Aprimorando a classificação semissupervisionada de séries temporais extraídas de imagens de satélite

B. F. do Amaral[1], R. R. V. Gonçalves[2], L. A. S. Romani[3], E. P. M. de Sousa[1]

[1] Universidade de São Paulo, Brasil {amaral, parros}@icmc.usp.br
[2] Centro de Pesquisas Meteorológicas e Climáticas Aplicadas a Agricultura, Universidade de Campinas, Brasil
renata@cpa.unicamp.br
[3] Laboratório de Novas Tecnologias, Embrapa Informática Agropecuária, Brasil
luciana.romani@embrapa.br

Abstract. A quantidade de dados gerados e armazenados tem crescido muito nos últimos anos, assim como o interesse na descoberta de conhecimento e extração de padrões em tais bases de dados. No entanto, em domínios como sensoriamento remoto, grandes volumes de dados complexos, por exemplo imagens de satélite, requerem uso de recursos computacionais para análise de dados. Nesse contexto, a classificação semissupervisionada é uma alternativa adequada para extração de conhecimento de conjuntos de dados com poucas instâncias previamente rotuladas e grandes quantidades de instâncias não rotuladas. Neste artigo, é proposta uma técnica de classificação semissupervisionada para análise de séries temporais obtidas a partir de imagens de satélites, visando a identificação de áreas de plantio de cana-de-açúcar. Os resultados dos experimentos realizados indicam que a técnica proposta alcança melhor qualidade na classificação de áreas de cana-de-açúcar quando comparada a duas outras técnicas, supervisionada e semissupervisionada, propostas na literatura.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: classificação baseada em grafos, classificação semissupervisionada, mineração de dados, séries temporais de imagens de satélite

1. INTRODUÇÃO

Nas últimas décadas, o volume de dados gerados e armazenados em bases de dados tem crescido muito, principalmente devido a avanços nas tecnologias de geração, coleta e armazenamento de dados. No entanto, a quantidade de dados armazenada, assim como a complexidade de tais dados, extrapolam a capacidade humana de análise.

Na literatura de mineração de dados, diversas técnicas têm sido propostas com o objetivo de extrair padrões e informações, em princípio escondidas, a partir de grandes quantidades de dados. Uma das tarefas mais amplamente estudadas e aplicadas em diversos contextos é a classificação. Na tarefa de classificação, a partir de um conjunto de instâncias previamente classificadas, ou rotuladas, chamado conjunto de treinamento, um algoritmo é treinado para classificar instâncias não rotuladas. Na maioria dos casos, o conjunto de treinamento é elaborado com auxílio de especialistas no domínio de aplicação, que rotulam manualmente as instâncias do conjunto.

Entretanto, em alguns domínios de aplicação, a quantidade de dados previamente rotulados e disponíveis para classificação pode ser muito reduzida. Por exemplo, em sensoriamento remoto, uma área bastante estudada é a de análise de séries temporais de imagens de satélite (SITS, do inglês

Agradecemos à FAPESP, ao CNPq, à Capes pelo apoio financeiro, à Embrapa - Campinas e ao Cepagri/Unicamp pelo fornecimento das imagens de satélite e ao Canasat/Inpe pelo fornecimento da máscara de cana-de-açúcar.

Satellite Image Time Series). SITS é um termo usado para definir uma sequência de imagens de satélite que capturam a mesma cena, em intervalos de tempo iguais. Apesar da grande quantidade de imagens de satélite disponíveis para análise, devido à alta complexidade desses dados, o processo manual de rotulação é muito custoso, em questão de tempo e esforço humano, o que resulta em pequenas quantidades de dados rotulados para classificação.

Nesse cenário, uma alternativa para classificação de SITS é a classificação semissupervisionada. Diferentemente da classificação supervisionada, a abordagem semissupervisionada considera tanto os dados previamente rotulados por especialistas, quanto grandes quantidades de dados não rotulados para construção do classificador, o que é adequado no contexto de análise de SITS.

Como um conjunto de SITS consiste em uma sequência de imagens de satélite que referenciam uma área real, é possível obter informações do comportamento de cada área das imagens ao longo do período de tempo compreendido pela SITS. Essas informações são extremamente importantes em análises de uso de terra, em que deseja identificar qual é o conteúdo de uma área real, a partir de uma sequência de imagens de satélite. Com a informação do comportamento dessa área ao longo do tempo é possível identificar áreas de rios, florestas, pastagens e culturas agrícolas, através da classificação.

Neste artigo, é proposta uma técnica de classificação semissupervisionada de séries temporais de índices de vegetação obtidas a partir de SITS, visando a identificação de áreas de plantio de cana-de-açúcar, no estado de São Paulo, Brasil. As imagens de satélite utilizadas são do tipo NDVI (*Normalized Difference Vegetation Index*) [Rouse et al. 1973], do satélite AVHRR/NOAA[1]. O NDVI é um índice que mede níveis de biomassa e vigor vegetativo de plantas e é amplamente utilizado em estudos do meio ambiente. Além disso, as imagens de satélite apresentam baixa resolução espacial, sendo cada pixel representativo de uma área de $1km^2$. Devido à baixa resolução, os dados apresentam mistura espectral, o que torna o processo de classificação desafiador. Na Seção 2, são apresentados os conceitos e trabalhos correlatos, a técnica proposta é descrita em detalhes na Seção 3, os experimentos e resultados são discutidos na Seção 4, e as considerações finais são apresentadas na Seção 5.

2. CONCEITOS E TRABALHOS RELACIONADOS

Uma das tarefas mais amplamente utilizadas na literatura de mineração de dados, a classificação pode ser separada em duas abordagens de aprendizado: supervisionada e semissupervisionada. Em ambas as abordagens, informações externas aos dados analisados, provenientes de um processo de supervisão desses dados, são utilizadas para treinamento do classificador. A principal diferença é que a classificação semissupervisionada considera também o uso de dados não supervisionados para construção e treinamento do classificador [Zhu et al. 2009]. Os principais conceitos relacionados às abordagens de classificação e trabalhos relacionados são apresentados a seguir.

2.1 Classificação supervisionada

Em classificação supervisionada, a partir de um conjunto de treinamento composto por instâncias previamente rotuladas, o objetivo é identificar a relação entre os atributos dos dados de entrada e o atributo de classe, visando obter um modelo que permita a classificação de instâncias não rotuladas [Maimon and Rokach 2005]. Dado o conjunto de treinamento $D = \{d_1, d_2, ..., d_m\}$ de classes conhecidas $C = \{c_1, c_2, ..., c_m\}$, e um conjunto de instâncias não rotuladas $D' = \{d'_1, d'_2, ..., d'_n\}$, o objetivo é mapear $\mathbf{D} \rightarrow \mathbf{C}$, a fim de predizer os valores das classes $C' = \{c'_1, c'_2, ..., c'_n\}$ [Chapelle et al. 2006].

A técnica *K-Nearest-Neighbors* (KNN) é uma das mais tradicionais e amplamente utilizadas no contexto de classificação supervisionada. Apesar de sua simplicidade de algoritmo, é considerada um dos classificadores mais competitivos no cenário de classificação de séries temporais [Xi et al. 2006].

[1] http://noaasis.noaa.gov/NOAASIS/ml/avhrr.html

Para cada instância não rotulada d'_i, são selecionadas as k instâncias rotuladas de D mais similares a ela, de acordo com uma função de distância $Dist(d'_i, d_j)$. Então, a classe majoritária, dentre aquelas dos k vizinhas de d'_i, é definida como sua classe c'_i. O processo é repetido para todas as instâncias em D'. A técnica KNN foi utilizada no estudo experimental detalhado na Seção 4.

2.2 Classificação semissupervisionada

A classificação semissupervisionada consiste no processo de aprendizagem que envolve dados rotulados e não rotulados na etapa de treinamento do classificador [Chapelle et al. 2006]. Métodos de classificação semissupervisionada são adequados para problemas nos quais há pouca disponibilidade de dados rotulados, e estudos mostram que os resultados obtidos por métodos semissupervisionados superam métodos puramente supervisionados nesse cenário [Nigam et al. 1999] [Zhou et al. 2004].

Em classificação semissupervisionada, o conjunto de treinamento $X = \{x_1, x_2, ..., x_m\}$ é composto por um subconjunto de instâncias rotuladas $X_L = \{x_1, x_2, ..., x_l\}$, com valores de classe $Y_L = \{y_1, y_2, ..., y_l\}$, e um subconjunto de instâncias não rotuladas $X_U = \{x_{(l+1)}, x_{(l+2)}, ..., x_{(l+u)}\}$, de classes desconhecidas $Y_U = \{y_{(l+1)}, y_{(l+2)}, ..., y_{(l+u)}\}$ $(u + l = m)$. O algoritmo de classificação semissupervisionada pode ser de dois tipos: transdutivo ou indutivo [Zhu et al. 2009]. No caso transdutivo, o objetivo é predizer os rótulos Y_U das instâncias não rotuladas usadas no treinamento do classificador (X_U). Métodos indutivos, por outro lado, visam treinar o classificador de modo a encontrar uma função $f(x) = y$ capaz de classificar não apenas X_U, mas também dados nunca vistos.

As principais abordagens de classificação semissupervisionada são [Zhu et al. 2009]:

—Modelos generativos: assume que existe uma mistura de distribuições identificáveis, a partir das quais os dados são gerados [Nigam et al. 1999].

—Separação por regiões de baixa densidade: assume que a fronteira de separação das classes reside em uma região de baixa densidade, considerando dados rotulados e não rotulados [Joachims 1999].

—Baseada em grafos: o conjunto de treinamento é modelado por um grafo que conecta vértices representando instâncias rotuladas e não rotuladas. O objetivo é estimar uma função f cujo valor deve ser muito próximo dos valores de classe real em vértices rotulados e deve ser suave ao longo de todo o grafo [Zhu 2005]. Após a construção do grafo, existem diversas técnicas para predição das classes de vértices não rotulados, como *Min-cut* [Blum et al. 2004], *Harmonic functions* [Zhu et al. 2003] e Propagação de rótulos [Wang and Zhang 2008].

—*Self-training*: uma instância não rotulada é incorporada ao conjunto de instância rotuladas por iteração, sendo utilizada na classificação das demais instâncias não rotuladas [Wei and Keogh 2006].

—*Co-Training*: utiliza dois classificadores para os mesmos dados, sendo que eles trabalham em subconjuntos distintos dos atributos do conjunto de dados [Blum and Mitchell 1998].

—*Cluster-then-label*: é um processo de duas etapas: (1) o conjunto de treinamento é agrupado usando uma técnica de agrupamento de dados; e (2) as informações do agrupamento são utilizados para auxiliar a tarefa de classificação [Zeng et al. 2003]

Para o estudo experimental, detalhado na Seção 4, além da técnica proposta, foi implementada a técnica de classificação semissupervisionada baseada em grafos *Linear Neighborhood Propagation* [Wang and Zhang 2008], que utiliza a ideia de vizinhança linear para modelagem do conjunto de treinamento como um grafo. Assim como a técnica proposta, detalhada na Seção 3, a LNP é uma técnica baseada em grafos e que utiliza propagação de rótulos para classificação dos vértices do grafo.

Na técnica *Linear Neighborhood Propagation*, inicialmente, é construído um grafo KNN, em que uma aresta conecta dois vértices (x_i, x_j) apenas se x_i estiver entre os k vértices mais próximos de x_j, ou vice-versa, de acordo com uma função de distância. No grafo, cada vértice corresponde a uma instância do conjunto de treinamento X. Para determinação do peso da aresta w_{ij}, assume-se que

cada vértice pode ser otimamente reconstruído usando a combinação linear de seus k vértices vizinhos. Seja $N(x_i) = \{x_{i1}, x_{i2}, ..., x_{ik}\}$ a vizinhança do vértice x_i, deseja-se minimizar o erro ϵ:

$$\epsilon = \sum_i \left\| x_i - \sum_{i_j : x_{i_j} \in N(x_i)} w_{ii_j} x_{i_j} \right\|^2 \tag{1}$$

Assim, o peso $w_{ij} \in [0; 1]$ representa a participação do vértice x_j na combinação linear para reconstrução de x_i. Tais valores podem ser calculados utilizando-se programação quadrática [Wang and Zhang 2008]. Após a construção do grafo, a técnica de propagação de rótulos é utilizada para predição dos rótulos Y_U. A propagação de rótulos consiste em um processo iterativo, em que cada vértice rotulado $x_i \in X_L$ transmite um sinal com valor de classe y_i pelas arestas que o conectam a outros vértices. Assim, cada vértice recebe e armazena tais sinais e, na iteração seguinte, os transmite a outros vértices conectados a ele. Desse modo, os rótulos representados pelos sinais transmitidos são propagados por todo o grafo. Ao final do processo, cada vértice não rotulado é classificado de acordo com o sinal de rótulo armazenado por ele em maior quantidade.

O processo de propagação de rótulos é definido por:

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

em que W é a matriz de pesos do grafo ($W(i, j) = w_{ij}$), γ a matriz de rótulos iniciais, F a matriz que armazena os rótulos propagados pelo grafo e $\alpha \in (0; 1)$ é a fração de informação recebida a partir de nós vizinhos. O valor $(1 - \alpha)$ representa a contribuição dos rótulos inciais para o cálculo de F.

Seja c o número de classes do problema, a matriz F^t de dimensões $(l + u) \times c$, armazena os sinais recebidos de cada classe. Por exemplo, f_{ij}^t corresponde à quantidade de sinal da classe j armazenado no vértice x_i, na iteração t. Ao final do processo, o vértice $x_i \in X_U$ recebe o rótulo $y_i = argmax_{\ 1 \leq j \leq c}\ f_{ij}$.

3. TÉCNICA PROPOSTA

A "**C**lassificação **S**emissupervisionada usando **P**roximidade **G**eospacial" (CSPG), proposta neste artigo, utiliza a abordagem baseada em grafos para classificação semissupervisionada e a técnica de propagação de rótulos da LNP. No entanto, é importante reassaltar que a modelagem e construção do grafo a partir do conjunto de treinamento difere da proposta da LNP, como descrito a seguir.

Inicialmente, o algoritmo recebe como entrada o conjunto de treinamento X. O objetivo da técnica CSPG é predizer os valores de classe das instâncias em X_U. Por isso, a técnica pode ser considerada transdutiva, pois visa a classificação das próprias instâncias não rotuladas usadas no treinamento e não fornece fórmula fechada para dados nunca vistos (caso indutivo).

O conjunto de treinamento é gerado a partir da extração de dados de uma sequência de imagens de satélite. Dado um par de coordenadas (latitude, longitude), uma instância $x_i \in X$ é criada e possui três atributos:

—*lat*: coordenada de latitude da área real representada pelo pixel na imagem de satélite.

—*long*: coordenada de longitude da área real representada pelo pixel na imagem de satélite.

—S: uma série temporal definida por $S = \{s_1, s_2, ..., s_n\}$, em que s_i, $i \in [1, n]$, corresponde ao valor extraído do pixel representante das coordenadas *lat* e *long* na imagem i, da sequência de n imagens de satélite.

Um grafo $G = (V, E)$ é construído tal que cada vértice corresponde a uma instância $x_i \in X$. Para criação das arestas do grafo e atribuição de pesos, são necessárias duas informações: (1) a vizinhança $N(x_i)$ de um vértice; e (2) a proximidade geoespacial entre pares de vértices $Prox(x_i, x_j)$.

Um vértice x_j pertence à vizinhança $N(x_i)$ se x_j for um dos k vizinhos mais próximos de x_i, de acordo com uma função de distância $Dist(x_i.S, x_j.S)$. Qualquer função de distância compatível com séries temporais pode ser utilizada para determinação da vizinhança dos vértices, como por exemplo, a distância Euclidiana:

$$L_2(A, B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{3}$$

em que A e B são as séries temporais de dois vértices do grafo G e n é o comprimento das séries temporais. Como a mesma sequência de imagens de satélite é usada para geração de todas instâncias de X, o valor de n será o mesmo para todas as séries temporais.

A proximidade geoespacial consiste na distância Euclidiana entre os pares de coordenadas de dois vértices x_i e x_j e é expressa pela equação:

$$Prox(x_i, x_j) = \sqrt{(x_i.lat - x_j.lat)^2 + (x_i.long - x_j.long)^2} \tag{4}$$

Assim, o peso de uma aresta w_{ij} que conecta os vértices x_i e x_j, $i \neq j$ ($w_{ii} = 0$), é definido como:

$$w_{ij} = \begin{cases} 1, & \text{se } x_j \in N(x_i) \text{ e } Prox(x_i, x_j) \leq r \\ 0.5, & \text{se } (x_j \in N(x_i) \text{ e } Prox(x_i, x_j) > r) \text{ ou } (x_j \notin N(x_i) \text{ e } Prox(x_i, x_j) \leq r) \\ 0, & \text{caso contrário} \end{cases} \tag{5}$$

em que r é o raio de alcance que determina quais vértices são conectados a um vértice central, considerando a proximidade geoespacial.

Após a construção do grafo G, a técnica CSPG, assim como a LNP, utiliza a técnica de propagação de rótulos para predição dos rótulos Y_U, de acordo com a Equação 2 apresentada na Seção 2. O resultado final é a classificação das instâncias não rotuladas X_U, de acordo com os rótulos propagados pelo grafo G. Os resultados obtidos pela técnica proposta em experimentos realizados com séries temporais reais são discutidos na Seção 4.

4. EXPERIMENTOS E RESULTADOS

Nessa seção, são descritos os experimentos iniciais realizados visando a análise dos resultados alcançados pela técnica proposta (CSPG) quando comparada às técnicas KNN e LNP. O conjunto de dados usado nos experimentos é composto por séries temporais reais, extraídas de imagens de satélites geradas pelo sensor AVHRR (*Advanced Very High Resolution Radiometer*), a bordo do satélite NOAA (*National Oceanic and Atmospheric Administration*). Os satélites AVHRR/NOAA têm sido utilizados em pesquisas do ecossistema terrestre por apresentarem cobertura global, disponibilidade de longas séries de imagens e distribuição gratuita dos dados [Kampel 2004].

As imagens apresentam resolução espacial de 1km × 1km, ou seja, cada pixel da imagem representa uma área real de 1km^2. Além disso, as imagens de satélite são organizadas em uma sequência de 12 imagens mensais, de abril/2004 a março/2005, período que compreende um ciclo completo da safra de cana-de-açúcar. Para cada área real, é gerada uma série temporal de 12 valores, na qual cada observação corresponde ao valor do pixel que referencia essa área, em uma imagem da sequência.

Nesse experimento, foram usadas imagens NDVI (*Normalized Difference Vegetation Index*) [Rouse et al. 1973]. Cada série temporal é composta por 12 valores de NDVI, compreendidos no intervalo $[-1.0; +1.0]$, e indicam o estado da vegetação na área referenciada, em cada mês da safra de cana.

A região de interesse definida para esse experimento corresponde a 174.156 áreas reais, que é o número de séries temporais de NDVI que compõem o conjunto de treinamento utilizado. A região de interesse, ilustrada em azul na Figura 1, consiste na região do estado de São Paulo, Brasil, com exceção das áreas litorâneas. O litoral não foi considerado, pois não se tem informações de áreas de plantio de cana-de-açúcar nessa região, o que impossibilita o cálculo da taxa de erro na avaliação dos resultados de classificação para essas áreas. Para obtenção do conjunto de instâncias rotuladas e posterior avaliação dos resultados, foi utilizada uma máscara do mapeamento geográfico de áreas de cana-de-açúcar na região de interesse, referente à safra de 2004/2005, fornecida pelo Canasat/Inpe[2]. A partir da máscara, do total de 174.156 instâncias, 27.675 foram identificadas como áreas de cana-de-açúcar (classe positiva) e 146.481 como áreas que não possuem cana (classe negativa). Para esse experimento, a classe positiva ou negativa das séries temporais correspondentes às áreas reais foram utilizadas apenas para criação do conjunto de dados rotulados, como descrito a seguir, e para validação dos dados.

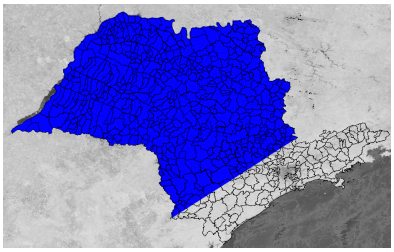

Fig. 1: Região de interesse (em azul) na imagem de satélite.

O conjunto de instâncias rotuladas foi criado com auxílio de especialistas em agrometeorologia do Cepagri/Unicamp e da Embrapa Informática Agropecuária, em duas etapas: 1) obtenção de amostras de instâncias de classe positiva e negativa a partir das áreas identificadas usando a máscara de cana-de-açúcar; 2) seleção de instâncias que representam séries temporais características de tipos de áreas conhecidos. Dentre as instâncias que não representam áreas de cana-de-açúcar foram selecionadas as que apresentam perfis de NDVI características de áreas de lagos e rios, áreas urbanas, florestas e matas, pastagem e outras culturas agrícolas. Os especialistas também selecionaram instâncias de cana-de-açúcar que representam subtipos conhecidos de culturas de cana-de-açúcar. Foram selecionadas 65 séries temporais de classe positiva e 100 de classe negativa, totalizando 165 séries temporais rotuladas X_L. Das demais 173.991 instâncias a informação de classe positiva ou negativa foram omitidas para construção de X_U (Tabela I).

Tabela I: Distribuição de instâncias de classe positiva e negativa.

	Cana (Pos)	Não Cana (Neg)	Total
Conjunto de Dados Não Rotulados (X_U)	27.610	146.381	173991
Conjunto de Dados Rotulados (X_L)	65	100	165
Total (X)	27.675	146.481	174156

As três técnicas foram parametrizadas com mesmo valor de k, e as duas técnicas com abordagem baseadas em grafos receberam o mesmo valor de α. Após a realização de diversos testes variando os valores dos parâmetros, $k = 5$ foi escolhido, pois foi o valor com o qual a técnica KNN apresentou o

[2]http://www.dsr.inpe.br/laf/canasat

melhor resultado. Utilizando o valor de k, as técnicas LNP e CSPG foram executadas com $\alpha = 0.90$, $\alpha = 0.95$ e $\alpha = 0.99$, valores comumente usados na literatura. As duas técnicas apresentaram melhores resultados com $\alpha = 0.95$. O melhor resultado obtido para a técnica CSPG foi utilizando $r = 0.02$, o que corresponde a conectar uma área da imagem de satélite às áreas vizinhas adjacentes.

Para validação dos resultados, foram utilizadas as informações de classe das instâncias de X_U, obtidas a partir da máscara de cana-de-açúcar. Os resultados são ilustrados na Tabela II.

Tabela II: Resultados de classificação obtidos pelas técnica CSPG, LNP e KNN.

	CSPG		LNP		KNN	
Classe Atribuída / Classe Real	Cana	Não Cana	Cana	Não Cana	Cana	Não Cana
Cana	**72.09**%	27.91%	67.55%	32.45%	60.77%	39.23%
Não Cana	32.36%	**67.64**%	42.26%	57.74%	43.90%	56.10%

Dentre as três técnicas, a CSPG foi a que obteve o maior número de áreas de cana e não cana corretamente classificadas (72.09% e 67.64%, respectivamente). Apesar disso, a quantidade de áreas que não possuem cana rotuladas como áreas de cana-de-açúcar ainda é elevado (32.36%). Segundo os especialistas, como as séries temporais foram extraídas a partir de imagens de baixa resolução espacial (1km^2), há muita mistura espectral e perda de informação, e por isso diversas áreas de pastagem e outras culturas agrícolas foram incorretamente classificadas como cana-de-açúcar.

Utilizando os dados mostrados na Tabela II, foi calculada uma taxa de erro E:

$$E = \frac{VP + VN}{VP + FP + FN + VN} \tag{6}$$

—VP = verdadeiro positivo (área de cana classificada como cana)
—FP = falso positivo (área de não cana classificada como cana)
—FN = falso negativo (área de cana classificada como não cana)
—VN = verdadeiro negativo (área de não cana classificada como não cana)

As taxas de erro para as técnicas CSPG, LNP e KNN são mostradas na Tabela III.

Tabela III: Taxas de erro de classificação para as técnicas CSPG, LNP e KNN.

	CSPG	LNP	KNN
E	**0.32**	0.41	0.43

Como a técnica KNN é puramente supervisionada, o classificador é construído com base apenas no conjunto de instâncias rotuladas, enquanto que as técnicas LNP e CSPG utilizam tanto instâncias rotuladas quanto não rotuladas para a classificação. Como o conjunto de treinamento é significativamente menor que o conjunto de teste (0.009%), as técnicas de classificação semissupervisionada obtiveram taxas de erros inferiores à abordagem supervisionada do KNN.

No entanto, mesmo utilizando os dados não rotulados, a técnica LNP obteve resultados inferiores aos da técnica proposta. A principal diferença entre as duas técnicas é a construção do grafo que representa o relacionamento entre as séries temporais do conjunto de dados. Enquanto a técnica LNP conecta os vértices do grafo apenas pela similaridade entre as séries temporais, a técnica CSPG também conecta as séries temporais pela proximidade geográfica das áreas reais representadas por elas. Portanto, pode-se concluir que, para o conjunto de séries temporais de NDVI analisado, o grafo construído pela técnica CSPG é capaz de modelar mais precisamente a maneira como as séries temporais de cada classe se organizam e se relacionam no conjunto de dados.

5. CONCLUSÕES

Neste artigo, foi proposta a técnica CSPG para classificação semissupervisionada de séries temporais obtidas a partir de imagens de satélite. A CSPG segue a abordagem baseada em grafos com propagação de rótulos, utilizando a proximidade geoespacial e a similaridade entre as séries temporais para construção do grafo e ponderação das arestas.

Os resultados dos experimentos realizados com séries temporais reais utilizando a CSPG são promissores e, quando comparados aos resultados obtidos com KNN e LNP, apresentam uma melhoria significativa na precisão de classificação. Vale destacar que as séries foram extraídas de imagens de baixa resolução espacial, com mistura espectral e perda de informação, o que torna o processo de classificação ainda mais impreciso.

Em trabalhos futuros, pretende-se aprimorar a técnica proposta, com foco na construção do grafo que modela os dados, visando representar mais fielmente as relações existentes entre as séries temporais analisadas, aperfeiçoando o processo de identificação de áreas produtoras de cana-de-açúcar. Além disso, a técnica desenvolvida será aplicada a outros conjuntos de dados compostos por imagens de satélite com maior resolução espacial, como do satélite MODIS[3]. Além disso, pretende-se ainda avaliar mais detalhadamente a influência dos parâmetros k, α e r no processo de classificação, assim como a relação entre eles, e realizar um extenso estudo comparativo entre a CSPG e outras técnicas de classificação semissupervisionadas propostas na literatura.

REFERÊNCIAS

Blum, A., Lafferty, J., Rwebangira, M. R., and Reddy, R. Semi-supervised learning using randomized mincuts. In *Proceedings of the 21st International Conference on Machine learning*. ACM, New York, USA, pp. 13–20, 2004.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, pp. 92–100, 1998.

Chapelle, O., Schölkopf, B., and Zien, A. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

Joachims, T. Advances in kernel methods. MIT Press, Cambridge, MA, USA, Making large-scale support vector machine learning practical, pp. 169–184, 1999.

Kampel, M. Características gerais dos satélites noaa: histórico, instrumentos e comunicação de dados. In *Aplicações ambientais brasileiras dos satélites NOAA e Tiros-N*, N. J. Ferreira (Ed.). São Paulo: Oficina de Textos, 2004.

Maimon, O. and Rokach, L. *The Data Mining and Knowledge Discovery Handbook*. Springer, Heidelberg, 2005.

Nigam, K., Mccallum, A., Thrun, S., and Mitchell, T. Text classification from labeled and unlabeled documents using em. In *Machine Learning*. pp. 103–134, 1999.

Rouse, J. W., Haas, R. H., Schell, J. A., and Deering, D. W. Monitoring vegetation systems in the great plains with erts. In *Third ERTS Symposium*, S. C. Freden and M. A. Becker (Eds.). Vol. 1. NASA, pp. 309–317, 1973.

Wang, F. and Zhang, C. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering* 20 (1): 55–67, Jan., 2008.

Wei, L. and Keogh, E. Semi-supervised time series classification. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*. KDD '06. ACM, New York, NY, USA, pp. 748–753, 2006.

Xi, X., Keogh, E., Shelton, C., Wei, L., and Ratanamahatana, C. A. Fast time series classification using numerosity reduction. In *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. ACM, New York, NY, USA, pp. 1033–1040, 2006.

Zeng, H.-J., Wang, X.-H., Chen, Z., Lu, H., and Ma, W.-Y. Cbc: Clustering based text classification requiring minimal labeled data. In *Proceedings of the Third IEEE International Conference on Data Mining*. ICDM '03. IEEE Computer Society, Washington, DC, USA, pp. 443–450, 2003.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. Learning with local and global consistency. *Advances in neural information processing systems* 16 (16): 321–328, 2004.

Zhu, X. Semi-supervised learning literature survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

Zhu, X., Ghahramani, Z., and Lafferty, J. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*. ICML '03. pp. 912–919, 2003.

Zhu, X., Goldberg, A. B., Brachman, R., and Dietterich, T. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 2009.

[3] http://modis.gsfc.nasa.gov/