Anais do 6º Congresso Brasileiro de Agroinformática – SBIAgro 2007, São Pedro, Brasil, 8-11 outubro 2007,
Embrapa Informática Agropecuária, p.161-165.

AVALIAÇÃO DE MÉTODOS DE SELEÇÃO DE ATRIBUTOS PARA CLASSIFICAÇÃO DE SOLOS

FERNANDO ATTIQUE MÁXIMO[1]
STANLEY ROBSON DE MEDEIROS OLIVEIRA[2]
MARIA LEONOR LOPES-ASSAD[3]

RESUMO: Técnicas de mineração de dados têm sido usadas, estrategicamente, para transformar dados em informações e conhecimentos visando subsidiar o processo decisório em vários domínios. Na agricultura, em particular, essas técnicas são eficientes para selecionar o conjunto de atributos relevantes para a criação de um modelo de classificação de dados em bancos com muitas variáveis. Este trabalho tem por objetivo avaliar a eficiência de diferentes métodos para a seleção de atributos visando a classificação de solos no 1º nível categórico do Sistema Brasileiro de Classificação de Solos (SiBCS). Uma vez selecionados os principais atributos, um algoritmo de aprendizado de máquina (árvore de decisão) é usado para a classificação de solos de acordo com o SiBCS. Os resultados obtidos abrem perspectivas para a classificação automática de solos, a partir de critérios definidos e de informações organizadas em bancos de dados.

PALAVRAS-CHAVE: árvores de decisão; mineração de dados; atributos de solos; métodos de discriminação de parâmetros.

AN EVALUATION OF FEATURE SELECTION METHODS FOR SOIL CLASSIFICATION

ABSTRACT: Data mining techniques have been used strategically to transform data into information and knowledge to support decision making in various domains. In agriculture, particularly, these techniques have shown to be efficient for selecting a set of attributes to create a data classification model in databases containing several variables. The objective of this work is to evaluate the efficiency of different methods for feature selection to support soil classification in the first categorical level of the Brazilian System for Soil Classification (BSSC). After selecting the main attributes, a machine learning algorithm (e.g., decision tree) is used for soil classification according to the BSSC. The results indicate interesting perspectives for soil automatic classification from proper criteria and information organized in databases.

KEY-WORDS: decision trees; data mining; soil attributes**;** discriminant methods.

1. INTRODUÇÃO

O desenvolvimento de um sistema de classificação de solos no Brasil está associado aos trabalhos de levantamento pedológicos e estes contribuem para a estruturação do sistema, em *feedback*, a partir da ampliação da base de conhecimento sobre os solos brasileiros. A adequada classificação de um solo permite estabelecer correlações com sua gênese e

[1] Bacharel em Matemática Aplicada e Computacional, pesquisador da Embrapa Informática Agropecuária, Campinas, SP, e-mail: fernando@cnptia.embrapa.br.
[2] PhD em Ciência da Computação, pesquisador da Embrapa Informática Agropecuária, Campinas, SP, e-mail: stanley@cnptia.embrapa.br.
[3] Doutora em Ciência do Solo, professora da Universidade Federal de São Carlos (UFSCar), Campus de Araras, SP, e-mail: assad@cca.ufscar.br.

evolução, assim como com fatores ambientais e econômicos relativos à sua ocupação, manejo, aptidão agrícola, entre outros (OLIVEIRA et al., 1992).

O Sistema Brasileiro de Classificação de Solos (SiBCS) apresenta uma estrutura hierárquica, multicategórica e, atualmente, encontra-se estruturado até o 4º nível categórico (SANTOS et al., 2006). Na classificação de um solo considera-se uma ampla gama de dados morfológicos, físicos, químicos e mineralógicos do perfil que o representa bem como aspectos ambientais do local do perfil, tais como clima, vegetação, relevo, material originário, condições hídricas, características externas ao solo e relações solo-paisagem. Trata-se de um sistema aberto, que admite a inclusão de novas classes que permitirão classificar todos os solos existentes no território nacional. Nesse sistema, utiliza-se uma chave de classificação cuja estrutura se apóia em atributos ditos de diagnósticos e em outros atributos gerais e na presença de horizontes diagnósticos superficiais e subsuperficiais. Em geral, esse trabalho é feito a partir da discussão de especialistas que estabelecem, com base em larga experiência e conhecimento sobre solos brasileiros, os critérios de definição de cada uma das classes previstas no sistema.

A seleção de atributos desempenha uma tarefa essencial dentro do processo de descoberta de conhecimento em banco de dados (GUYON & ELISSEEFF, 2003). Os alvos principais do processo de seleção de atributos são: i) melhorar a performance dos algoritmos de aprendizado de máquina; ii) simplificar os modelos de predição e reduzir o custo computacional para "rodar" esses modelos; e iii) fornecer um melhor entendimento sobre os resultados encontrados, uma vez que existe um estudo prévio sobre o relacionamento dos atributos.

Quando se tem uma grande quantidade de atributos envolvidos, torna-se necessária a minimização do grau de subjetividade inerente a processos classificatórios nos quais a seleção de variáveis se apóia na experiência de quem classifica. Isto pode se alcançado adotando-se procedimentos de mineração de dados. Os métodos mais utilizados em mineração de dados são: i) **Teste do Qui-quadrado (χ^2):** avalia os atributos individualmente por meio do teste do Qui-quadrado com relação à classe (atributo-alvo). Se o valor do teste for próximo a zero, o atributo analisado e o atributo-classe são independentes. Caso contrário, o atributo analisado é dependente do atributo-classe e deve, portanto, ser selecionado como um dos atributos relevantes (LIU & SETIONO, 1995); ii) **Seleção Baseada na Correlação (SBC):** considera um conjunto de atributos ao invés de avaliar a relevância individual de um atributo. A idéia é identificar subconjuntos de atributos não correlacionados entre si, mas que sejam altamente correlacionados com o atributo-classe. Para cada subconjunto é associada uma razão (mérito), em que o numerador pode ser visto como indicador do poder preditivo do conjunto de atributos e o denominador indica o grau de redundância que existe entre os atributos. O subconjunto com o maior mérito é, então, selecionado (HALL, 1998); iii) **Ganho de Informação (GI):** a seleção de atributos é feita medindo-se o ganho de informação de um atributo em relação ao atributo-classe. O ganho mede o quanto um atributo é capaz de separar um conjunto de exemplos em categorias. Atributos que possuem os maiores ganhos de informação são selecionados e posicionados em uma árvore de decisão, da raiz até as folhas, em ordem decrescente. A definição de ganho de informação vem da teoria da informação e é baseada na entropia, que mede a incerteza de uma variável (QUINLAN, 1986). Portanto, quanto menor a entropia, maior será o ganho de informação; e iv) **abordagem Wrapper (W):** define um subconjunto ótimo de atributos usando um algoritmo de apredizado de máquina (árvore de decisão, redes neurais, algoritmos genéticos, etc), levando em conta a tendência (*bias)* indutiva do algoritmo e sua interação com o conjunto de treinamento (KOHAVI & JOHN, 1998). Essa abordagem normalmente aumenta consideravelmente o tempo de execução do algoritmo, mas a precisão obtida no modelo de classificação tende a ser, em muitos casos, superior às demais abordagens.

Este trabalho tem por objetivo avaliar a eficiência de diferentes métodos para a seleção e/ou ranqueamento de atributos visando a classificação de solos no 1º nível categórico do Sistema Brasileiro de Classificação de Solos. Após a seleção dos principais atributos de solos, um algoritmo de aprendizado de máquina (árvore de decisão) é usado para a classificação de solos de acordo com o SiBCS.

2. MATERIAL E MÉTODOS

Dados de solos de diferentes localidades do Brasil foram compilados e, após uma avaliação dos métodos analíticos empregados, foi selecionado um conjunto de variáveis de 52 perfis de solos (Tabela 1). Cada perfil apresentava um ou mais horizontes de solos, perfazendo um total de 237 horizontes. De cada perfil foram considerados dados de local, posição no relevo, declividade, altitude, litologia, relevo local, erosão, drenagem e uso. Dos horizontes foram considerados dados referentes ao tipo, profundidade, cor úmida, textura, estrutura (tipo, tamanho e grau), superfície de fricção, cerosidade, consistência (seca, úmida e molhada), densidade de partícula, densidade do solo, quantidade de calhaus, cascalhos, areia, silte e argila, argila natural, grau de floculação, grau de dispersão, condutividade elétrica, pH em água, pH em cloreto de potássio, teores de alumínio, cálcio, magnésio e sódio, soma de bases solúveis, capacidade de troca catiônica, saturação por bases, saturação por alumínio, teores de fósforo, carbono, nitrogênio e matéria orgânica, quantidade de água em diferentes potenciais bem como teores totais de sílica, ferro, alumínio, titânio e fósforo e suas relações moleculares.

Tabela 1: Classes de solos utilizadas

Classe de Solo	Número de perfis utilizados
Cambissolo	4
Neossolo	10
Latossolo	32
Argissolo	4
Nitossolo	1
Planossolo	1

Nesse trabalho, foram considerados o teste do qui-quadrado (χ^2), a seleção baseada na correlação (SBC) e o ganho de informação (GI) para mineração de dados e a abordagem Wrapper (W) para a definição de um subconjunto ótimo de atributos.

Na avaliação da metodologia de seleção de parâmetros em classificação de solos, foram utilizados os algoritmos do *software* Weka, versão 3.4.4 (WITTEN & FRANK, 2005). Weka é um ambiente de software usado em problemas de descoberta do conhecimento, composto de uma coleção de algoritmos nas áreas de aprendizado de máquina e mineração de dados. Trata-se de um *software* livre que está disponível sob licença GNU (General Public License).

A classificação é uma tarefa de mineração de dados que tem por objetivo classificar itens de dados em uma entre diversas classes previamente definidas, com base em propriedades comuns, entre um conjunto de objetos no banco de dados. A técnica de classificação utiliza um conjunto de exemplos para desenvolver um modelo, conhecido como conjunto de treinamento. Em geral, o conjunto de treinamento contém dois terços dos exemplos disponíveis em um banco de dados. Um terço dos exemplos é então usado para testar a precisão do modelo. Após a construção do modelo de classificação, esse é usado para predizer classes de novos casos que estão para ser inseridos no banco de dados.

Para o estudo de caso de classificação de solos, foi utilizado um dos principais métodos propostos na literatura, que é a árvore de decisão (HAN & KAMBER, 2006).

3. RESULTADOS E DISCUSSÃO

O número de membros por classe no conjunto de dados utilizado não é balanceado (Tabela 1). Por exemplo, a classe Latossolo tem 32 membros, enquanto a classe Nitossolo tem apenas1. Se o modelo for construído considerando esse número de membros por classe, a precisão será comprometida devido a um possível enviesamento do modelo em direção às classes com um maior número de representantes. Para tentar amenizar esse problema, pode-se usar a técnica estatística de amostragem com reposição para balancear o número de membros por classe, na seleção de amostras para os conjuntos de treinamento e teste do modelo de classificação (BREIMAN, 1996). Duas restrições foram consideradas: a) preservação do número de membros por classe (classificação sem filtro); e b) amostragem com reposição seguindo a distribuição uniforme (classificação com filtro). Constata-se (Tabelas 2 e 3) que a precisão da classificação com filtro é superior àquela sem o uso do filtro.

Os métodos SBC e W (Tabela 2) não fazem o ranqueamento de atributos, mas seleciona, por meio de heurísticas, um conjunto de atributos candidato para a criação do modelo de classificação de dados de solos. Os resultados disponíveis referem-se à árvore de decisão sem e com a seleção de atributos.

Os resultados da árvore de decisão, para os métodos χ^2 e GI, sem o efeito dos métodos de seleção de atributos e também com a aplicação destes métodos para indicar a contribuição (ranqueamento) dos melhores atributos antes da geração da árvore de decisão (Tabela 3) produz um modelo simplificado. Após a classificação dos atributos em ordem de importância, o algoritmo de árvore de decisão foi aplicado ao conjunto de dados com 10% dos atributos que apresentaram uma maior contribuição para o modelo. O mesmo processo foi repetido para 20%, 30%, 40% e 50% dos atributos com maior peso para a geração do modelo (Tabela 3).

Tabela 2. Precisão da árvore de decisão após a aplicação dos métodos SBC e W.

Método	Sem seleção	Com seleção
SBC (s/ filtro)	91,98%	96,62%
SBC (c/ filtro)	99,58%	98,31%
W (s/ filtro)	91,98%	96,62%
W (c/ filtro)	99,58%	98,73%

Tabela 3. Precisão da árvore de decisão após a aplicação dos métodos χ^2 e GI.

Método	Sem seleção	10%	20%	30%	40%	50%
χ^2 (s/ filtro)	91,98%	64,56%	78,48%	80,17%	89,87%	92,41%
χ^2 (c/ filtro)	99,58%	95,78%	97,05%	99,16%	99,16%	99,16%
GI (s/ filtro)	91,98%	64,56%	64,56%	80,17%	92,41%	94,09%
GI (c/ filtro)	99,58%	83,12%	97,05%	99,16%	99,16%	99,16%

Nos métodos de seleção de atributos SBC e W (Tabela 2), quando o filtro não é usado, a precisão da árvore de decisão, com seleção de atributos, apresentou os melhores resultados para a classificação de solos no 1º nível categórico. Por outro lado, os métodos χ^2 e GI (Tabela 3) têm uma abordagem diferente dos métodos SBC e W, isto é, χ^2 e GI ranqueiam atributos com maior peso para a geração do modelo. Nota-se que, se até 50% dos atributos forem eliminados, usando χ^2 e GI, a precisão do modelo apresentado pela árvore de decisão é ainda bastante aceitável. Esse resultado indica que a classificação de solos no 1º nível categórico pode ser realizada com um pequeno número de atributos, simplificando o trabalho feito por um pedólogo, o que abre perspectivas para a classificação automática de solos, a partir de critérios definidos e de informações organizadas em bancos de dados.

4

5 4. CONCLUSÕES

Este trabalho avaliou a eficiência de vários métodos de seleção de atributos para a criação de um modelo para classificação de solos no 1º nível categórico do Sistema Brasileiro de Classificação de Solos (SiBCS). Os resultados obtidos revelam que a mineração de dados é uma ferramenta útil para dar suporte ao trabalho de classificação de solos. Em particular, foi observado que o modelo de classificação de dados (árvore de decisão), criado após a etapa de seleção de atributos, é bastante eficiente, alcançando uma precisão acima de 90% para todos os métodos analisados.

Dentre os resultados analisados, destaca-se a seleção de atributos usando o método Ganho de Informação (GI), com filtro (amostragem com reposição para balancear o número de membros por classe de solo antes da criação do modelo) que apresentou uma árvore de decisão com precisão de 99,16%, considerando apenas 50% dos atributos disponíveis no banco de dados original.

O uso de técnicas de mineração de dados para classificação de solos se constitui numa ação de pesquisa promissora e merece maior investigação. Como continuação desse trabalho de pesquisa, planeja-se: (a) analisar a eficiência de outros métodos de seleção de atributos, em conjunto com outros métodos de classificação de dados da literatura, tais como: redes neurais, modelos probabilísticos, entre outros; (b) explorar a eficiência dos métodos de seleção de atributos para a classificação de solos, além do 1º nível categórico; (c) identificar qual é o método de seleção de atributos e o método de classificação de dados mais apropriado para cada nível categórico do Sistema Brasileiro de Classificação de Solos (SiBCS).

5. REFERÊNCIAS BIBLIOGRÁFICAS

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 26, p. 123-140, 1996.

GUYON, I.; ELISSEEFF, A. An Introduction to variable and feature selection. **Journal of Machine Learning Research,** 3, p.1157-1182, 2003.

FAYYAD, U.; Irani, K. **Multi-interval discretization of continuous-valued attributes for classification learning**. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, p. 1022–1029, 1993.

HALL, M.A. ***Correlation-based feature selection for machine learning***. PhD thesis, Department of Computer Science, University of Waikato, Hamilto, New Zealand, 1998.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. 2ª ed. Morgan Kaufmann Publishers, San Francisco, USA, 2006. 770p.

KOHAVI, R.; JOHN, G. H. **The Wrapper Approach**, In: H. Liu & H. Motoda (Eds.) Feature Extraction, Construction and Selection: a data mining perspective, 33-49. Kluwer, 1998.

LIU, H.; SETIONO, R. **Chi2: Feature selection and discretization of numeric attributes**. *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence*, (1995) pp.388–391.

OLIVEIRA, J. B.; JACOMINE, P. K. T.; CAMARGO, M. N. **Classes gerais de solos do Brasil**: guia auxiliar para seu reconhecimento. 2ª ed. Jaboticabal: FUNEP, 1992. 201p.

QUINLAN, J.R. **Induction of decision trees**. Machine Learning, 1:81–106, 1986.

SANTOS, H. G. dos; JACOMINE, P. K. T.; ANJOS, L. H. C. dos; OLIVEIRA, V. A. de; OLIVEIRA, J. B. de; COELHO, M. R.; LUMBRERAS, J. F.; CUNHA, T. J. F. da. **Sistema brasileiro de classificação de solos**, 2 ed. Rio de Janeiro: Embrapa Solos, 2006. 306p.

WITTEN, I. H.; FRANK, E. **Data mining: practical machine learning tools and techniques**. 2nd ed. San Francisco: Morgan Kaufmann, 2005.