Aplicação de técnicas de classificação semissupervisionada para análise de séries multitemporais de imagens de satélite

Luciana Alvim Santos Romani[1]
Bruno Ferraz do Amaral[2]
Renata Ribeiro do Valle Gonçalves[3]
Jurandir Zullo Jr.[3]
Elaine Parros Machado de Sousa[2]

[1] Embrapa Informática Agropecuária
Caixa Postal 6041 – 13083-886 – Campinas – SP, Brasil
luciana.romani@embrapa.br

[2] Universidade de São Paulo – ICMC/USP
13566-590 – São Carlos - SP, Brasil
brunoslash@grad.icmc.usp.br; parros@icmc.usp.br

[3] Universidade Estadual de Campinas – Cepagri/Unicamp
13083-970 – Campinas – SP, Brasil
renata, jurandir@cpa.unicamp.br

Abstract. In the last years remote sensing has become an important tool to support the agricultural crop monitoring in the whole world. In the most of studies involving agriculture specialists have applied medium and high spatial resolution satellites, preferentially. However, an important question is: can low spatial resolution satellites also be used in this monitoring task? In this context, the focus of this work is how to take advantage of the high temporal resolution which is characteristic of this kind of satellite to identify agricultural fields such as sugarcane. Accordingly, this paper proposes to apply two semi-supervisioned classification methods to classify multitemporal satellite images. Thus, we adapted the LNP (Linear Neighborhood Propagation) and HC-LGT (Hierarchical Clustering and Local Graph Transduction) methods to be employed in classification of time series extracted from NDVI-NOAA images. The training dataset was generated with time series of six (agricultural crops, sugarcane, urban areas, forest, pasture and perennial crops) different classes defined by agrometeorologists. The study area encloses state of São Paulo in Brazil where is cultivated large sugarcane fields. Results with both classification methods showed that time series from low spatial resolution satellites can be satisfactorily used to identify regions of agricultural fields as well as forests, pasture and urban areas. Additionally, the classification generated by both methods can also be used to identify the vegetation cover of a specific region as an initial step before applying more complex strategies.

Keywords: remote sensing, image processing, agriculture sensoriamento remoto, processamento de imagens, agricultura

1. Introdução

O agronegócio brasileiro é responsável por mais de 20% do Produto Interno Bruto (PIB) e valor superior a 35% das exportações, segundo dados oficiais (IBGE). Dessa forma, é de fundamental importância o investimento em pesquisas tecnológicas para aprimorar o rendimento das culturas agrícolas uma vez que a agricultura é economicamente estratégica para o país. Além disso, um monitoramento eficiente das safras agrícolas também contribui para apoiar a tomada de decisão no campo.

Uma das técnicas que vêm se destacando no acompanhamento de áreas de cultivo agrícola

no Brasil, que é um país de grandes dimensões, é o sensoriamento remoto. Dados de sensores orbitais têm se mostrado adequados para a identificação de áreas agrícolas, acompanhamento de safras dentre outras aplicações. Os sensores que imageiam a superfície terrestre produzindo séries de imagens multitemporais como o MODIS/TERRA (a partir de 2000) e AVHRR/NOAA (a partir de 1970) permitem a geração de índices de vegetação que possibilitam a identificação do estágio da cultura no campo pois indicam a biomassa da vegetação. Além disso, a aquisição diária permite o uso de imagens compostas temporais, o que remove em grande parte a interferência de nuvens.

Ao mesmo tempo que a disponibilidade de grandes volumes de dados de sensores orbitais é uma vantagem em pesquisas envolvendo agricultura, a capacidade humana de processar grandes quantidades de dados complexos (como conjuntos volumosos de séries temporais extraídas de imagens de satélite) torna esse processo de análise demorado e, por vezes, inviável. Nesse cenário, a utilização de técnicas computacionais que facilitem e agilizem o processo de análise dos dados se torna um importante auxílio aos especialistas (GANGULY; STEINHAEUSER, 2008).

Além disso, a incorporação dessas técnicas e recursos computacionais em ferramentas que permitam a manipulação, análise e visualização de dados de sensoriamento remoto é de grande importância para tornar pesquisas em agrometeorologia mais ágeis e flexíveis (DATCU et al., 2003; LI; NARAYANAN, 2004; DASCHIEL; DATCU, 2005; ROMANI et al., 2009, 2011). Uma das técnicas que vem sendo utilizadas é a classificação (HAN; KAMBER, 2001; RATANAMAHATANA; KEOGH, 2004). Usualmente, em problemas de classificação, um algoritmo de aprendizado supervisionado considera um conjunto de dados rotulados por especialistas (conjunto de treinamento), e visa associar cada instância desconhecida de um conjunto de dados a uma ou mais classes pré-definidas. Entretanto, o processo de supervisão de dados realizado pelo especialista para obtenção do conjunto de treinamento (rotulação dos dados) é, em geral, muito custoso. Assim, em muitos problemas reais, a quantidade de dados rotulados, ou supervisionados, pode ser extremamente pequena em relação à quantidade de dados desconhecidos a serem analisados, dificultando a tarefa de classificação e podendo levar a resultados menos precisos.

Nesse cenário, o aprendizado semissupervisionado visa construir e treinar um classificador baseado tanto nos conjuntos de dados rotulados por especialistas, quanto de dados não rotulados. Técnicas de classificação semissupervisionada têm sido utilizadas em problemas reais em que a quantidade de dados rotulados é muito pequena, com maior acurácia em relação à aplicação de técnicas de classificação puramente supervisionadas (WEI; KEOGH, 2006; NIGAM; GHANI, 2000).

No contexto da análise de séries temporais de dados extraídos de imagens de satélite, a supervisão dos dados por especialistas é custosa e muito demorada, configurando-se um problema adequado para aplicação de técnicas de classificação semissupervisionada. Logo, este trabalho apresenta uma comparação de dois algoritmos de classificação semissupervisionada utilizados para auxiliar na identificação de áreas de cultivo de cana-de-açúcar, uma importante commoditie brasileira. As técnicas foram incorporadas ao software SatImagExplorer, que foi desenvolvido para auxiliar na extração de séries temporais de imagens de satélite (CHINO; ROMANI; TRAINA, 2010). Os resultados indicam que ambas as técnicas apresentaram resultados satisfatórios para classificação de diferentes classes usando séries de imagens de baixa resolução espacial.

O restante desse artigo descreve a Metodologia do Trabalho na Seção 2, discute os Resultados na Seção 3 e finalmente apresenta as Conclusões na Seção 4.

2. Metodologia de Trabalho

A Figura 1 apresenta o fluxograma com a metodologia do trabalho que envolve (i) a seleção da área de estudo, (ii) a extração das séries temporais de NDVI das imagens do satélite NOAA, (iii) a classificação das séries usando duas técnicas semissupervisionadas implementadas na ferramenta SatImagExplorer e finalmente (iv) a visualização da classificação e a análise dos resultados. Todas essas etapas são detalhadas a seguir.

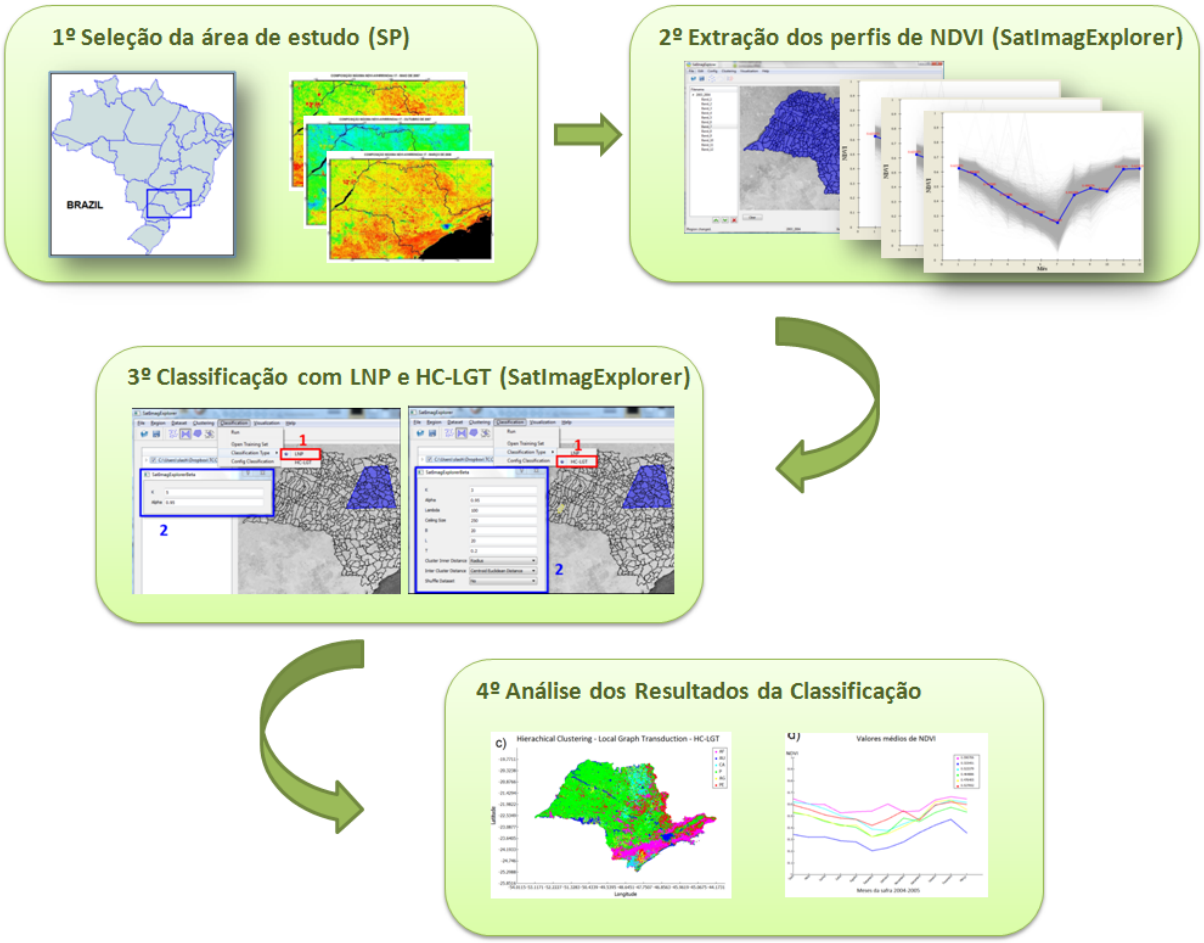

Figura 1: Fluxograma com as etapas da metodologia executada no trabalho.

O estado de São Paulo, que está situado entre as coordenadas geográficas 54º e 43º30' de longitude oeste e 25º30' e 19º30' de latitude sul, foi escolhido como área de estudo para este trabalho, por ser responsável por grande parte da produção de cana-de-açúcar do país. Além disso, a cultura está em expansão sendo cultivada também em regiões no oeste do estado. Neste trabalho, foi utilizada uma série de doze imagens mensais de NDVI, obtidas do satélite NOAA-17, que possui baixa resolução temporal (pixel de 1Km $\times$ 1Km). A série tem início em abril de 2004 e término em março de 2005, correspondendo ao ano safra da cana-de-açúcar de 2004/2005 e foram obtidas do banco de imagens do satélite AVHRR/NOAA do Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura, da Universidade Estadual de Campinas (Cepagri/Unicamp). As imagens AVHRR/NOAA foram automaticamente processadas realizando a calibração radiométrica, correção geométrica (georreferenciamento preciso) e geração de produtos, como o NDVI. O cálculo do NDVI foi feito a partir de imagens diárias do AVHRR/NOAA utilizando as bandas 1 e 2, que correspondem ao vermelho e infravermelho próximo, respectivamente. Nesse processo, foram

excluídos os pixels com ângulo zenital solar maior que 70° e ângulos de varredura maiores que 42°. Para amenizar o efeito da atmosfera nas imagens, geraram-se Composições de Valor Máximo (MVC) mensal de NDVI.

O conjunto de dados possui 220.235 séries temporais. Cada série temporal corresponde a um pixel da região analisada, sendo que cada observação indica o valor de NDVI daquela área em um determinado mês da safra 2004/2005, perfazendo um total de 12 valores por série. O NDVI varia no intervalo [-1,0;+1,0], onde os valores mais baixos indicam áreas com vegetação esparsa e baixo vigor vegetativo, enquanto valores mais altos indicam áreas com vegetação com maior biomassa.

Para a classificação, foi utilizado um conjunto de treinamento com dados rotulados pelos especialistas no domínio. Esse conjunto é composto por 114 séries temporais classificadas em seis diferentes classes:

- (AF): Área Florestal
- (AU): Água ou Área Urbana
- (CA): Cana-de-Açucar
- (P) : Pasto
- (AG): Área Agrícola
- (PE): Cultura Perene (café, citros ou outra)

O conjunto de treino possui igualmente 19 séries temporais para cada classe. Os algoritmos de classificação semissupervisionada foram implementados e incorporados à ferramenta SatImagExplorer.

Linear Neighborhood Propagation (LNP) (WANG; ZHANG, 2008) é uma técnica de classificação semissupervisonada baseada em grafos. Ao contrário de outros métodos que adotam a mesma abordagem, a LNP não constrói o grafo completo, ou seja, que possui arestas conectando todos os pontos. A suposição feita pela LNP é que cada um dos N elementos do conjunto de dados pode ser reescrito como combinação linear de seus K vizinhos mais próximos. Dessa forma, cada nó x_i possui arestas que o ligam aos K nós mais próximos a ele, e os pesos dessas arestas w_{ij} correspondem a valores reais positivos, cuja soma totalize 1. Para o cálculo dos pesos w_{ij} do grafo, são resolvidos N problemas de programação quadrática como apresentado na Equação 1:

$$min_{w_{ij}} \sum_{j,k:x_j x_k \in N(x_i)} w_{ij} G^i_{jk} w_{ik} \tag{1}$$

dado que $\sum_i w_{ij} = 1$ e $w_{ij} \geq 0$, onde G^i_{jk} é o elemento $G^i(j,k) = (x_i..x_j)^T(x_i..x_k)$ da matriz *Gram* G_i do nó x_i. Uma vez solucionados os problemas de minimização, é criada a matriz esparsa $W(i,j) = w_{ij}$, que representa as similaridades entre os nós do grafo.

O algoritmo LNP implementa a técnica de propagação de rótulos em que, considerando que os nós possuam valores numéricos que representem sua pertinência a cada uma das c classes, cada nó do grafo iterativamente propaga seus valores de pertinência aos seus vizinhos, e estes para outros nós. Na LNP, essa propagação de rótulos por meio de vizinhanças lineares pode ser escrita na forma da iteração:

$$F^{m+1} = \alpha W F^m + (1-\alpha)Y \tag{2}$$

em que F e Y são matrizes de dimensões $n \times c$, sendo n o tamanho total do grafo e c o número de classes. O vetor F_i^m, corresponde aos valores de pertinência do nó x_i para cada classe, na iteração m. A matriz Y contém as pertinências para os dados já rotulados, ou seja, $Y_{ij} = 1$ se o exemplo i pertence à classe j, e 0 caso contrário. O parâmetro a indica a importância dada aos rótulos dos vizinhos e o complemento $(1-a)$ é a importância dada aos rótulos dos dados já classificados inicialmente. Após a convergência, a classe y_i de cada nó x_i recebe o valor $y_i = argmax_{1<j<c} F_{ij}$. A técnica LNP também permite encontrar uma solução fechada (indução) para a fórmula iterativa (transdução).

A outra técnica utilizada neste trabalho é o *Hierarchical Clustering and Local Graph Transduction* (HC-LGT) que é um método desenvolvido para classificação semissupervisionada em grandes bases de dados (WU et al., 2012). Ao contrário de outras técnicas que constróem um grafo para todo o conjunto de dados analisado, o HC-LGT divide os dados em subregiões com um método de clustering hierárquico, e constrói grafos menores, localmente para cada subregião dos dados. Além disso, o algoritmo de clustering utilizado, o ***BIRCH*** (ZHANG; RAMAKRISHNAN; LIVNY, 1996) é projetado para agrupamentos de grandes volumes de dados.

O HC-LGT se divide em dois módulos principais: o de agrupamento e o de classificação. No módulo de agupamento, o algoritmo BIRCH realiza o clustering hierárquico dos dados não rotulados construindo uma árvore, a CF-Tree. O elemento principal da CF-Tree é a Clustering Feature (CF). A Clustering Feature é uma estrutura utilizada para sumarização de dados de um cluster ou subcluster. Dados N elementos multidimensionais $\vec{X_i}$ em um cluster, a CF é definida como:

$$CF = (N, \vec{LS}, SS) \tag{3}$$

onde $\vec{LS}$ é a soma linear dos elementos ($\sum_{i=1}^{N} \vec{X_i}$) e SS é a soma quadrática dos mesmos elementos ($\sum_{i=1}^{N} \vec{X_i^2}$). A CF-Tree organiza os dados em uma estrutura de árvore, cujos nós representam clusters ou subclusters. Cada nó é representado por uma Clustering Feature e armazena as informações de cada nó filho também na forma de CFs. O algoritmo BIRCH é adequado para processamento de grandes volumes de dados, pois utiliza métricas que trabalham diretamente com as Clustering Features, ao invés de calcular a distância entre cada elemento dos dados.

Após a construção completa da CF-Tree, os dados são divididos em subregiões, cada uma contendo no máximo s elementos, valor definido pelo usuário. O módulo de classificaçãao consiste na construção de grafos locais para cada subregião de dados não rotulados obtidos após o processo de agrupamento. Para cada grafo, são computados valores de similaridade para as arestas entre os elementos de $L \cup U_p$, onde L é o conjunto de dados rotulados e U_p é a subregião analisada. Os autores adotam um framework de regularização denominado Normalized Graph Laplacian Regularizer, utilizado para propagar os valores de classe de cada nó para os vizinhos mais próximos. Usando esse regularizador, é obtida a solução fechada (indução) pela equação 4.

$$F^* = (I - \alpha S)^{-1} Y \tag{4}$$

onde S é a matriz normalizada de pesos do grafo, F e Y são matrizes de dimensões $n \times c$, sendo n o tamanho do conjunto $(L \cup U_p)$ e c o número de classes pré-definidas.

A matriz Y armazena as informações de classe dos dados rotulados inicialmente. Na propagação de rótulos, cada nó recebe a informação de classe de seus vizinhos com peso α e de Y com peso $(1 - \alpha)$. Cada elemento x_i de U_p recebe valor de classe $y_i = argmax_{1<j<c} f_i^* j$, ou seja, é associado à classe da qual recebeu mais informação durante a propagação. O processo se repete para cada subregião U_p, até que todo o conjunto não rotulado seja classificado.

3. Resultados e Discussões

O conjunto de mais de 220.000 séries extraídas das imagens do ano safra 2004/2005 foram classificadas nas 6 classes para as quais foi construído o conjunto de treinamento utilizando os dois métodos: LNP e HC-LGT. Ambos os métodos estão implementados no sistema SatImagExplorer, como mencionado anteriormente e além do resultado em arquivo texto, também apresenta os resultados em formato espacializado.

Os especialistas em agrometeorologia analisaram a classificação das áreas pela técnica LNP, ilustrada na Figura 2a), e concluíram que a técnica obteve êxito ao identificar áreas de mata fechada, água e pasto ao longo de todo o mapa. No entanto, considerando a principal cultura de interesse neste trabalho, a cana-de-açúcar, a técnica praticamente não identificou áreas de cultivo em quantidade signifitavas, como a região ao nordeste do estado, onde se concentra a maior produção estadual da cultura agrícola.

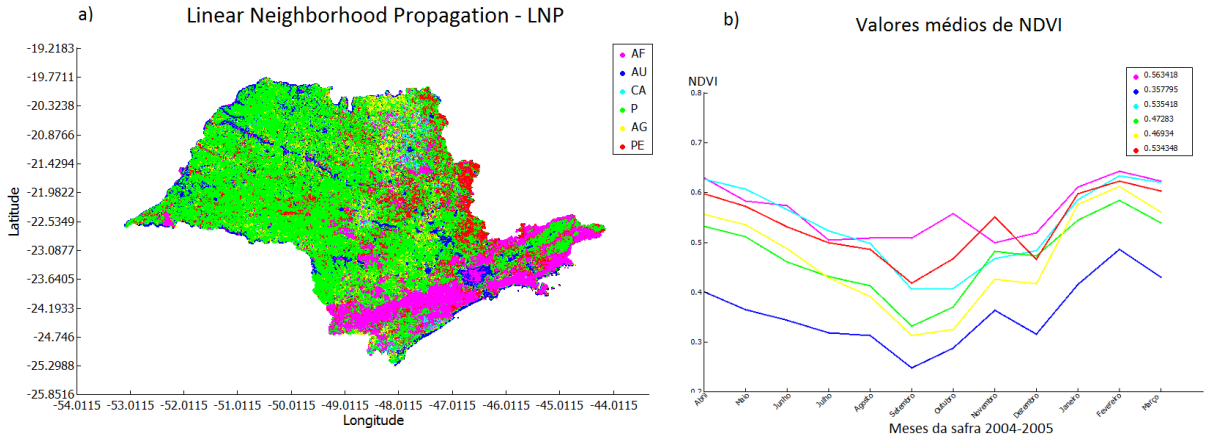

Figura 2: Classificação resultante da aplicação da técnica LNP e gráfico com perfil de NDVI para as 6 classes.

Além disso, a região centralizada no ponto de latitude $-23,55^o$ e longitude $-46,63^o$, correspondente à área central e urbana da capital São Paulo, foi classificada incorretamente como Área Florestal (AF). Como ilustrado na Figura 2b), o perfil de NDVI das áreas classificadas como AF é o mais elevado, indicando áreas de grande concentração de plantas e biomassa, o que não ocorre na região central da capital paulista.

Os resultados de classificação da HC-LGT, que são apresentados na Figura 3a), foram mais satisfatórios que os da LNP, principalmente pela identificação das áreas produtoras de cana-de-açúcar na região localizada ao nordeste do estado. O perfil de NDVI das áreas classificadas como cana-de-açúcar (Figura 3b)) apresenta valores elevados no mês de abril, grande variação entre abril e outubro, e um rápido crescimento até o fim da safra, perfil de NDVI característico da cultura de cana-de-açúcar. No entanto, apenas poucas áreas foram classificadas como cana-de-açúcar no sul do estado onde a cultura também é cultivada. Diferentemente do classificador

LNP, não houve confusão na classificação de grandes áreas urbanas, como na cidade de São Paulo.

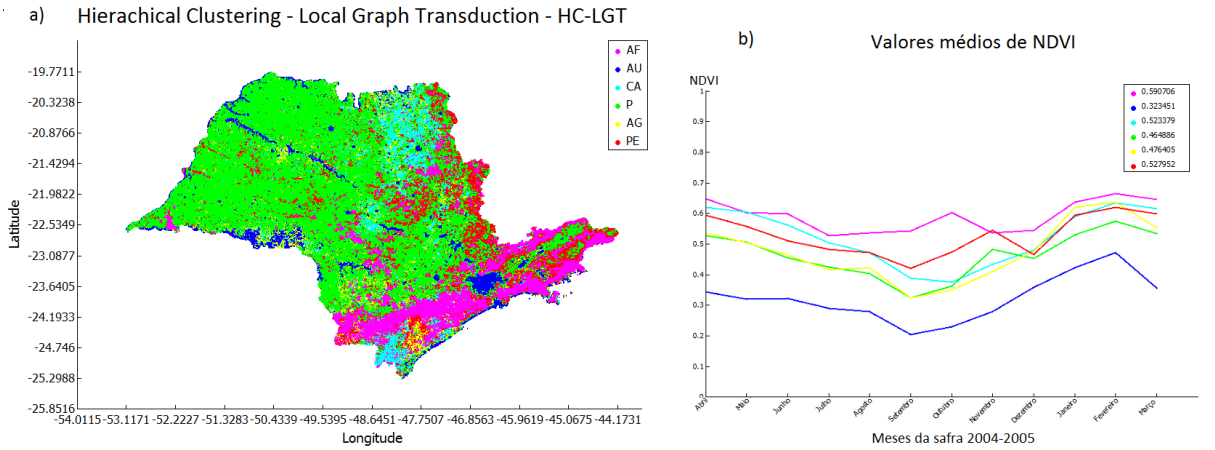

Figura 3: Classificação resultante da aplicação da técnica HC-LGT e gráfico com perfil de NDVI para as 6 classes.

Do ponto de vista computacional, embora as duas técnicas sejam baseadas em grafos, a HC-LGT apresentou um tempo de execução, aproximadamente, vinte vezes menor que a LNP. Isso se deve ao fato da técnica LNP modelar um grafo com base em todo o conjunto de dados para realizar a propagação de rótulos, enquanto a HC-LGT subdivive o conjunto total em um grande número de subregiões, por meio da análise de clustering do algoritmo BIRCH. Assim, cada grafo modelado pela técnica HC-LGT baseia-se em um subcluster de tamanho muito inferior ao do conjunto de dados original.

4. Conclusão

Este trabalho apresentou a aplicação de dois métodos de classificação HC-LGT (*Hierachical Clustering - Local Graph Transduction*) e LNP (*Linear Neighborhood Propagation*) integrados ao sistema SatImagExplorer para classificar séries temporais de NDVI extraídas do satélite AVHRR-NOAA. A principal contribuição do trabalho foi mostrar que mesmo utilizando um satélite de baixa resolução temporal foi possível classificar a região de estudo em 6 classes principais de forma automática. Além disso, uma das técnicas classificou áreas produtoras de cana-de-açúcar de forma adequada de acordo com a avaliação dos especialistas. De forma complementar, esta classificação também pode servir de base para estudos de cobertura da terra como um passo inicial à aplicação de técnicas mais complexas.

Para o especialista, a possibilidade de extrair séries temporais de dados a partir de imagens de satélite, analisá-las utilizando técnicas de mineração de dados, como de classificação, e visualizar os resultados de forma geoespacial, em um único ambiente integrado, se mostra um importante suporte em pesquisas envolvendo agricultura. Além disso, o potencial dessa análise pode ser destacado pelo uso de imagens de baixa resolução espacial, pois apesar da perda de informação na captura das imagens, as técnicas aplicadas permitem a extração de conhecimento relevante, com bons resultados, como apontam os especialistas que apoiaram a avaliação dos resultados. Como trabalhos futuros pretende-se incluir um método para validação dos resultados de forma menos subjetiva.

5. Agradecimentos

Agradecemos às agências Fapesp, CNPq, Capes e Embrapa pelo apoio financeiro e ao Cepagri/Unicamp pela base de imagens de satélite AVHRR/NOAA.

Referências

CHINO, D. Y. T.; ROMANI, L. A. S.; TRAINA, A. J. M. Construindo séries temporais de imagens de satélite para sumarização de dados climáticos e monitoramento de safras agrícolas. *Revista Eletrônica de Iniciação Científica*, v. 10, p. 1–16, 2010.

DASCHIEL, H.; DATCU, M. Information mining in remote sensing image archives: System evaluation. *IEEE Transactions on Geoscience and Remote Sensing*, v. 43, n. 1, p. 188–199, 2005.

DATCU, M. et al. Information mining in remote sensing image archives: System concepts. *IEEE Transactions on Geoscience and Remote Sensing*, v. 41, n. 12, p. 2923–2936, 2003.

GANGULY, A. R.; STEINHAEUSER, K. Data mining for climate change and impacts. In: *Proceedings of the 8th International Conference on Data Mining Workshops (ICDM'2008)*. Pisa, Italy: IEEE Computer Society, 2008. p. 385–394.

HAN, J.; KAMBER, M. *Data Mining - Concepts and Techniques*. 1st edition. ed. New York, NY, USA: Morgan Kaufmann Publishers, 2001.

LI, J.; NARAYANAN, R. M. Integrated spectral and spatial information mining in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, v. 42, n. 3, p. 673–685, 2004.

NIGAM, K.; GHANI, R. Analyzing the effectiveness and applicability of co-training. In: *Proceedings of the ninth international conference on information and knowledge management*. New York, USA: ACM Press, 2000. p. 86–93.

RATANAMAHATANA, C. A.; KEOGH, E. Making time-series classification more accurate using learned constraints. In: *Proceedings of SIAM International Conference on Data Mining*. Florida, USA: SIAM, 2004. p. 11–22.

ROMANI, L. A. S. et al. Clustering analysis applied to ndvi/noaa multitemporal images to improve the monitoring process of sugarcane crops. In: *Proceedings of the The 7th International Workshop on the Analysis of Multitemporal Remote Sensing images (Multitemp'2011)*. Trento, Italy: IEEE, 2011. p. 33–36.

ROMANI, L. A. S. et al. Monitoring sugar cane crops through dtw-based method for similarity search in ndvi time series. In: *Proceedings of the 5th International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multitemp'2009)*. Groton, Connecticut, USA: IEEE, 2009. p. 171–178.

WANG, F.; ZHANG, C. Label propagation through linear neighborhoods. *IEEE Trans. on Knowl. and Data Eng.*, v. 20, n. 1, p. 55–67, 2008.

WEI, L.; KEOGH, E. Semi-supervised time series classification. In: *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*. New York, USA: ACM Press, 2006. p. 748–753.

WU, G. et al. Hierarchical clustering and local graph transduction for large scale semi-supervised classification. *Journal of Computational Information Systems*, v. 8, p. 1165–1175, 2012.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. Birch: an efficient data clustering method for very large databases. *SIGMOD Rec.*, v. 25, p. 103–114, 1996.