COMPARAÇÃO DOS ATRIBUTOS ESCOLHIDOS PELO TREINAMENTO DE CLASSIFICADORES DE ÁRVORES DE DECISÃO COM SELEÇÃO DE ATRIBUTOS POR FILTRO

Comparison between the features selected by decision tree classifiers' training and feature selection filters

Maurício Carvalho Mathias de Paulo[1]
Stefano Bacciuylis Bluyus Rodrigues Pansardis Mathias[1]
Marielcio Lacerda[1]
Thales Sehn Korting[1]
Leila Maria Garcia Fonseca[1]

[1]Instituto Nacional de Pesquisas Espaciais - INPE
Departamento de Processamento de Imagens
Caixa Postal 515 -- 12245-970 -- São José dos Campos - SP, Brasil
mauricio@dpi.inpe.br

RESUMO

Os resultados obtidos através da classificação de imagens de sensoriamento remoto orientada a objetos dependem de todas as etapas do processo. Após a criação ou definição dos polígonos que representam cada objeto vários atributos podem ser extraídos do comportamento dos píxels que eles contém. O objetivo deste trabalho é apresentar técnicas para avaliar os atributos relevantes para utilização durante o processo de classificação supervisionada. Estas técnicas foram disponibilizadas no aplicativo de sistemas de informação geográficas TerraView através da extensão GeoDMA. Para ilustrar as capacidades de cada método, foram comparados os atributos escolhidos por diferentes técnicas de poda das árvores de decisão e de seleção de atributos por filtro, mostrando as diferenças nos conceitos e nos resultados.

Palavras chaves: Classificação de imagens orientada a objetos, árvore de decisão, seleção de atributos.

ABSTRACT

The results obtained through the object based image classification of remote sensing imagery rely on every process' step. After creating or defining the polygons that represent each object many attributes can be measured from the contained pixels' behavior. This paper's goal is to present methods to evaluate the relevant features to be used during the supervised classification process. These techniques were made available in the geographic information system TerraView through the GeoDMA extension. In order to illustrate each method's capabilities , the feature selected through some decision tree pruning and filter feature selection techniques, showing the difference in the results and concepts.

Keywords: Object based image classification, decision tree, feature selection.

1. INTRODUÇÃO

O *GeoDMA* é uma extensão do aplicativo de sistema de informações geográficas *TerraView* , voltada para a mineração de dados geoespaciais. Com este pacote é possível realizar as principais fases de processamento necessárias para manipular dados de sensoriamento remoto (KORTING *et al.*, 2009).

Um dos procedimentos permitidos pelo aplicativo é a classificação de imagens orientada a objetos utilizando técnicas de mineração de dados. A Figura 1.1 ilustra as principais etapas do processo.

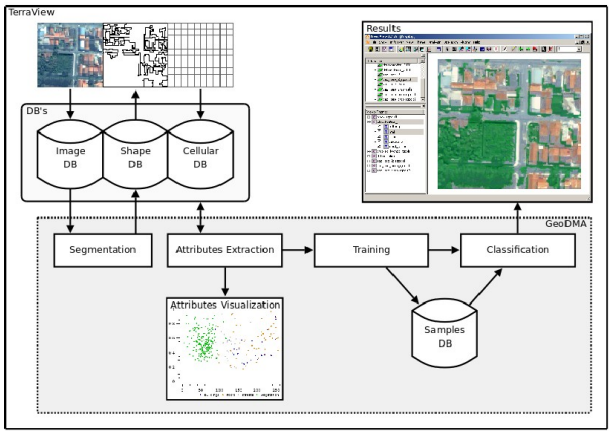

Figura 1.1 – Processamento no *GeoDMA* (KORTING *et al.*, 2010)

O processo tem início com a segmentação de uma imagem, que produz representações vetoriais (objetos) do agrupamento dos píxels com respostas semelhantes. A qualidade do agrupamento depende da técnica de segmentação, dos parâmetros definidos pelo usuário e das características das bandas disponíveis na imagem.

Para cada objeto, é possível produzir medidas que representam as características espectrais e espaciais do agrupamento de píxels. Esta etapa é chamada de extração de atributos.

O *GeoDMA* possui algoritmos de classificação supervisionada (por árvores de decisão e redes neurais) e não supervisionada (mapas auto-organizáveis) (KORTING *et al.*, 2009). Nos algoritmos supervisionados é necessária a escolha de objetos para compor as amostras de treinamento. Cada classe deve possuir um número mínimo de objetos no conjunto de treinamento que possibilite descrever o seu comportamento.

Como é impraticável calcular exatamente quantos exemplos são necessários para descrever corretamente o comportamento da classe utilizando os atributos extraídos, podem ser utilizadas técnicas que avaliem a separabilidade das classes utilizando subconjuntos de atributos. Desta forma o analista pode extrair todos os atributos disponíveis no *GeoDMA* e avaliar quais são os mais indicados para a solução de seu problema. Este processo é conhecido como seleção de atributos.

A presença de uma quantidade excessiva de atributos pode prejudicar a qualidade dos resultados da classificação. O algoritmo de classificação baseado em redes neurais (HAYKIN, 2009) contido no *GeoDMA* é susceptível a este fenômeno, portanto a escolha dos melhores atributos para separar as classes de treinamento é fundamental.

Por outro lado, o algoritmo de classificação baseado em árvore de decisão escolhe, durante o processo de treinamento, os melhores atributos para separar as classes utilizando o critério da razão de ganho baseada em entropia (QUINLAN, 1993).

A árvore é produzida escolhendo as melhores decisões que separam as amostras de treinamento das classes. Com efeito, as decisões escolhidas espelham características das amostras escolhidas, podendo não ser as melhores decisões para classificar os demais objetos da imagem. Para minimizar este problema podem ser utilizados algoritmos de poda, que verificam durante o processo de treinamento, e após o seu término, se todas as decisões estão contribuindo para o aumento da qualidade de classificação conforme a heurística de cada um.

Este trabalho apresenta uma comparação entre os atributos escolhidos utilizando técnicas de seleção de atributos por filtro e os atributos escolhidos durante o treinamento de árvores de decisão. Como a árvore exige parâmetros para controlar o processo de treinamento, são apresentados resultados com diferentes tipos de poda.

2. SELEÇÃO DE ATRIBUTOS

Com a extensa quantidade de atributos possíveis de se extrair com o *GeoDMA*, eliminar atributos desnecessários pode melhorar a qualidade da classificação, o tempo de processamento e a memória utilizada.

O algoritmo de classificação por árvore de decisão efetua uma seleção de atributos robusta intrínseca ao processo de classificação. Portanto os operadores não precisam se preocupar com quais atributos estão sendo repassados ao classificador.

Os classificadores por redes neurais, por outro lado, podem se beneficiar de uma seleção de atributos. Este processo pode ser feito imediatamente antes do treinamento do classificador, para avaliar a separabilidade das classes usando subconjuntos de atributos.

Existem métricas para avaliar a separabilidade de um conjunto de amostras de treinamento. Foram implementadas no *GeoDMA* os seguintes critérios de separabilidade: distâncias Euclideana, distância de Mahalanobis e J3.

As funções objetivo dos filtros Euclidiano e Malahanobis medem as distâncias entre cada par de classes, verificando a menor dado um subconjunto de atributos, conforme formalizado na Equação 1. Desta forma a técnica busca maximizar a menor separabilidade entre as classes (e) (THEODORIDIS; KOUTROUMBAS, 2006).

(1)

A distância Euclideana é medida entre os vetores de atributos médios (e)de cada classe, conforme apresentado na Equação 2.

(2)

A distância de Mahalanobis leva em consideração a correlação entre os atributos, através da matriz de covariância (e), conforme a Equação 3. Tanto a distância Euclideana quanto a de Mahalanobis assumem a distribuição de probabilidade normal para o comportamento dos atributos.

(3)

O algoritmo J3 utiliza o conceito de matrizes de espalhamento interno de cada classe (S_w) e de espalhamento global (S_m).

O espalhamento S_w trata-se da média das matrizes de covariância de cada classe ponderada pela quantidade de amostras. O espalhamento S_m é a matriz de covariância de todo o conjunto, sem distinguir classes. O filtro J3 avalia a separabilidade entre as classes utilizando a função J apresentada na Equação 4.

(4)

Para descobrir quais são os melhores subconjuntos de atributos utilizando as métricas apresentadas, é necessária uma técnica de busca. A busca por força bruta pode ser inviável pois a quantidade de possíveis subconjuntos é 2^n e estão disponíveis para cada banda mais de 20 atributos no *GeoDMA*. As técnicas de busca, portanto procuram encontrar o maior valor de separabilidade possível, partindo de um heurística.

A técnica de busca sequencial para frente consiste em partir de conjuntos com um atributo e a cada iteração adicionar o atributo que atinge o maior valor de separabilidade. Uma variante deste conceito é a busca sequencial flutuante para frente, que ao fim de cada iteração, procura se existe algum atributo que pode ser removido do conjunto dos escolhidos que venha a aumentar a separabilidade. Isto resolve uma limitação da busca sequencial que prioriza atributos que são escolhidos primeiro (THEODORIDIS; KOUTROUMBAS, 2006).

3. CLASSIFICAÇÃO POR ÁRVORE DE DECISÃO

3.1 O algoritmo de treinamento

As árvores de decisão são modelos estatísticos que utilizam um conjunto de treinamento para classificar e/ou "prever" algum dado, a "função" de aprendizado é a própria árvore resultante.

Em resumo uma árvore de decisão também pode ser representada como um conjunto de regras "se-senão" para assim melhorar a percepção humana.

Uma árvore de decisão é composta por nós (podendo ser representados por círculos) interconectados por ramos (representados por linhas) e por folhas (representadas por quadrados), cada nó possui um teste em um determinado atributo e cada ramo descendente corresponde a um possível valor deste atributo. As folhas estão relacionadas sempre com uma classe. (João Gama, 2002).

Existem diversos algoritmos clássicos de arvores de decisão tais como o ID3, CART e C 4.5. Este último é a evolução do ID3 e encontra-se disponível no *GeoDMA*.

As árvores de decisão depois de construídas podem possuir muitas arestas que reflitam em erros ou ruídos, isto é um problema conhecido chamado "sobreajuste". Sobreajuste significa que a árvore adquiriu um conhecimento muito especifico do conjunto de treinamento, perdendo a capacidade de generalização.

Para minimizar este tipo de problema é necessário "podar" a árvore, as podas podem ser realizadas durante a construção da árvore (pré-poda) ou após a construção da árvore (pós-poda).

3.2 Pré-poda

A pré-poda consiste em interromper o crescimento da árvore quando um critério for atingido. O critério de pré-poda objetiva modelar se haverá ganho em realizar a divisão. Havendo ganho a divisão é realizada, caso contrário é gerada uma folha com a classe de maior número de exemplos de treinamento.

Por funcionar como um teste preliminar durante a construção da árvore, a pré-poda é mais rápida, que a pós-poda. Apesar disso, pode ser menos eficiente, pois pode ocorrer a interrupção do crescimento da árvore antes do ideal, selecionando uma árvore sub-ótima.

Existem diversos critérios para interromper o crescimento. Foram implementadas no *GeoDMA* a contagem de mínimo de amostras de treinamento em cada nó e o percentual de itens não pertencentes à classe majoritária no nó.

A contagem de mínimo de amostras de treinamento em cada nó permite que não sejam criadas decisões com quantidades insignificantes de amostras. Para isso o parâmetro do número mínimo de amostras deve ser introduzido pelo analista.

O percentual de itens não pertencentes à classe majoritária insere uma tolerância a erros de amostragem. Caso um exemplo de treinamento seja escolhido errado, ele não será considerado uma decisão se ele representar menos do que a porcentagem inserida pelo usuário.

3.3 Pós-poda

Utilizando a pós-poda, a árvore cresce e atinge seu tamanho máximo e somente após este crescimento a mesma é podada, objetivando escolher a melhor sub-árvore. Com isso este processo pode ser computacionalmente ineficiente, pois a árvore cresce até o seu limite e depois é reduzida, indicando que o uso da pós-poda deve coexistir com o uso da pré-poda. Foram disponibilizadas no GeoDMA duas técnicas para a realização da pós-poda: EBP e REP.

O EBP (*Error Based Pruning*) utiliza os conceitos do PEP (*Pessimistic Error Pruning*) aplicando um erro tolerável à comparação com as subárvores. O PEP utiliza o próprio conjunto de treinamento para realizar a poda da arvore, esta técnica realiza uma varredura do tipo "Top-Down" (Tso, 2009).

Para cada folha é calculado o Erro Pep do nó t conforme a Equação 5, na qual N(t) é o número de amostras no nó e $n_j(t)$ é o número de amostras da classe majoritária no nó.

$$Error_{PEP}(t) = \frac{N(t) - n_j(t) + 0.5}{N(t)} \quad (5)$$

O Erro Pep da subárvore é obtido através da somatória dos erros da árvore:

$$Error_{PEP}(S) = \frac{\sum_{i \in \text{leaf node of } S} \left(N(i) - n_j(i) + 0.5\right)}{\sum_{i \in \text{leaf node of } S} N(i)}. \quad (6)$$

$$std\_error(S) = \sqrt{\frac{Error_{PEP}(S) \times \left(1 - Error_{PEP}(S)\right)}{N(t)}}. \quad (7)$$

No EBP são comparados o erro PEP de t for e o Erro PEP de s mais o desvio padrão (acerto realizado por BREIMAN em 1993 chamado de "EBP"), conforme a Equação 8. Caso o critério seja atendido a poda é realizada.

(8)

Já o REP (*Reduced error pruning*) divide o conjunto de amostras em treinamento e validação. As amostras de treinamento são utilizadas para construir a árvore de decisão. As amostras de validação são utilizadas para verificar os erros de classificação cometidos ao utilizar sub-árvores da árvore gerada. Por dividir o conjunto em dois, faz-se necessário uma quantidade maior de amostras.

O REP é um procedimento "*botom-up*", pois primeiro são classificadas as amostras de validação e depois são verificados os erros encontrados em cada decisão. A Equação 9 apresenta o teste feito. Se o erro de classificação dos dados de treinamento nos nós inferiores (filhos) for maior que o erro de classificação no nó atual (pai), poda-se a decisão, transformando-a em uma folha.

(9)

4. MATERIAIS E MÉTODOS

Para ilustrar os efeitos positivos que as técnicas apresentadas podem causar no processo de classificação, foi feito um estudo de caso que envolve o processo de avaliação dos atributos e de classificação de classes representadas em uma imagem de satélite

4.1 Materiais

Durante os testes, foi utilizada uma imagem Spot 5 *SuperMode* com resolução espacial 2.5m com bandas nas faixas do vermelho, infravermelho e verde. A Fig. 1 apresenta uma visão geral da imagem utilizada.

Fig. 1 – Imagem SPOT *SuperMode* canais RGB: vermelho, infravermelho e verde.

4.2 Extração de atributos

Utilizando as funções previamente existentes no GeoDMA, a imagem foi segmentada, utilizando o algoritmo de crescimento de regiões com os parâmetros: área mínima 50 e distância euclideana 20. Foram extraídos atributos geométricos e espectrais dos polígonos gerados e foi efetuada a normalização deste utilizando *z-Score*. Foram coletadas amostras de treinamento para as classes pasto, cultura, mata nativa, reflorestamento e urbanização.

4.3 Árvores de decisão

Foram geradas três árvores de decisão para classificar o mesmo conjunto de treinamento criado utilizando a segmentação da imagem Spot 5. Uma sem pós-poda, uma utilizando a poda baseada em erros (*Error Based Pruning -EBP)* e outra utilizando a poda de erro reduzido (*Reduced Error Pruning - REP*). A Tab. 1 descreve quantitativamente as árvores geradas com cada tipo de pós-poda.

TABELA 1 – RESUMO DOS EFEITOS DE PODAS

Método de Poda	Quantidade de nós	Quantidade de atributos	Altura
Sem poda	20	7	4
EBP	18	7	4
REP	8	3	3

A árvore gerada utilizando *REP* descartou o atributo largura, que foi utilizado tanto na árvore gerada utilizando *EBP* quanto na gerada sem pós-poda. Este atributo foi utilizado para separar as classes Reflorestamento e Mata Nativa. Não há indícios na literatura que este atributo seja realmente uma forma de diferenciar estas classes e é possível encontrar contra-exemplos para esta regra na própria imagem.

Durante o processamento do REP, a classificação dos dados de validação teve maior acurácia ao não utilizar este atributo, caracterizando o sobreajustamento. Este é um exemplo do comportamento do algoritmo de treinamento frente à uma característica das amostras de treinamento que não reflete em uma característica da população.

Por sua característica pessimista, a árvore gerada utilizando *EBP* possibilitou a redução de um nó que não aumenta significativamente a acurácia da árvore. A moda da banda 3 foi utilizada em duas decisões consecutivas para separar duas classes.

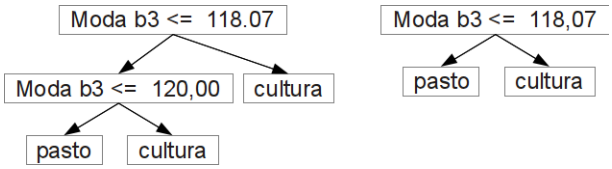

Fig. 2 – Decisão removida pela técnica *EBP*.

4.4 Seleção de atributos

Para avaliar os resultados da seleção de atributos por filtro, foram comparados os atributos escolhidos utilizando os filtros de Mahalnobis, Euclidiano e J3, com os utilizados nas árvores de decisão. A Tabela 2 apresenta os resultados encontrados. Na coluna "Qtd" estão representadas as quantidades de atributos escolhidos. Os atributos de 0 a 7 são geométricos, de 9 a 24 são espectrais.

TABELA 2 – ATRIBUTOS ESCOLHIDOS

Técnica	Qtd	Atributos
Árvore - Sem poda	7	7, 9, 15, 16, 17, 21, 22
Árvore - *EBP*	7	7, 9, 15, 16, 17, 21, 22
Árvore - *REP*	3	15, 20, 22
Filtro Euclideano	9	1, 2, 4, 5, 7, 12, 14, 16, 18
Filtro Mahalanobis	16	3, 6, 7, 8, 9, 10, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23
Filtro J3	19	0, 2, 3, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 18, 19, 20, 21, 22, 23

Devido às características dos filtros, que modelam os atributos seguindo distribuições de probabilidade normais unimodais, diversos atributos geométricos foram considerados importantes. Como a árvore de decisão não assume hipóteses sobre o comportamento estatístico dos atributos, esta técnica descartou a maior parte dos atributos geométricos.

5. CONCLUSÃO

Conforme apresentado na Tab. 1, os algoritmos de poda da árvore de decisão influenciam diretamente as decisões geradas na árvore e, consequentemente, a qualidade da classificação produzida por cada uma.

Teoricamente, a poda de erro reduzido(*REP*) busca resolver o problema de sobreajustamento, minimizando características existentes apenas no conjunto de treinamento. Para que isso seja possível, esta técnica exige mais amostras que as demais, permitindo assim, dividir em dados de treinamento e validação.

A técnica de poda baseada em erro (*EBP*) pode resolver problemas pequenos de sobreajustamento, mas, por não possuir um conjunto de testes, não atinge os mesmos resultados que o *REP*. Por outro lado, esta técnica não divide as amostras de treinamento, portanto pode ser aplicada em situações onde não é possível coletar uma quantidade significativa de amostras.

Os resultados das escolhas de atributos indicam que os critérios da árvore de decisão podem não ser úteis para outros classificadores. Nas árvores a escolha dos atributos é feita até atingir critérios de parada do treinamento. Por este motivo alguns atributos importantes podem não ter sido escolhidos por limitações das amostras de treinamento ou dos parâmetros inseridos.

A seleção de atributos pode ser um critério preliminar para guiar os analistas antes de iniciar a classificação por redes neurais. Por haver discordância entre os métodos, há indícios de que testar os três métodos e avaliar quais atributos foram escolhidos por cada técnica pode ser um procedimento para avaliar a relevância de cada atributo.

REFERÊNCIAS BIBLIOGRÁFICAS

HAYKIN, S. S. **Neural networks and learning machines**. [S.l.]: Prentice Hall, 2009. v. 3

KORTING, T. S.; FONSECA, L. M. G.; ESCADA, M. I. S.; CÂMARA, G. GeoDMA-Um sistema para mineração de dados de sensoriamento remoto. **Simpósio Brasileiro de Sensoriamento Remoto. Natal, RN, Brasil**, 2009.

KORTING, T. S.; FONSECA, L. M. G.; CAMARA, G. **Interpreting images with GeoDMA**. GEOBIA. **Anais**... [S.l.]: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. , 2010

QUINLAN, J. R. **C4. 5: programs for machine learning**. [S.l.]: Morgan kaufmann, 1993.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. [S.l.]: Elsevier/Academic Press, 2006.

TSO, B. **Classification Methods for Remotely Sensed Data**. 2ed., 207p. CRC Press, 2009.

MITCHELLI, T. **Machine Learning.** The McGraw-Hill Companies, 1997.

GAMA J. **Árvores de decisão**. Disponível em <http://www.liaad.up.pt/~jgama/Aulas_ECD/arv.pdf>. Acessado em 14 de junho de 2011.