# Alternativas para construção de classificadores de solos brasileiros

Matheus Agostini Ferraciolli[1]
Luiz Manoel Silva Cunha[2]

**Resumo**: Este trabalho avalia os algoritmos J48, JRip e PART como alternativas para obtenção de modelos de classificação de solos, todos baseados em regras de classificação. Foram utilizadas observações das classes Nitossolos Brunos e Latossolos Brunos, extraídas do Sistema de Informação de Solos Brasileiros, Embrapa Solos. O algoritmo J48 apresentou modelos com maior acurácia e número de regras, em contraste com o JRip, que apresentou menor acurácia, com menor número de regras, o algoritmo PART se manteve no nível intermediário em ambas as métricas.

Palavras-chave: regras de classificação, pedologia, sistema de suporte à decisão, classificação de solos.

[1] Estudante de Engenharia Agrícola da Universidade Estadual de Campinas (Unicamp), estagiário da Embrapa Informática Agropecuária, Campinas, SP..

[2] Estatístico, mestre em Engenharia de Software, analista da Embrapa Informática Agropecuária, Campinas, SP.

## Introdução

A classificação de solos tem seu início em campo, durante o levantamento pedológico executado em um local, e termina após análises, em laboratório, das amostras coletadas (SANTOS et al., 2006). Centenas de atributos numéricos e categóricos descrevendo as características morfológicas, químicas e físicas dos solos, e do ambiente em que as amostras de solos se encontram, são descritas e quantificadas (MANUAL..., 2007).

O grande volume de dados disponível para classificação de solos, torna possível a utilização de ferramentas automáticas para esta finalidade. Algoritmos de Mineração de Dados, do inglês *Data Mining* (DM), têm demostrado um grande potencial para o aperfeiçoamento do processo de classificação de solos, por sua capacidade de extrair padrões de grandes quantidades de dados.

O objetivo deste trabalho é avaliar a capacidade de classificação dos algoritmos: a) J48 (SALZBERG, 1994), que cria uma árvore de decisão, podendo esta, também, ser descrita utilizando regras de classificação; b) JRIP (COHEN, 1995) e PART (FRANK; WITTEN, 1998) que criam uma lista de regras, somente.

O Sistema Brasileiro de Classificação de Solos (SiBCS) é estruturado em 4 níveis de classificação: a) ordem; b) subordem; c) grupo; e d) subgrupo. Para a classificação de uma amostra, neste trabalho foram utilizados atributos que caracterizam a ordem.

## Materiais e Métodos

O conjunto de dados consiste de 81 amostras de solos, sendo 38 classificadas como NB[3] e 43 como LB[4] , extraídas de perfis de solo localizados nos estados da Região Sul do Brasil. Os principais atributos para a modelagem (Tabela 1) foram escolhidos com os métodos de seleção de atributos automáticos: a) seleção de atributos por correlação (CFS; b) qui-quadrado (QQ); c) ganho de informação (GI; d) taxa de ganho de informação (TGI);

[3] Nitossolos Brunos.
[4] Latossolos Brunos.

e) wrapper (WR); f) relief (RF). E com os métodos não automáticos: a) lista unificada de atributos automática (LUAA), onde um atributo era selecionado caso três ou mais métodos o escolhessem; e b) lista híbrida unificada de atributos (LUAH), onde foram incluídos na LUAA atributos sugeridos por pedólogos por meio de um questionário aplicado.

Na construção dos modelos, foram adotadas duas abordagens: a) Situação A, utilizando os 35 atributos selecionados no processo de seleção e b) Situação B, suprimindo 8 atributos por sugestão do pedólogo responsável. Os atributos excluídos foram: 14, 16, 18, 22, 23, 24, 30 e 34, relacionados na Tabela 1, totalizando 27 atributos.

Os algoritmos de classificação têm diferentes formas de construção de regras. O J48 cria uma árvore de decisão com um atributo principal no chamado nó raiz, e termina em nós folha, com a classificação da observação em LB ou NB. O JRip cria uma lista de regras, onde cada regra é gerada individualmente, abrangendo um certo grupo de observações. O PART une elementos das duas abordagens e também retorna uma lista de regras, construídas de forma diferente das do JRip, utilizando árvores de decisão parciais na criação de cada regra.

Dada a necessidade de aplicação do mesmo algoritmo nas duas situações propostas, com diferentes configurações de hiperparâmetros, adotou-se a abordagem *Workflow Científico* (WC) (LUDÄSCHER et al., 2006), via o *Knowledge Flow Interface*, disponível no software Weka (HALL et al., 2009).

Para avaliação dos modelos, as métricas utilizadas foram: acurácia de classificação, dada pela razão entre observações classificadas corretamente e o total de observações; o número de regras geradas, pois menos regras implicam em um modelo mais generalizado; e estatística K (LANDIS; KOCH, 1977)

## Resultados e Discussão

A Figura 1 exibe uma visão parcial do WC construído para geração dos modelos. Este workflow trouxe facilidades: a) para visualização e entendimento do processo, por parte de profissionais que não são da área de computação; e b) de reprodução do processo empregando os três algoritmos, sobre diferentes conjuntos de dados e/ou parâmetros de configuração. A Figura 2 exibe as acurácias de cada modelo gerado utilizando o resultado do método

**Tabela 1.** Relação de atributos componentes das observações.

| No | Atributo | Ud[5] | No | Atributo | Ud |
|---|---|---|---|---|---|
| **1** | Grau de textura do solo | Ad[6] | **19** | Quantidade de silte no solo | g/Kg |
| **2** | Matiz | Ad | **20** | Quantidade de areia grossa no solo | g/Kg |
| **3** | Croma | N[7] | **21** | Quantidade de areia total no solo | g/Kg |
| **4** | Valor | N | **22** | Relação silte argila | % |
| **5** | Grau de desenvolvimento da estrutura | Ad | **23** | Teor de alumínio trocável existente no solo | g/Kg |
| **6** | Tamanho da estrutura do solo | Ad | **24** | Índice de saturação por alumínio trocável | |
| **7** | Formato da estrutura do solo. | Ad | **25** | Quantidade de Sílica (SiO2) no solo por ataque sulfúrico | g/Kg |
| **8** | Grau da consistência do solo quando úmido | Ad | **26** | Quantidade de Óxido de Alumínio (Al2O3) no solo por ataque sulfúrico. | g/Kg |
| **9** | Grau de plasticidade do solo | Ad | **27** | Quantidade de Óxido de Ferro (Fe2O3) no solo por ataque sulfúrico | g/Kg |
| **10** | Grau de pegajosidade do solo | Ad | **28** | Quantidade de Iodeto de potássio no solo. | g/Kg |
| **11** | Grau de cerosidade no solo | Ad | **29** | Teor de carbono orgânico no solo | g/Kg |
| **12** | Quantidade de cerosidade no solo | Ad | **30** | Relação carbono orgânico e nitrogênio | g/Kg |
| **13** | Grau de nitidez do solo | Ad | **31** | Relação Óxido de Ferro e Argila | Ad |
| **14** | Altitude do ponto de coleta do solo | Ad | **32** | Relação Óxido de Alumínio e Argila | Ad |
| **15** | Horizonte diagnóstico de superfície | Ad | **33** | Relação Silício e Argila | Ad |
| **16** | Material de origem do solo | --- | **34** | Relação Dióxido de Titânio e Argila | Ad |
| **17** | Quantidade de argila no solo | --- | **35** | Relação Iodeto de Potássio Argila | Ad |
| **18** | Capacidade de troca de cátions da argila | $cmol_c/dm^3$ | **36** | Rótulo da classe | Ad |

de seleção, e a Figura 3, o número de regras em cada um deles. As acurácias obtidas com os modelos variaram de 67 a 80,3%, considerando as duas situações.


**Figura 1.** Workflow Científico para automação do processo de geração de modelos.

[5] Unidade de medida.
[6] Admiensional.
[7] Numérico.

Nota-se os modelos obtidos com o J48, em geral, possuem maior acurácia do que outros algoritmos, acompanhados de mais regras. O JRip gera modelos menos precisos, porém, com número de regras inferior aos outros dois algoritmos. O PART fica no nível intermediário em ambas as métricas.

Quanto à estatística K, os valores dos modelos com acurácia superior a 75%, tiveram média de 0,53 para ambas as situações, considerado nível de moderada concordância, porém, quatro modelos, na situação A, obtiveram valores maiores do que 0,55, chegando próximo ao limite de 0,60, considerado de alta concordância (LANDIS; KOCH, 1977).

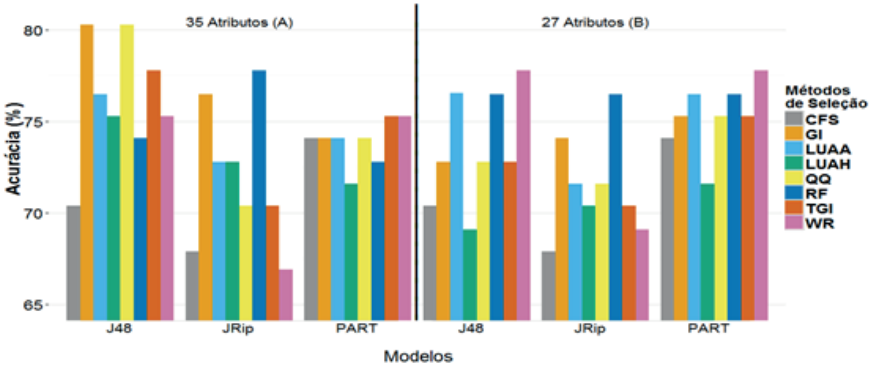


**Figura 2.** Nível de acurácia dos modelos obtidos.

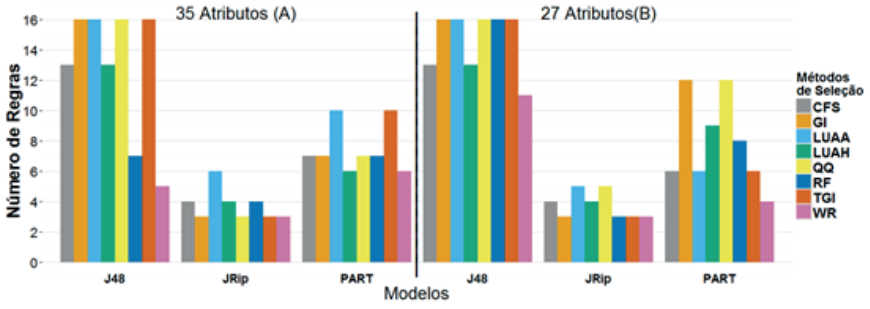


**Figura 3.** Número de regras de cada modelo.

## Considerações Finais

Com base nos resultados das métricas utilizadas, pode-se considerar que os algoritmos avaliados são ferramentas valiosas para o SiBCS por apresentar novas regras de classificação e por meio delas derivar novos conhecimentos. Na situação A, o de maior acurácia teve 80% de acerto com 16 regras de classificação e, na situação B, o modelo de destaque teve 78% de acerto com apenas 4 regras. Mesmo os modelos apresentando desempenhos semelhantes, a escolha pelo melhor não deve se basear no resultado de uma única métrica, e sim, em um conjunto delas. Além disso, deve-se levar em conta a capacidade do modelo em classificar uma observação em sua respectiva classe.

## Referências

COHEN, W. W. Fast effective rule induction. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 12.,1995, Tahoe. **Proceedings**... San Francisco: Morgan Kaufmann,1995. p. 115-123.

FRANK, E.; WITTEN, I. H. **Generating accurate rule sets without global optimization**. Hamilton: University of Waikato, 1998. p.144-151.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **ACM SIGKDD Explorations Newsletter**, v. 11, n.1, p.10-18, 2009. DOI: 10.1145/1656274.1656278.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159-174, Mar. 1977.

LUDÄSCHER, B.; ALTINTAS, I.; BERKLEY, C.; HIGGINS, D.; JAEGER, E.; JONES, M.; LEE, E. A.; TAO, J.; ZHAO, Y. Scientific workflow management and the Kepler system. **Concurrency and Computation: Practice and Experience**, v. 18, n. 10, p. 1039-1065, Aug. 2006. DOI: 10.1002/cpe.994.

MANUAL técnico de pedologia. 2. ed. Rio de Janeiro: IBGE, 2007. 320 p. (Manuais técnicos em geociências, 4).

SALZBERG, S. L. C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers, inc., 1993. **Machine Learning**, v. 16, n. 3, p. 235-240, 1994.

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. (Ed.). **Sistema brasileiro de classificação de solos**. 2. ed. Rio de Janeiro: Embrapa Solos, 2006. 306 p.