Revista Brasileira de Meteorologia, **v.25**, **n.1**, 13 - 23, 2010

ESTIMATIVA DE OCORRÊNCIA DE PRECIPITAÇÃO EM ÁREAS AGRÍCOLAS UTILIZANDO FLORESTA DE CAMINHOS ÓTIMOS

GREICE MARTINS DE FREITAS[1], JOÃO PAULO PAPA[2], ANA MARIA HEUMINSKI DE AVILA[3], ALEXANDRE XAVIER FALCÃO HILTON SILVEIRA PINTO[4] , HILTON SILVEIRA PINTO[3]

[1]Universidade Estadual de Campinas (UNICAMP), Faculdade de Engenharia Elétrica, Departamento de Engenharia de Computação e Automação Industrial, Campinas, SP, Brasil
[2]Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) Faculdade de Ciências, Departamento de Computação, Bauru, SP, Brasil
[3]UNICAMP, Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura, Campinas, SP, Brasil
[4]UNICAMP, Instituto de Computação, Departamento de Sistemas de Informação, Campinas, SP, Brasil

greice@dca.fee.unicamp.br, papa@fc.unesp.br, {avila,hilton}@cpa.unicamp.br, afalcao@ic.unicamp.br

Recebido Fevereiro 2008 – Aceito Dezembro 2009

RESUMO

As condições meteorológicas são determinantes para a produção agrícola; a precipitação, em particular, pode ser citada como a mais influente por sua relação direta com o balanço hídrico. Neste sentido, modelos agrometeorológicos, os quais se baseiam nas respostas das culturas às condições meteorológicas, vêm sendo cada vez mais utilizados para a estimativa de rendimentos agrícolas. Devido às dificuldades de obtenção de dados para abastecer tais modelos, métodos de estimativa de precipitação utilizando imagens dos canais espectrais dos satélites meteorológicos têm sido empregados para esta finalidade. O presente trabalho tem por objetivo utilizar o classificador de padrões "floresta de caminhos ótimos" para correlacionar informações disponíveis no canal espectral infravermelho do satélite meteorológico GOES-12 com a refletividade obtida pelo radar do IPMET/UNESP localizado no município de Bauru, visando o desenvolvimento de um modelo para a detecção de ocorrência de precipitação. Nos experimentos foram comparados quatro algoritmos de classificação: redes neurais artificiais (ANN), k-vizinhos mais próximos (k-NN), máquinas de vetores de suporte (SVM) e floresta de caminhos ótimos (OPF). Este último obteve melhor resultado, tanto em eficiência quanto em precisão.
Palavras-chave: Classificadores Supervisionados, Floresta de Caminhos Ótimos, GOES, Estimativa de Ocorrência de Precipitação.

ABSTRACT: AGRICULTURAL AREAS PRECIPITATION OCCURRENCE ESTIMATION USING OPTIMUM PATH FOREST

Meteorological conditions are determinant for the agricultural production; in particular, rainfall may be cited as the most important because having direct relation with water balance. To estimate agricultural production, agrometeorological models based on the cultures behavior under meteorological conditions, have been used. Since it is difficult to obtain the required data to these models, rainfall estimation techniques using meteorological satellites images from spectral channels have been used. The objective of the present work is to apply the Optimum-Path Forest pattern classifier to the agrometeorological research field in order to correlate the available information from GOES-12 satellite infrared spectral channel images, to the reflectivity data obtained by the IPMet/UNESP radar located at Bauru, aiming

to develop a model for precipitation occurrence identification. In the experiments we compared four classification algorithms: Artificial Neural Networks (ANN), k-Nearest Neighbors (k-NN), Support Vector Machines (SVM) and Optimum-Path Forest (OPF). This last one shows the best results in terms of accuracy rate and running time.
KEYWORDS: Supervised Classifiers, Optimum-Path Forest, GOES, Precipitation Occurrence Identification

1. INTRODUÇÃO

A agricultura, dentre todas atividades econômicas, é, sem dúvida, a mais afetada pelas condições meteorológicas, sendo estas determinantes para o sucesso ou fracasso dos rendimentos agrícolas (Moraes, 1998). Neste sentido, modelos agrometeorológicos (Camargo, 1999; Fonseca 2007; Klering et al, 2008), abastecidos principalmente por dados de estações meteorológicas, têm obtido grande destaque na estimativa da produção agrícola com o objetivo de auxiliar na tomada de decisões como, por exemplo, planejamento do uso do solo, adaptação de culturas, monitoramento e previsão de safras, controle de pragas e doenças dentre outros.

Trabalhos baseados em estimativa de precipitação por meio de imagens de satélite, têm-se tornado uma alternativa de destaque aos modelos tradicionais pela facilidade de aquisição e abundância de dados. Entretanto, a dificuldade deste tipo de abordagem, encontra-se principalmente, em correlacionar características espectrais das imagens com dados de precipitação. Como exemplo pode-se citar o trabalho de Adler & Negri (1988), no qual demonstram que precipitações intensas sempre estão associadas com os topos frios das nuvens, mas a recíproca não é verdadeira. Desta forma, soluções encontradas através do uso de técnicas de processamento de imagens aliadas à inteligência artificial têm chamado atenção pelos bons resultados e capacidade de generalização dos dados.

Mccullagh et al. (1995) foram os primeiros a desenvolver um modelo de redes neurais artificiais (*Artificial Neural Networks* - ANN) (Haykin, 1994) para estimação de precipitação, utilizando três métodos diferentes. Os testes foram feitos com imagens da costa da Tasmânia entre os meses de agosto e setembro de 1994, obtendo uma acurácia de 77,4% utilizando imagens do canal visível e infravermelho, 75,2% combinando redes neurais com modelos de previsão e 78,6% utilizando os dois métodos combinados. Bellerby et al. (2000) utilizaram um modelo de ANN para o satélite GOES-8 e o radar TRMM (*Tropical Rainfall Measuring Mission*) com imagens da região da Amazônia. Para ativar o algoritmo, utilizaram 45 características obtidas das imagens do satélite, tais como os valores das quatro bandas, textura, mudanças ocorridas em imagens consecutivas e o horário das imagens. Como resultado, obtiveram coeficientes de correlação variando entre 0.4 e 0.5; um valor moderado. Palmeira et al. (2004) utilizaram uma abordagem probabilística de precipitação utilizando dados do canal infravermelho do satélite GOES. A presença de raios e partículas de gelo são associadas à presença de precipitação.

Umehara et al. (2005) propôs um sistema para estimação de ocorrência de chuvas baseado em Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM) (Boser, 1992). Uma SVM propõe resolver algum problema de classificação qualquer (classificar um dado como sendo precipitação ou não, por exemplo) em um espaço amostral de alta dimensão, ou seja, sua teoria é comprovada em espaços de dimensão infinita, o que nem sempre é válido na prática. Freitas et al. (2007b) e Freitas et al. (2008) propuseram também a utilização de outras variações do SVM para estimativa de ocorrência de precipitação utilizando imagens de satélite.

Na área de aprendizado de máquina, um novo método de classificação de padrões denominado Floresta de Caminhos Ótimos *(Optimum Path Forest* – OPF*)* (Papa et al., 2009) têm demonstrado superioridade em relação aos classificadores ANN e SVM, quando aplicado em bases de imagens contendo formas (Papa et al., 2007; Papa et al., 2009), texturas (Montoya et al., 2007) e identificação de disfagias em humanos (Spadotto et al., 20008). Esta técnica de classificação tem como principais vantagens a rapidez e a habilidade em trabalhar com um grande volume de dados.

O presente trabalho propõe a utilização do classificador OPF na área de agrometeorologia, visando o desenvolvimento de um modelo para a estimativa de ocorrência de precipitação utilizando informações disponíveis no canal espectral infravermelho do satélite meteorológico GOES-12, tendo como referência terrestre os dados de refletividade obtida pelo radar do IPMet/UNESP localizado no município de Bauru-SP. Foram realizados experimentos de modo a comparar o classificador OPF a classificadores supervisionados conhecidos na literatura: SVM, ANN-MLP e *k*-NN. Uma versão prévia deste trabalho pode ser encontrada em Freitas et al. (2007a). O presente estudo contempla mais um classificador na seção experimental, bem como análises mais detalhadas a respeito das taxas de acerto dos classificadores e uma fundamentação teórica mais abrangente que o anterior. O trabalho está organizado da seguinte forma: a

Seção 2 apresenta os classificadores estudados e implementados. As Seções 3 e 4 apresentam, respectivamente, os resultados e a conclusão.

2. MATERIAL E MÉTODOS

A presente seção descreve os classificadores supervisionados de padrões utilizados nos experimentos. Entendem-se, também, por problemas de classificação supervisionada, as situações nas quais se possuí os conhecimento *a priori* de todo o conjunto de dados. Tal informação pode, assim, ser utilizada para a classificação do conjunto de dados não vistos (conjunto de teste). Outras situações nas quais se tem pouca (classificação semi-supervisioanda) ou nenhuma informação sobre o conjunto de dados (classificação não supervisionada), não serão abordadas no presente trabalho.

3.1. Classificadores de Padrões

3.1.1. k-Vizinhos mais Próximos (*k*-NN)

O classificador de padrões *k*-NN (*k-Nearest Neighbor Algorithm*) (Fukunaga, 1975), é um dos algoritmos de classificação mais conhecidos, principalmente pela sua simplicidade de implementação. Cada *pixel x* da imagem é representado num espaço n-dimensional pelo seu respectivo vetor de características $\vec{x} = (x^1, \ldots, x^n)$, onde cada x^i, $1 \le i \le n$, representa uma informação relevante de cada *pixel* como, por exemplo, seu tom de cinza, média da vizinhança ou temperatura que o *pixel* representa no canal termal.

Durante o processo de treinamento do classificador, os *pixel*s pertencentes às imagens utilizadas para treinamento são atribuídos à classe ϖ_j, $1 \le j \le c$, sendo que c representa o número de classes, de acordo com conhecimento *a priori* do usuário. No caso do presente estudo, os vetores de características foram divididos em duas classes, ou seja, $c = 2$, onde $\varpi_1 = 1$ representa a presença de chuva e $\varpi_2 = -1$ a ausência de chuva, de acordo com as imagens de radar, as quais foram tomadas como verdade terrestre.

Durante a etapa de classificação, um vetor de características não rotulado $\vec{x_i}$ é atribuído à classe ϖ_j se esta for a classe de maior incidência entre os *k*-vizinhos mais próximos, segundo algum tipo de distância pré-estabelecido (euclidiana, por exemplo).

3.1.2. Redes Neurais Artificiais (ANN)

Algoritmos que utilizam técnicas de inteligência artificial, tais como redes neurais, são baseados em uma analogia feita ao sistema neuronal humano, onde neurônios dispostos em várias camadas trocam informações entre si. Em sua forma mais básica, o algoritmo ANN aprende uma função de decisão que dicotomiza dois agrupamentos de dados (precipitação ou não precipitação, por exemplo) linearmente separáveis. Neste caso, a fronteira de decisão *d(x)* é dada por:

$$\mathrm{d(x)} = \sum_{i=1}^{N} w_i \vec{x}_i + w_{n+1} \qquad (1)$$

onde os coeficientes w_i, $1 \le i \le N$, são os pesos que modificam as entradas (x_i) antes de serem somadas.

Quando *d(x) >0*, a saída da máquina é 1, indicando que o padrão *x* foi reconhecido com pertencente à classe ϖ_1. Quando *d(x) < 0*, a saída será -1, indicando que *x* pertence à classe ϖ_2. Quando *d(x) = 0*, então *x* encontra-se sobre a superfície de decisão. Logo, a fronteira de decisão é obtida igualando-se (1) a zero:

$$\mathrm{d(x)} = \sum_{i=1}^{N} w_i \vec{x}_i + w_{n+1} \qquad (2)$$

que é a equação do hiperplano no espaço $\Re^n$ de padrões.

Durante o treinamento, *w(1)* é um vetor inicial de pesos que pode ser escolhido arbitrariamente. Então, no *k*-ésimo passo iterativo, troca-se *w*(*k*) por *w*(*k+1*) = *w*(*k*) + *cy*(*k*), onde *c* é um incremento positivo de correção. Entretanto, se $x(k) \in \varpi_2$ e $w(k)\, x(k) \ge 0$, então troca-se *w(k)* por *w(k+1) = w(k) - cy(k)*. Caso contrário, *w(k+1) = w(k)*. Este algoritmo é também conhecido por *Perceptron* (Haykin, 1994).

Contudo, na maioria dos casos as amostras não são linearmente separáveis, e por este motivo, algoritmos de redes neurais multicamadas são freqüentemente utilizados. A idéia consiste, basicamente, em aninhar vários classificadores do tipo Perceptron em camadas e distribuí-las, formando uma arquitetura de rede neural. Espera-se que um número razoável de neurônios e de camadas resolva problemas separáveis lineares e não lineares.

3.1.3. Máquinas de Vetores de Suporte (SVM)

Máquinas de Vetores de Suporte (SVM - *Support Vector Machines*) são um conjunto de métodos de aprendizagem supervisionados utilizados para classificação e regressão. O classificador SVM mapeia os vetores de características das amostras dadas (*pixel*s, por exemplo), como entrada para o algoritmo em um espaço de maior dimensão, onde se supõe que elas sejam separáveis linearmente. Neste espaço de alta dimensão, um hiperplano separador das amostras será construído, juntamente com dois outros hiperplanos paralelos a este (Figura 3). O hiperplano separador ($w \cdot \vec{x} - b = 0$) é o

hiperplano que maximiza a distância entre os dois hiperplanos paralelos ($w \cdot \vec{x} - b = 1$ e $w \cdot \vec{x} - b = -1$). Quanto maior essa distância, também chamada de margem, maior será o poder de generalização do classificador.

No caso linearmente separável, o algoritmo SVM rotula os vetores de características como sendo 1 e -1, indicando presença ou ausência de chuva, respectivamente. Assim, cada vetor de características $\vec{x}_i$, $i = 1, \ldots, N$ é representado pelo conjunto $\{\vec{x}_i, y_i\}$, $y_i \in \{-1, 1\}$, $\vec{x}_i \in R^n$.

Suponha que exista um hiperplano separador no espaço *n*-dimensional, que separe os exemplos positivos dos negativos; então qualquer ponto $\vec{x}_i$ que pertença ao hiperplano satisfaz a equação $w \cdot \vec{x} - b = 0$, sendo w o vetor normal ao hiperplano separador e *b* um escalar. Sejam $w \cdot \vec{x} - b = 1$ e $w \cdot \vec{x} - b = -1$ os hiperplanos paralelos ao separador. Note que, caso os dados sejam linearmente separáveis, pode-se selecionar esses dois hiperplanos paralelos de tal modo, que nenhuma amostra insida no espaço entre eles. Então, pode-se maximizar a distância (margem) entre eles. Seja d_+ / d_- a distância do hiperplano ao exemplo positivo/negativo (quadrados/círculos na Figura 1) mais próximo, então a margem de separação do hiperplano é dada por

$$d = d_+ + d_- = \frac{2}{|w|}.$$

O algoritmo SVM busca traçar o hiperplano com maior margem de separação, isto é, maximizar d, ou seja, minimizar w. Forçando todas as amostras $\vec{x}_i, \forall i$, temos:

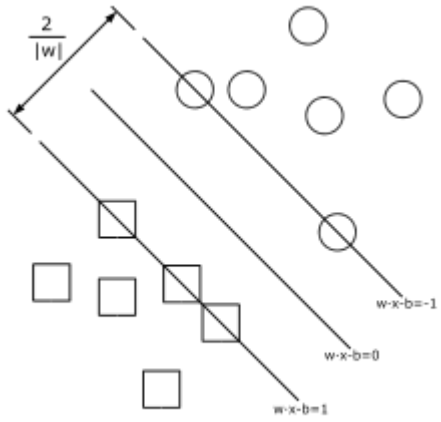

Figura 1 - Hiperplano separador e hiperplanos paralelos.

$$w \cdot \vec{x}_i - b \geq 1 \quad (3)$$

e

$$w \cdot \vec{x}_i - b \leq -1 \quad (4)$$

Assim, garante-se que todas as amostras não insidam na região pertencente à margem. Seja N o número de amostras, pode-se escrever as Equações 1 e 2 como

$$c_i \left(\vec{w} \cdot \vec{x}_i - b \right) \geq 1, \ 1 \leq i \leq N \quad (5)$$

O que nos remete ao seguinte problema de minimização:

minimizar w,b
sujeito a

$$c_i \left(w \cdot \vec{x}_i - b \right) \geq 1, \ 1 \leq i \leq N \quad (6)$$

Entretanto, este problema de otimização soluciona somente casos onde as amostras são linearmente separáveis, o que dificilmente ocorre na prática. Assim sendo, classificadores não lineares foram criados utilizando o truque do *kernel* (Vapnik, 1995). O algoritmo resultante é muito similar, exceto que todo produto interno da Equação 9 é trocado por uma função de *kernel* (φ), a qual permite mapear as amostras do seu espaço original (não linearmente separável) a um espaço de maior dimensão, onde supõe-se que elas sejam agora linearmente separáveis. A Figura 2 ilustra este procedimento.

3.1.4. Floresta de Caminhos Ótimos (OPF)

A técnica de classificação supervisionada, baseada em florestas de caminhos ótimos, modela as amostras como sendo os nós de um grafo completo. Os elementos mais representativos de cada classe do conjunto de treinamento, isto é, os protótipos, são escolhidos utilizando a abordagem da Árvore de Espalhamento Mínima (*Minimum Spanning Tree* - MST), o que garante erro zero de classificação nesta fase. Os protótipos participam de um processo de competição disputando as outras amostras oferecendo-lhes caminhos de menor custo e seus respectivos rótulos. Ao final deste processo, obtem-se um conjunto de treinamento particionado em árvores de caminhos ótimos, sendo que a união das mesmas nos remete a uma floresta de caminhos ótimos. Esta abordagem apresenta vários benefícios com relação a outros métodos de classificação de padrões supervisionados: (i) é livre de parâmetros, (ii) consegue erro zero de classificação

no conjunto de treinamento sem super treinamento dos dados, (iii) tratamento nativo de problemas multiclasses e (iv) não faz alusão sobre forma e/ou separabilidade das classes.

Seja então Z uma base de dados, e Z_1 e Z_2 os conjuntos de treinamento e teste, respectivamente, com $|Z_1|$ e $|Z_2|$ amostras, tais que $Z = Z_1 \cup Z_2$. Seja $\lambda(s)$ uma função que associa o rótulo correto i, $i = 1,\ldots,c$ da classe i a qualquer amostra $s \in Z_1 \cup Z_2$. Seja $S \in Z_1$ um conjunto de protótipos de todas as classes (isto é, amostras importantes que melhor representam as classes). A distância $d(s,t)$ entre duas amostras, s e t, é dada pela distância entre seus vetores de características $\vec{s}$ e $\vec{t}$. Pode-se utilizar qualquer métrica válida (euclidiana, por exemplo).

Nosso problema consiste em usar S, $(\vec{v},d)$ e Z_1, para projetar um classificador otimal, o qual pode predizer o rótulo correto $\lambda(s)$ de qualquer amostra $s \in Z_2$, onde $\vec{v}$ corresponde ao vetor de características de uma amostra (*pixel*) qualquer. Assim sendo, o classificador OPF cria uma partição discreta otimal, a qual é uma floresta de caminhos ótimos computada em $\Re^n$ pelo algoritmo da transformada imagem floresta (Falcão et al., 2004).

Seja (Z_1, A) um grafo completo (existe um arco entre quaisquer dois nós) cujos nós são as amostras em Z_1, onde qualquer par de amostras define um arco em A (isto é $A = Z_1 \times Z_1$) (Figura 3a). Note que os arcos não precisam ser armazenados e o grafo não precisa ser explicitamente representado.

Figura 3: (a) Grafo completo ponderado nas arestas para um determinado conjunto de treinamento. (b) MST do grafo completo. (c) Protótipos escolhidos como sendo os elementos adjacentes de classes diferentes na MST (nós circulados). (d) Floresta de caminhos ótimos resultante para a função de valor de caminho $f_{\max}$ e dois protótipos. Os identificadores (x, y) acima dos nós são, respectivamente, o custo e o rótulo dos mesmos. A seta indica o nó predecessor no caminho ótimo. (e) Uma amostra de teste (triângulo) da classe 2 e suas conexões (linhas pontilhadas) com os nós do conjunto de treinamento. (f) O caminho ótimo do protótipo mais fortemente conexo, seu rótulo 2 e o custo de classificação 0.4 são associados a amostra de teste. Note que, mesmo a mostra de teste estando mais próxima de um nó da classe 1, ela foi classificada como sendo da classe 2.

Um caminho é uma seqüência de amostras $\pi = \langle s_1, s_2, \ldots, s_k \rangle$, onde $(s_i, s_{i+1}) \in A$ para $1 \le i \le k-1$. Um caminho é denominado trivial se $\pi = \langle s_1 \rangle$. Associa-se a cada caminho π, o custo dado por uma função de custos suave f, denotada por $f(\pi)$. Define-se que um caminho π é ótimo se $f(\pi) \le f(\tau)$ para qualquer

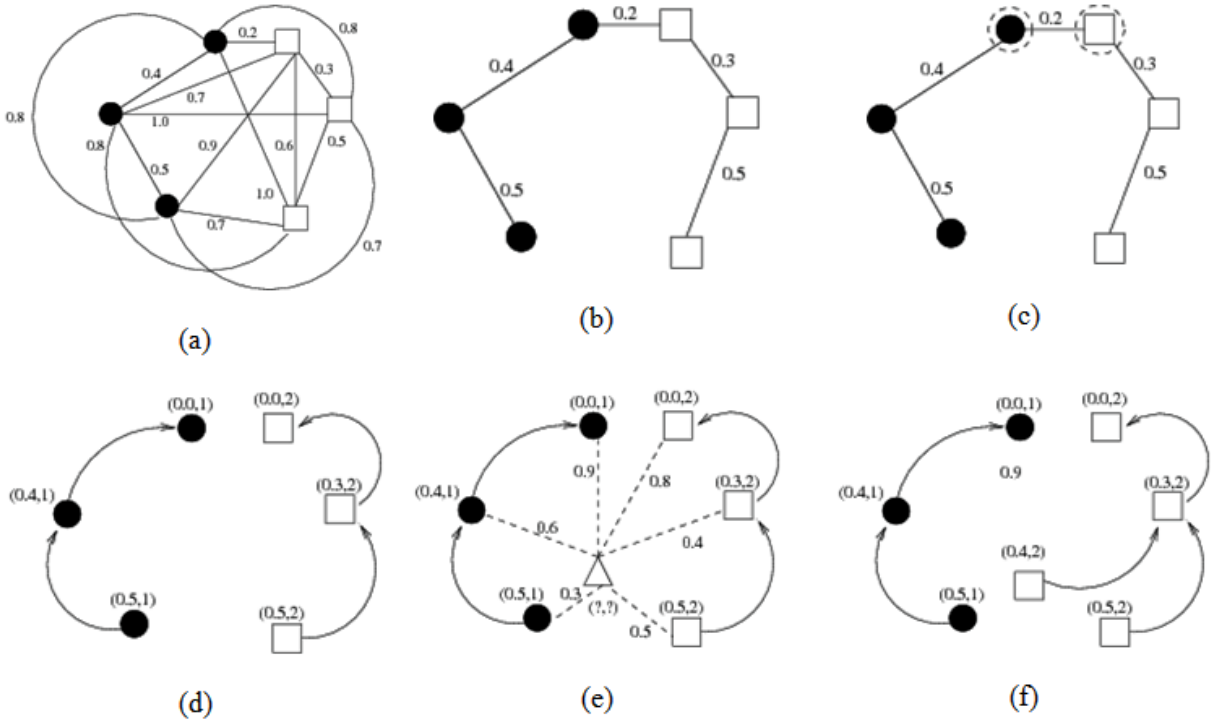

Figura 2 - Mapeamento do espaço de características originais para outro de maior dimensão.

caminho τ, onde π e τ terminam na mesma amostra S, independente de sua origem. Também se denota $\pi \cdot \langle s,t \rangle$ a concatenação do caminho π com término em s e o arco (s,t).

O algoritmo OPF pode ser utilizado com qualquer função de custo suave que pode agrupar amostras com propriedades similares (Falcão et al., 2004). Entretanto, o OPF foi projetado abordando a função de custo $f_{\max}$, por causa de suas propriedades teóricas para estimar protótipos ótimos (Allène et al. 2007):

$$f_{\max}(\langle s \rangle) = \begin{cases} 0 \; se \; s \in S \\ +\infty \; caso \; contrário \end{cases} \qquad (7)$$
$$f_{\max}(\pi \cdot \langle s,t \rangle) = \max\{f_{\max}(\pi, d(s,t))\}$$

sendo que $f_{\max}(\pi)$ computa a distância máxima entre amostras adjacentes em π, quando π não é um caminho trivial.

O algoritmo OPF associa um caminho ótimo $P^*(s)$ de S a toda amostra $s \in Z_1$, formando uma floresta de caminhos ótimos P (uma função sem ciclos, a qual associa a todo $s \in Z_1$ seu predecessor $P(s)$ em $P^*(s)$, ou uma marca nil, quando $s \in S$, como mostrado na Figura 3d. Seja $R(s) \in S$ a raiz de $P^*(s)$ a qual pode ser alcançada por $P(s)$. O algoritmo OPF computa, para cada $s \in Z_1$, o custo $V(s)$ de $P^*(s)$, o rótulo $L(s) = \lambda(R(s))$ e o seu predecessor $P(s)$ no caminho ótimo.

2.1.4.1. Treinamento

Chama-se S^* o um conjunto ótimo de protótipos quando o classificador OPF propaga os rótulos $L(s) = \lambda(s)$ para todo $s \in Z_1$. Desta forma, S^* pode ser encontrado explorando a relação teórica entre a MST (Allène et al. 2007) e a árvore de caminhos mínimos para $f_{\max}$. O treinamento consiste essencialmente em encontrar S^* e um classificador OPF enraizado em S^*.

Computando uma MST no grafo completo (Z_1, A), se obtém um grafo conexo acíclico cujos nós são todas as amostras em Z_1, e os arcos são não direcionados e ponderados (Figura 3b). Seus pesos são dados pela distância d entre os vetores de atributos de amostras adjacentes. Esta árvore de espalhamento é ótima no sentido em que a soma dos pesos de seus arcos é mínima se comparada a outras árvores de espalhamento no grafo completo. Na MST, cada par de amostras é conectado por um caminho o qual é ótimo de acordo com $f_{\max}$, ou seja, para qualquer amostra $s \in Z_1$, é possível direcionar os arcos da MST tais que o resultado será uma árvore de caminhos mínimos P utilizando $f_{\max}$ enraizada em S.

Os protótipos ótimos são os elementos conectados na MST com diferentes rótulos em Z_1, isto é, elementos mais próximos de classes diferentes (Figura 3c). Removendo-se os arcos entre classes diferentes, tais amostras adjacentes tornam-se protótipos em S^*, e o algoritmo OPF pode computar uma floresta de caminhos ótimos sem erros de classificação em Z_1 (Figura 3d). Note que uma dada classe pode ser representada por múltiplos protótipos (isto é, árvores de caminhos ótimos) e deve existir pelo menos um protótipo por classe.

3.1.4.3. Classificação

Para qualquer amostra $t \in Z_2$, são considerados todos os arcos que conectam t com as amostras $s \in Z_1$, tornando t como se fosse parte do grafo (ver Figura 3e, onde a amostra t é representada pelo triângulo no grafo). Considerando todos os possíveis caminhos entre S^* e t, deseja-se encontrar o caminho ótimo $P^*(t)$ de S^* até t com a classe $\lambda(R(t))$ de seu protótipo mais fortemente conexo $R(t) \in S^*$. Este caminho pode ser identificado incrementalmente, avaliando o valor do custo ótimo $V(t)$ como

$$V(t) = \min\{\max\{V(s), d(s,t)\}\}, \forall s \in Z_1 \qquad (8)$$

Seja $s^* \in Z_1$ o nó que satisfaz a equação acima (isto é, o predecessor $P(t)$ no caminho ótimo $P^*(t)$. Dado que $L(s^*) = \lambda(R(t))$, a classificação simplesmente associa $L(s^*)$ como a classe de t (Figura 3f). Um erro ocorre quando $L(s^*) \neq \lambda(t)$.

3.2. Experimentos

Para a realização dos experimentos foram utilizadas imagens da área localizada dentro do raio de cobertura quantitativa, de 240 km, do Radar Meteorológico do IPMet/ UNESP (Instituto de Pesquisas Meteorológicas da Universidade Estadual Paulista Júlio de Mesquita Filho), instalado no município de Bauru, região central do Estado de São Paulo, coordenadas 22°21'30'' de Latitude Sul e 49°01'38'' de Longitude Oeste. Os dados utilizados foram do tipo CAPPI (*Constant Altitude Plan Position Indicator*). Estes dados de radar, com resolução espacial de 1 km, forneceram ao estudo a referência terrestre.

Dados do canal infravermelho termal do satélite meteorológico GOES 12 (*Geoestationary Operational Environmental Satellites*), com comprimentos de onda entre 10,2 e 11,2 μm e resolução espacial de 4km, no período de dezembro/2006 a janeiro/2007, foram utilizados em horários tão próximos quanto possíveis dos horários dos dados obtidos pelo radar, com no máximo 5 minutos de defasagem. Devido às diferentes resoluções espaciais das imagens de radar e de satélite, os dados foram redimensionados, utilizando a técnica de interpolação vizinho mais próximo, para grades quadradas de mesma dimensão.

Uma dificuldade inerente à estimativa de precipitação ocorre, quando são utilizadas informações apenas do canal infravermelho. Nuvens cirrus podem levar à superestimação da área de precipitação, sendo identificadas como precipitáveis, quando na verdade não são por apresentar temperaturas mais frias, aparecendo facilmente no canal infravermelho (Rao et al., 1990). Um método desenvolvido por Adler & Negri (1988) foi utilizado para identificar este tipo de nuvem. Neste método um parâmetro S*lope* é calculado para cada *pixel*. Seja:

$$S = Tm - T, \tag{9}$$

onde T_m é a média das temperaturas dos oito *pixels* ao redor do *pixel* analisado T. O parâmetro *Slope* é calculado através da equação linear desenvolvida por Panofsky & Brier (1968):

$$Slope = 0.568(T - 217). \tag{10}$$

Esta equação linear delimita se a temperatura calculada é de uma nuvem cirrus ou de uma nuvem de precipitação.

Os modelos descritos acima foram utilizados como entrada para os algoritmos de classificação de padrões avaliados: ANN, SVM, OPF e *k*-NN. Assim sendo, cada *pixel* p_i da imagem de satélite foi representado por seu vetor $\vec{v}_i = (x^1, x^2)$, onde x^1 e x^2 são, respectivamente, a temperatura de brilho do *pixel* no canal termal e o valor de *S*. Valores de *S* menores que o *Slope,* são geralmente associados a *pixels* de nuvens cirrus, já os valores de *S* maiores ou iguais ao *Slope,* são associados como sendo prováveis *pixels* de nuvens de precipitação.

Os testes foram realizados de acordo com o seguinte procedimento: para cada conjunto de imagens (cada conjunto representa uma resolução espacial), os dados obtidos foram divididos em dois subconjuntos: treinamento e teste, com a proporção de 50% cada. Os subconjuntos de dados gerados foram os mesmos para as quatro metodologias em questão, ou seja, OPF, SVM, ANN-MLP e *k*-NN. Com o intuito de minimizar o efeito da instabilidade ocasionado pelas redes neurais, os experimentos foram repetidos 10 vezes, com diferentes conjuntos de treinamento e teste, sendo calculados a média e o desvio padrão da exatidão destes classificadores no conjunto de teste. O tempo de execução médio também foi computado.

A exatidão foi calculada de maneira a considerar classes com diferentes tamanhos, visto que o número de áreas não chuvosas é significativamente maior que a quantidade de regiões com precipitação. Assim sendo, uma alta taxa de acerto em regiões chuvosas e uma baixa exatidão na classificação de regiões sem precipitação, ou vice-versa, penaliza os classificadores, diminuindo o seu desempenho.

Seja $NZ_2(i)$, $i = 1,2,\ldots,c$, o número de amostras em Z_2 da classe *i*. Define-se, então

$$e_{i,1} = \frac{FP(i)}{|Z_2| - |NZ_2(i)|} \quad \text{e} \quad e_{i,2} = \frac{FN(i)}{|NZ_2(i)|}, \; i = 1,2,\ldots,c \tag{11}$$

onde *FP(i)* e *FN(i)* denotam os falsos positivos e falsos negativos, respectivamente. Assim, *FP(i)* corresponde ao número de amostras de outras classes que foram classificadas como sendo da classe *i* e *FN(i)* corresponde ao número de amostras da classe *i,* que foram classificadas como sendo de outras classes. Os erros $e_{i,1}$ e $e_{i,21}$ são utilizados para definir

$$E(i) = e_{i,1} + e_{i,2} \tag{12}$$

onde $E(i)$ é a soma parcial do erro da classe *i*. Finalmente, a exatidão *Acc* da classificação é dada por

$$Acc = \frac{2c - \sum_{i=1}^{c} E(i)}{2c} = 1 - \frac{\sum_{i=1}^{c} E(i)}{2c} \tag{13}$$

4. RESULTADOS E DISCUSSÃO

As Tabelas 1 e 2 mostram, respectivamente, os valores médios de acurácia e desvio padrão e os tempos médios de execução (em segundos) das etapas de treinamento e teste após 10 execuções dos algoritmos de classificação.

Nota-se a superioridade de desempenho do OPF em relação aos demais algoritmos, o qual foi cerca de 1130 vezes mais rápido que uma ANN (etapa de treinamento), com uma taxa de acerto na classificação de uma região como sendo chuvosa ou não de aproximadamente 85%. O algoritmo SVM obteve um desempenho próximo, porém foi mais lento que o OPF. Vale ressaltar que uma taxa de acerto de 85% para a estimativa de

precipitação, pode ser considerada satisfatória, pois, além de superar os resultados da literatura, ainda tem-se que considerar os erros oriundos do tratamento das imagens, como a perda de informação no processo de interpolação dos dados.

Outra maneira de analisar o desempenho dos classificadores seria através da matriz de confusão (matriz de falsos positivos - FP e falsos negativos – FN), estruturada da seguinte maneira:

No exemplo acima, as classes 1 e 2 representam, respectivamente, a presença e a ausência de chuva em um determinado pixel. Através da matriz de confusão, consegue-se determinar qual a(s) classe(s) que prejudicou (aram) o desempenho de um classificador. A Figura 5 ilustra as matrizes de confusão para os classificadores OPF, SVM, ANN e k-NN.

Pode-se, ainda, interpretar os valores de *FP* e *FN* graficamente, através do espaço ROC (*Receiver Operating Characteristic*) (Fawcett, 2006). Sejam

$$TVP = \frac{VP}{P} \tag{14}$$

e

$$TFP = \frac{FP}{P}, \tag{15}$$

tais que *TVP* e *TFP* representam, respectivamente, as taxas de verdadeiro positivo e falso positivo. A variável *P* denota

Tabela 1 - Taxa de acerto médio e desvio padrão.

CLASSIFICADOR	Taxa de acerto médio
OPF	85.06 ± 0.60
SVM	80.69 ± 15.3
ANN	85.48 ± 3.28
k-NN	81.22 ± 2.34

Tabela 2 - Tempo médio de execução em segundos para as fases de treinamento e teste.

CLASSIFICADOR	Tempo médio de execução	
	Treinamento	Teste
OPF	0.128	0.0771
SVM	3440.42	0.0399
ANN	144.68	0.0008
k-NN	1.344	0.0003

		Rótulo atribuído pelo classificador	
		1	2
Rótulo verdadeiro	1	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	2	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Figura 4 - Matriz de confusão.

o número de amostras positivas (classe 1). O espaço ROC é definido como sendo um gráfico com os eixos *x* e *y* sendo representados, respectivamente, pelas variáveis *TFP* e *TVP*, como ilustra a Figura 6. Pode-se analisar o desempenho de 4 classificadores fictícios (A, B, C e D) de acordo com a sua posição no espaço ROC.

Assim, com o intuito de permitir uma avaliação mais detalhada do que a provida pela taxa média de acerto (Tabela 1), plota-se o espaço ROC para os classificadores OPF, SVM, ANN e k-NN, como mostra a Figura 7. Nota-se que os classificadores tiveram desempenhos similares e bastante satisfatórios, o que evidencia a robustez das características extraídas para a identificação de ocorrência de precipitação em imagens de satélite. Ainda assim, o classificador OPF foi o que mais se aproximou da posição imaginária (1,0) no gráfico acima, o que indicaria o classificador perfeito no espaço ROC.

5. CONCLUSÃO

Devido à grande importância da predição de precipitação para a estimativa de rendimentos agrícolas, no presente trabalho propõem-se a comparação do desempenho de quatro algoritmos de classificação supervisionados (*k*-NN, ANN, SVM e OPF), na identificação de áreas de precipitação utilizando imagens do satélite meteorológico GOES-12 em diferentes resoluções espaciais.

O estudo foi pioneiro na utilização do classificador Floresta de Caminhos Ótimos (OPF) nesta área de pesquisa e seus resultados evidenciaram a superioridade do referido classificador, o qual obteve melhor desempenho em relação aos demais algoritmos testados, tanto quanto à taxa média de acerto, quanto ao tempo de execução, obtendo-se em média 85% de acerto nas classificações, resultado que não varia notoriamente em diferentes resoluções espaciais, demonstrando o potencial do uso do OPF em estimações e predições meteorológicas.

A vantagem do OPF com relação ao tempo de execução faz-se necessária devido ao processamento em larga escala do grande volume de dados proveniente dos satélites meteorológicos. Aplicações de sistemas críticos, tais como alertas de precipitação em centros urbanos, necessitam de sistemas com respostas rápidas e precisas. Outro ponto a ser considerado, é que o desempenho do OPF pode ser melhorado

	1	2
1	361	40
2	40	341

(a)

	1	2
1	375	26
2	62	339

(b)

	1	2
1	400	1
2	103	298

(c)

	1	2
1	375	26
2	63	338

(d)

Figura 5 -Matrizes de confusão para os classificadores (a) OPF, (b) SVM, (c) ANN e (d) k-NN.

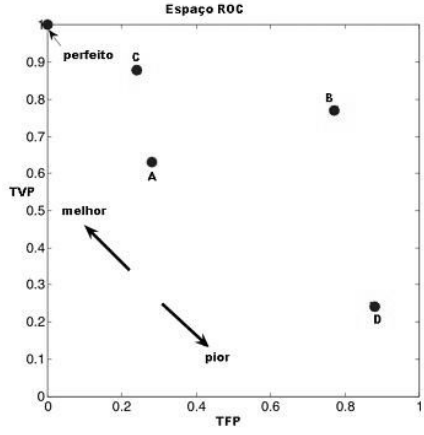

Figura 6 -Espaço ROC. Adaptado de http://en.wikipedia.org/wiki/Receiver_operating_characteristic.

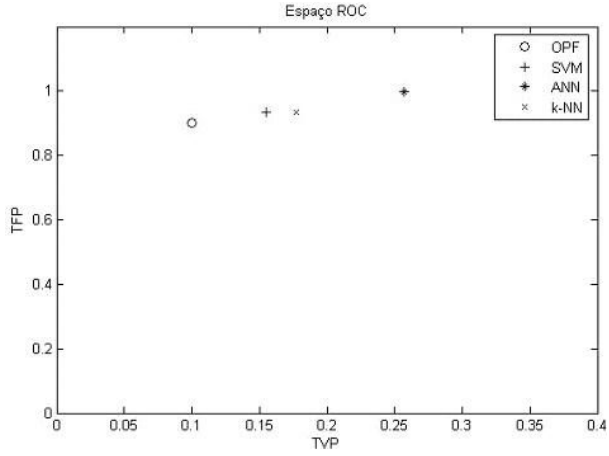

Figura 7 - Espaço ROC para os classificadores OPF, SVM, ANN e k-NN.

com o passar do tempo, visto que as imagens a serem classificadas podem ser armazenadas como parte do conjunto de treinamento e, periodicamente, o classificador pode ser re-treinado com as mesmas. Assim sendo, o OPF permite um sistema de realimentação permanente.

Trabalhos futuros devem utilizar o OPF para mensurar a quantidade de precipitação incidente em uma determinada área utilizando imagens de satélite. Serão estudadas também outras bandas que permitam a maximização da taxa de acerto do classificador, sem comprometer o seu tempo de execução.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ADLER, R. F.; NEGRI A. J. A satellite infrared technique to estimate tropical convective stratiform rainfall. **Journal of Applied Meteorology**, v. 27, p. 30-51, 1988.

ALLÈNE, C.; AUDIBERT, J.Y.; COUPRIE, M.; COUSTY, J. e KERIVEN, R. Some links between min-cuts, optimal spanning forests and watersheds. In: 8th INTERNATIONAL SYMPOSIUM ON MATHEMATICAL MORPHOLOGY, Rio de Janeiro. **Anais**. p.253-264, 2007.

ÁVILA, A. M. H. **Estimativa de Precipitação em Regiões Tropicais Utilizando Imagens do Satélite GOES 12**. 135p. Tese (Doutorado em Engenharia Agrícola). Universidade Estadual de Campinas – Faculdade de Engenharia Agrícola. Campinas, SP, 2006.

BELLERBY, T.; M. TODD; D. KNIVETON; C. KIDD. Rainfall Estimation from a Combination of TRMM Precipitation Radar and GOES Multispectral Satellite Imagery through the Use of an Artificial Neural Network. **Journal of Applied Meteorology** , v. **39**, p. 2115-2128, 2000.

BOSER, B.E.; GUYON, I.M.; VAPNIK, V.N., A training algorithm for optimal margin classifiers. In: 5th WORKSHOP ON COMPUTATIONAL LEARNING THEORY, ACM Press, p. 144-152, Pittsburgh, Pennsylvania, United States. **Anais**. 1992.

CAMARGO, M. B. P. DE A. A.; JÚNIOR, P.; ROSA, M. J.; MARCOS, S., Modelo agrometeorológico de estimativa de produtividade para o cultivar de laranja Valência. **Bragantia**, Campinas, v. 58, n. 1, 1999.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, p. 861-874, 2006.

FONSECA, E. L . DA; FORMAGGIO, A. R.; PONZONI, F. J., Estimativa da disponibilidade de forragem do bioma Campos Sulinos a partir de dados radiométricos orbitais: parametrização do submodelo espectral. **Ciência Rural**, v. 37, p. 1668-1674, 2007.

FONTANA, D.C.; WEBER, E.; DUCATI, J.R.; FIGUEIREDO, D.C.; BERGAMASCHI, H.; BERLATO, M.A. Monitoramento e previsão de safras no Brasil. [CDROM]. In: SIMPOSIO LATINO AMERICANO DE PERCEPCIÓN REMOTA. 9. Puerto Iguazú. **Anais**. 2000.

FALCÃO, A.X.; STOLFI, J. e LOTUFO, R.A. The image foresting transform: theory, algorithms, and applications. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 26, n. 1, p. 19-29, 2004.

FREITAS, G.M.; ÁVILA, A.M.H.; PAPA, J.P. Rainfall Estimation Using Transductive Learning. In: INTERNATIONAL CONGRESS ON IMAGE AND SIGNAL PROCESSING. Sanya, China. **Anais**. v. 4, p. 631-634, 2008.

FREITAS, G.M.; ÁVILA, A.M.H.; PINTO, H.S.; PAPA, J.P. Avaliação de classificadores para o monitoramento de precipitação em áreas agrícolas. In: XV CONGRESSO

BRASILEIRO DE AGROMETEOROLOGIA, Aracaju, SE. **Anais**. 2007a.

FREITAS, G.M.; ÁVILA, A.M.H.; PAPA, J.P. Semi-Supervised Support Vector Rainfall Estimation Using Satellite Images. In: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO GRÁFICA E PROCESSAMENTO DE IMAGENS, Belo Horizonte, MG. **Anais**. 2007b.

FUKUNAGA, K.; NARENDRA, P.M. A Branch and Bound Algorithms for Computing k-Nearest Neighbors. ***IEEE Transactions on Computers***. v. 24, n.7, p. 750-753, 1975.

HAYKIN, S**. Neural networks: a comprehensive foundation**. Prentice Hall PTR, 1st, 329 p., 1994.

KLERING, E. V.; FONTANA, D. C. ; BERLATO, M. A. ; CARGNELUTTI FILHO, A. Modelagem agrometeorológica do rendimento de arroz irrigado no Rio Grande do Sul. **Pesquisa Agropecuária Brasileira,** v. 43, p. 549-558, 2008.

MCCULLAGH, J.; BLUFF, K.; EBERT, E. A Neural Network Model for Rainfall Estimation. In: II NEW ZEALAND TWO-STREAM INTERNATIONAL CONFERENCE ON ARTIFICIAL NEURAL NETWORKS AND EXPERT SYSTEMS, New Zealand. **Anais**. 1995.

MELO, R.W.; FONTANA, D. C. Estimativa do rendimento de soja usando dados do modelo do ECMWF em um modelo agrometeorológico-espectral no estado do Rio Grande do Sul. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 13. Florianópolis. **Anais**. Florianópolis: INPE, 2007, p. 279-286, 2007.

MONTOYA-ZEGARRA, J.A.; PAPA, J.P.; LEITE, N.J.; TORRES, R.S.; FALCÃO A.X. Rotation-invariant Texture Recognition. In: 3RD INTERNATIONAL SYMPOSIUM ON VISUAL COMPUTING. Nevada, USA. **Anais**. Springer , Part II, LNCS 4842, p.193-204, 2007.

MORAES, A. V. DE C.; CAMARGO, M. B. P. DE; MASCARENHAS, H. A. A.; MIRANDA, M. A. C. DE; PEREIRA, J. C. V. N. A., Teste e análise de modelos agrometeorológicos de estimativa de produtividade para a cultura da soja na região de Ribeirão Preto. **Bragantia**, Campinas, v. 57, n. 2, 1998 .

PALMEIRA, F. L. B.; MORALES, C. A.; FRANÇA, G. B.; LANDAU, L. Rainfall Estimation Using Satellite Data For Paraíba do Sul Basin - Brazil. In: XXTH INTERNATIONAL SOCIETY FOR PHOTOGRAMMETRY AND REMOTE SENSING, p. 1-8, Istambul. **Anais**. 2004.

PANOFSKY, H.A.; BRIER, G.W. **Some Applications of Statistics to Meteorology**. The University of Pennsylvania, University of Park, PA, 224 p., 1968.

PAPA, J. P.; FALCÃO, A.X.; MIRANDA, P. A.V.; SUZUKI, C.T.N.; MASCARENHAS, N.D.A. Design of Robust Pattern Classifiers based on Optimum-path forests. In: MATHEMATICAL MORPHOLOGY AND ITS APPLICATIONS TO SIGNAL AND IMAGE PROCESSING, Rio de Janeiro. **Anais**. p.337-348, 2007. PAPA, J. P.; FALCÃO, A.X.; SUZUKI, C.T.N.; Supervised pattern classification based on optimum-path. **International Journal of Imaging Systems and Technology**. v. 19, n. 2, p. 120-131, 2009.

RAO, P. K.; HOLMES S. J.; ANDERSON R. K.; WINSTON J. S.; LEHR P. E. Weather Satellites: Systems, data, and environmental applications. **Bulletin of the American Meteorological Society**, Boston, 1990.

SPADOTTO, A.A.; PEREIRA, J.C.; GUIDO, R.C.; PAPA, J.P.; FALCÃO, A.X.; GATTO, A.R.; COLA, P.C.; SHELP, A.O. Oropharyngeal Dysphagia Identification Using Wavelets and Optimum Path Forest. In: 3TH IEEE INTERNATIONAL SYMPOSIUM ON COMMUNICATIONS, CONTROL AND SIGNAL PROCESSING, Malta. **Anais**. p. 735-740, 2008.

UMEHARA, S.; YAMAZAKI, T.; SUGAI, Y. *A Precipitation Estimation System Based on Support Vector Machine and Neural Network*. In: ELECTRONICS AND COMMUNICATIONS IN JAPAN, PART 3: FUNDAMENTAL ELECTRONIC SCIENCE. **Anais**. v. 89; n. 3, p. 38-47, 2006.

VAPNIK, V. **The Nature of Statistical Learning Theory**. Springer-Verlag, 1995.

VICENTE, G. A.; SCOFIELD, R. A.; MENZEL W. P. The Operational GOES Infrared Rainfall Estimation Technique. **Bulletin of the American Meteorological Society**, v. 79, n.9, p.1883-1898, 1998.