 20 e 40.

Na Figura 4 é apresentada a classificação de melhor acurácia para o método AHD+KM+DE e para o método KM+DE. Houve diferenças no comportamento dos dois classificadores, o KM+DE se mostrou mais eficiente na classificação considerando 10 pixels por classe, enquanto o AHD+KM+DE apresentou a classificação mais acurada para o treinamento contendo 25 pixels por classe. O KM+DE mostrou-se ainda mais acurado na classificação do que o AHD+KM+DE, sobretudo na representação da classe Floresta.



(a) KM+DE (40 centroides) - Treinamento composto por 10 pixels



(b) AHD+KM+DE - Treinamento composto por 25 pixels

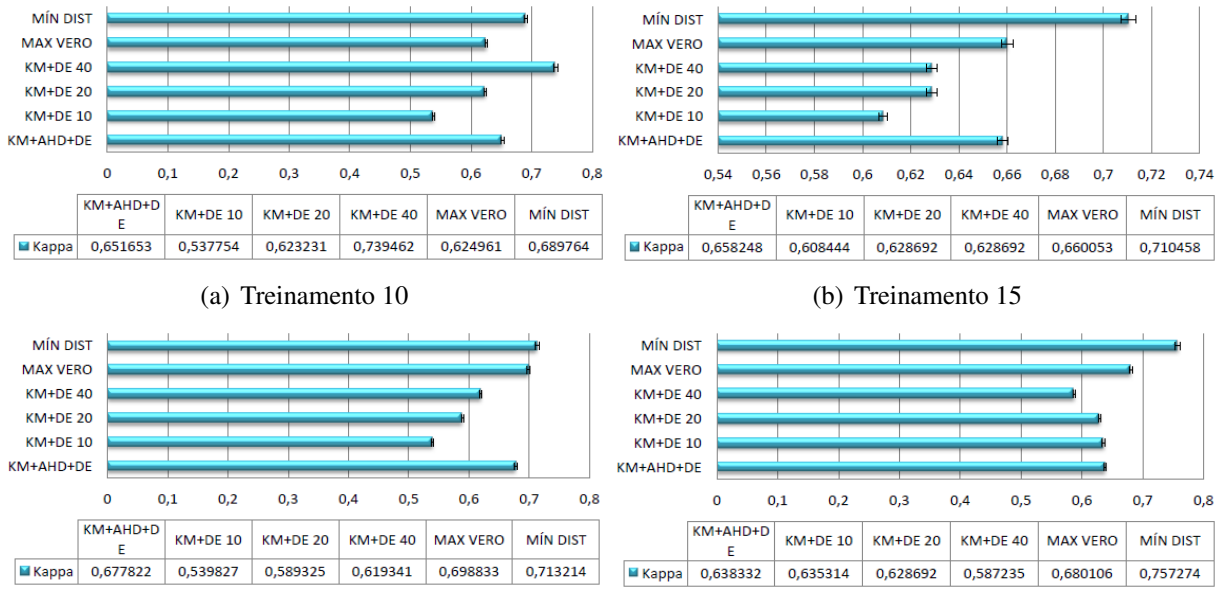
Figura 4: Resultados de classificação com melhor acurácia dos métodos KM+DE e AHD+KM+DE.

No método AHD+KM+DE, a qualidade da imagem classificada apresentou-se dependente do valor definido para o diâmetro dos agrupamenros e do número mínimo de elementos por agrupamento. Levando em consideração este fator, foram feitos testes para 5 diâmetros: 0,10, 0,25, 0,50, 0,75 e 0,90; para cada diâmetro testou-se 3 diferentes números mínimos de elementos por agrupamento: 200, 10000 e 100000. No total foram realizadas 60 classificações, sendo 15 para cada conjunto de treinamento (i.e., 10, 15, 25, 50). Para a comparação da eficiência deste método em relação aos outros, adotou-se os parâmetros responsáveis pela obtenção do maior Kappa, cujo diâmetro foi igual a 0,1 e o número mínimo de elementos por agrupamento igual a 10000. A Figura 5 representa a comparação entre os métodos,

segundo o coeficiente Kappa e de acordo com o número de pixels utilizados por classe em cada treinamento.

De acordo com os índices Kappas obtidos por cada método, verifica-se que o AHD+KM+DE não foi considerado o algoritmo de melhor desempenho para a classificação da área de estudo. Entretanto, percebe-se o seu potencial como classificador, uma vez que a classificação gerada apresentou valor próximo aos algoritmos já descritos na literatura. Devido à influência de parâmetros pré-estabelecidos na eficácia da classificação, pode-se afirmar que a busca por melhores atribuições de valores a estas variáveis podem gerar resultados de maior qualidade.

Todos os métodos empregados no estudo variaram de acordo com a quantidade de pixels por classe no conjunto de treinamento, o que é evidenciado pelos diferentes Kappas. Ainda, o KM+DE apresentou variações quanto ao número de centroides pré-definidos, o que indica que uma escolha de K ideal para a área de estudo possa gerar classificações com qualidade mais elevada.



(a) Treinamento 10 (b) Treinamento 15

(c) Treinamento 25 (d) Treinamento 50

Figura 5: Comparação entre os coeficientes Kappa.

5. Conclusões e Perspectivas Futuras

Os classificadores propostos, baseados em distâncias estocásticas para a classificação de agrupamentos determinados pelo algoritmo de agrupamento AHD e por K-Médias, mostraram-se eficientes na classificação de imagens. Embora os resultados obtidos segundo o coeficente Kappa nos quatros treinamentos não tenham sido maiores em todos os casos, acredita-se ainda que exista potencial para obtenção de classificações com qualidade superior.

Como perspectivas futuras, serão verificados novos parâmetros capazes de aprimorar os resultados. Essa busca por parâmetros ocorrerá concominantemente à realização de testes em diferentes cenários compostos pelos conjuntos de treinamento de 10, 15, 25 e 50 pixels por classe, visando conhecer as características e parâmteros ideais para uma maior acurácia nos resultados da classificação de imagens.

Agradecimentos

Os autores agradecem à FAPESP (procs. 2014/14830-8 e 2016/06242-4) pelo auxílio financeiro.

Referências

CONGALTON, R. G.; GREEN, K. *Assessing the accuracy of remotely sensed data:* principles and practices. New York: Lewis Publisher, 1999. 137 p.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. [S.l.: s.n.], 1988.

MANNING, C. D.; SCHUTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT Press, 1999.

MENESES, P. R.; ALMEIDA, T. Introdução ao processamento de imagens de sensoriamento remoto. *Brasília: UNB/CNPq*, 2012.

NEGRI, R. G. *Máquina de Vetores de Suporte Adaptativa ao Contexto: formalização e aplicações em Sensoriamento Remoto*. 166 p. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013.

RICHARDS, J. A.; RICHARDS, J. *Remote sensing digital image analysis*. [S.l.]: Springer, 1999.

SANTOS, A. d. M. Investigando a combinação de técnicas de aprendizado semissupervisionado e classificação hierárquica multirrótulo. Universidade Federal do Rio Grande do Norte, 2012.

WEBB, A. R. *Statistical Pattern Recognition*. 2nd. ed. Chichester: John Wiley & Sons, 2002. 514 p.

XAVIER, V. L. *Resolução do Problema de Agrupamento segundo o Critério de Minimização da Soma de Distâncias*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2012.