12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5

AVALIAÇÃO DA EFICIÊNCIA DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO AUTOMÁTICA DE SOLOS

Gabriel Teston **Vasconcelos**[1]**;** Stanley Robson de Medeiros **Oliveira**[2]

Nº 18603

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 um conjunto de atributos relevantes no processo de geração de modelos preditivos em bancos de dados com muitas variáveis. Este trabalho tem por objetivo avaliar a eficiência de diferentes algoritmos de Aprendizado de Máquina (AM) para classificação automática de solos, no 1º nível categórico do Sistema Brasileiro de Classificação de Solos (SiBCS). Os dados foram obtidos do projeto Mapeamento de Recursos Naturais do Brasil, liderado pelo Instituo Brasileiro de Geografia e Estatística (IBGE). Alguns algoritmos de AM (árvore de decisão, SVM, kNN) foram utilizados para classificação de solos de acordo com o SiBCS. Os resultados obtidos são promissores e 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, SVM, kNN, mineração de dados, atributos de solos.

1 Autor, Bolsista CNPq (PIBIC): Graduação em Engenharia da Computação, Unicamp, Campinas-SP; gabriel.vasconcelos@colaborador.embrapa.br

2 Orientador: Pesquisador da Embrapa Informática Agropecuária, Campinas-SP; stanley.oliveira@embrapa.br.

12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5

ABSTRACT – *Data mining techniques have been used strategically to transform data into information and knowledge to support decision making in various domains. In agriculture, in particular, these techniques are efficient to select a set of relevant attributes in the process of generating predictive models in databases with many variables. This work aims to evaluate the efficiency of different Machine Learning (ML) algorithms for automatic classification of soils, in the first categorical level of the Brazilian Soil Classification System (BSCS). The data were obtained from the Natural Resources Mapping Project in Brazil, led by the Brazilian Institute of Geography and Statistics (IBGE). Some ML algorithms (decision tree, SVM and kNN) were used to classify soil profiles according to SiBCS. The results obtained are promising and open perspectives for the automatic classification of soils, based on defined criteria and information organized in databases.*

Keywords: Decision trees, SVM, kNN, data mining, soil attributes.

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 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., 2013). 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 apoia 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

ISBN 978-85-7029-145-5

larga experiência e conhecimento sobre solos brasileiros, os critérios de definição de cada uma das classes previstas no sistema.

Embora o SiBCS seja atualizado periodicamente, ainda não existe um programa de computador disponível, que classifique perfis de solos, de tal forma que esse sistema computacional pudesse auxiliar profissionais em atividades de classificação de solos. Além de auxiliar o trabalho de pedólogos, o sistema poderia também ser utilizado como recurso didático, pois poderia explicar em detalhes o caminho que leva a uma determinada solução. Outros benefícios de uma ferramenta de software dessa natureza seriam: (a) auxiliar na validação de perfis previamente classificados; (b) classificar novos perfis de solo; (c) e subsidiar a evolução do SiBCS, por se tratar de um sistema taxonômico aberto com atualizações periódicas.

Uma alternativa para classificação automática de solos é a utilização de algoritmos de Aprendizado de Máquina (AM), já que o aprendizado automático explora o estudo e construção de algoritmos que podem aprender com dados, identificam padrões e tomam decisões com o mínimo de intervenção humana (MITCHELL, 1997). Tais algoritmos operam construindo um modelo a partir de *inputs* amostrais a fim de fazer previsões ou decisões guiadas pelos dados ao invés de simplesmente seguindo inflexíveis e estáticas instruções programadas. Enquanto que na inteligência artificial existem dois tipos de raciocínio (o indutivo, que extrai regras e padrões de grandes conjuntos de dados, e o dedutivo), o aprendizado de máquina só se preocupa com o indutivo (HASTIE et al., 2016).

Em particular, três classes de algoritmos de AM serão avaliadas neste trabalho: (a) **Simbólica**: algoritmo C4.5 para indução de árvores de decisão (QUILAM, 1993), por ser uma solução simples, intuitiva e amplamente utilizada na literatura; (b) **Baseada em instâncias** ou **Aprendizado *Lazy***, isso se deve ao fato do processamento ser atrasado até o momento de classificação de um novo exemplo, como é o caso do algoritmo k-vizinhos mais próximos (do inglês, kNN) (AHA e KIBLER, 1991); (c) **Aprendizado Estatístico**: SVM (Support Vector Machines ou Máquinas de Vetores Suporte), que possui boa capacidade de generalização na classificação de dados que não pertencem ao conjunto utilizado em seu treinamento (VAPNIK, 1995).

Assim, este trabalho tem por objetivo avaliar a eficiência de diferentes algoritmos de AM (C4.5, KNN e SVM) para classificação automática de solos no 1º nível categórico do SiBCS.

ISBN 978-85-7029-145-5

2. MATERIAL E MÉTODOS

2.1. ORIGEM DOS DADOS

Os dados de solos foram obtidos do Mapeamento de Recursos Naturais do Brasil, disponíveis no Instituto Brasileito de Geografia e Estatística (IBGE, 2018). Em particular, foram considerados os atributos de solos relacionados à pedologia.

Cada perfil apresentava um ou mais horizontes de solos, perfazendo um total de 23.534 de horizontes (instâncias). 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 aos atributos morfolóficos, físicos, químicos e mineralógicos. O conjunto de dados original foi composto de 23.534 instâncias e 95 atributos.

2.2. TRATAMENTO DOS DADOS

Foram removidos, do conjunto de dados originais, os perfis sem classificação e seus respectivos horizontes. Além disso, foram removidos atributos com mais de 80% dos valores faltantes. O conjunto de dados final foi constituído de 17.786 instâncias (horizontes) e 58 atributos de solos, sendo um deles a classificação no primeiro nível categórico do SiBCS. Foram considerados dados das 13 classes de solos, conforme Tabela 1.

Tabela 1. Classes de solos e seus respectivos números de instâncias.

CLASSE DE SOLO	NÚMERO DE INSTÂNCIAS POR CLASSE
LATOSSOLO	4.592
NEOSSOLO	1.384
ARGISSOLO	6.261
CAMBISSOLO	1.386
GLEISSOLO	819
ESPODOSSOLO	268
PLINTOSSOLO	992
PLANOSSOLO	698
NITOSSOLO	559
CHERNOSSOLO	328
VERTISSOLO	128
LUVISSOLO	338
ORGANOSSOLO	33

ISBN 978-85-7029-145-5

2.3. ALGORITMOS DE APRENDIZADO DE MÁQUINA

Conforme mencionado previamente, três classes de algoritmos de AM foram avaliadas neste trabalho:

- **Simbólica**: no campo do aprendizado simbólico, têm-se as árvores de decisão. O algoritmo C4.5 baseia-se em uma medida de separação de dados derivada da "entropia", aplicada sucessivas vezes aos dados, onde cada nível da árvore representa um atributo que particiona melhor o subconjunto segundo esta medida (QUILAM, 1993). Uma árvore de decisão é uma estrutura semelhante a um fluxograma na qual cada nó interno representa um "teste" em um atributo (por exemplo, no lançamento de uma moeda aparece cara ou coroa), cada ramo representa o resultado do teste e cada nó folha representa um rótulo de classe (decisão tomada após computar todos os atributos). Os caminhos da raiz para a folha representam regras de classificação.
- **Baseada em instâncias**: também conhecido como **aprendizado *Lazy***. Um dos principais algoritmos dessa classe de aprendizado é o kNN, proposto por Fukunaga e Narendra (1975). A ideia principal do kNN é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento. A variável k representa a quatidade de vizinhos mais próximos que serão utilizados para averiguar de qual classe a nova amostra pertence. Dentre os k exemplos, verifica-se a classe mais frequente. Essa classe é atribuída ao novo exemplo.
- **Aprendizado Estatístico**: a forma mais simples de particionar um espaço Euclidiano de n dimensões é através de hiperplanos. SVM (Support Vector Machines ou Máquinas de Vetores Suporte), baseia-se também nessa estratégia, porém, utiliza um tipo especial, o hiperplano de separação ótima. Trata-se de um hiperplano que divide as classes maximizando a margem de separação entre elas. A Figura 1 mostra um hiperplano de margem máxima. A linha mais espessa representa o hiperplano e as pontilhadas, as margens.

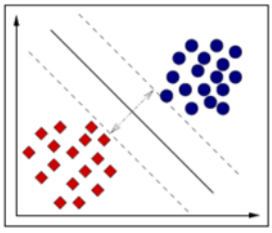

Figura 1. Hiperplano de serpação ótima para duas classes.

ISBN 978-85-7029-145-5

As SVM fundamentam-se em teorias estatísticas e matemáticas que possui boa capacidade de generalização na classificação de dados que não pertencem ao conjunto utilizado em seu treinamento (VAPNIK, 1995). Para se utilizar uma SVM para classificar múltiplas classes é necessário transformar o problema multi-classes em vários problemas de classes binárias (uma contra o resto) (Vapnik, 1998 ; Weston e Watkins , 1999).

2.4. VALIDAÇÃO DOS MODELOS PREDITIVOS

Para cada algoritmo de AM, foram gerados modelos preditivos para classificar os perfis de solos no primeiro nível categórico. Os modelos foram avaliados utilizando o método de validação cruzada (*cross-validation*) com 10-*folds*, e três métricas para cada uma das 13 classes de solo: (i) taxa de verdadeiros positivos (True Positive Rate); (ii) taxa de falsos positivos (False Positive Rate); (iii) precisão.

Para o algoritmo kNN foi utilizado um procedimento para ajuste de hiperparâmetros. Para isto, o valor de k foi avaliado de 1 a 10 com o objetivo de determinar o valor que maximiza as métricas do modelo.

Para avaliar os modelos, foram utilizados os algoritmos do *software* Weka, versão 3.8 (Witten e 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).

3. RESULTADOS E DISCUSSÃO

Para o algoritmo kNN foram testados os valores de k variando de 1 a 10, por meio validação cruzada. O valor de K = 1 foi selecionado pois apresentou os melhores ajustes para o modelo gerado.

A Tabela 2 apresenta os resultados das três métricas: TP Rate, FP Rate e Precision, para os três algoritmos em estudo (J48, kNN e SVM). Em particular, o algoritmo J48 é uma implementação do algoritmo C4.5 no Weka para indução de árvores de decisão.

ISBN 978-85-7029-145-5

Tabela 2. Resultados das métricas de avaliação dos modelos preditivos.

	J48 (C4.5)			kNN			SVM		
Classe	**TP Rate**	**FP Rate**	**Precision**	**TP Rate**	**FP Rate**	**Precision**	**TP Rate**	**FP Rate**	**Precision**
Latossolo	0,92	0,041	0,887	0,944	0,027	0,925	0,942	0,025	**0,93**
Neossolo	0,874	0,004	0,954	0,905	0,004	0,946	0,928	0,003	**0,96**
Argissolo	0,931	0,066	0,884	0,934	0,037	**0,932**	0,949	0,043	0,922
Cambissolo	0,83	0,006	**0,924**	0,9	0,007	0,917	0,874	0,006	0,922
Gleissolo	0,897	0,006	0,883	0,932	0,004	0,925	0,958	0,001	**0,973**
Espodossolo	0,907	0,001	**0,953**	0,937	0,002	0,893	0,944	0,001	0,937
Plintossolo	0,885	0,005	0,918	0,878	0,006	0,898	0,898	0,004	**0,923**
Planossolo	0,84	0,003	0,932	0,864	0,005	0,887	0,897	0,002	**0,953**
Nitossolo	0,758	0,001	0,959	0,896	0,003	0,911	0,939	0,001	**0,979**
Chernossolo	0,905	0,002	0,914	0,936	0,002	0,898	0,994	0,001	**0,964**
Vertissolo	0,773	0,001	0,908	0,914	0,001	0,929	0,961	0	**1**
Luvissolo	0,84	0,004	0,821	0,864	0,003	0,851	0,867	0,001	**0,942**
Organossolo	0,818	0	0,9	0,758	0	0,926	0,97	0	**1**

Os resultados revelaram que os três algoritmos apresentaram excelentes resultados para a classificação das 13 classes de solo. No entanto, o algoritmo SVM apresentou, em geral, valores levemente superiores aos demais. De acordo com Han et al. (2011), algumas das principais características das SVMs que tornam seu uso atrativo são: (a) boa capacidade de generalização, evitando o efeito de overfitting**;** (b) robustez em grandes dimensões; (c) apresenta bons resultados quando comparados a outros algoritmos de AM disponíveis na literatura**.**

Por outro lado, o algoritmo C4.5 tem uma grande vantagem em relação ao demais (SVM e kNN), já que uma árvore de decisão é intuitiva e suas regras de classificação podem ser facilmente entendidas por um pedólogo. Além disso, as regras de classificação podem ser utilizadas para retroalimentar o SiBCS e colaborar com sua evolução, uma vez que uma árvore de decisão pode apresentar regras mais curtas (sem a necessidade de considerar todos os atributos do SiBCS), o que pode representar uma redução expressiva não só no tempo para se classificar um determinado solo, mas também uma redução em termos financeiros pois nem todos os atributos precisam ser testados, evitando assim a necessidade de se analisar exaustivamente todos os atributos físicos, químicos e mineralógicos em laboratórios credenciados para esta finalidade. Maximo et al. (2007) utilizaram o algoritmo C4.5 em seus experimentos e obtiveram bons resultados quando classificaram solos no primeiro nível categórico do SiBCS. No entanto, os autores não compararam C4.5 com outros algoritmos de classificação. Eles estudaram a eficiência de métodos de seleção de atributos para melhorar o desempenho do algoritmo C4.5 na classificação de solos.

ISBN 978-85-7029-145-5

Já no caso do algoritmo kNN, foi observado que seu tempo de execução é muito grande quando comparado aos demais, pois trata-se de um algoritmo *Lazy*, que calcula todas as distâncias quando um novo é classificado. Além disso, o algoritmo kNN exige um procedimento de ajuste de hiperparâmetros para maximizar os valores das métricas do modelo gerado.

4. CONCLUSÃO

Este trabalho avaliou a eficiência de alguns algoritmos de Aprendizado de Máquina classificação de solos, no 1º nível categórico do Sistema Brasileiro de Classificação de Solos (SiBCS). Os resultados obtidos revelam que tais algoritmos são úteis para dar suporte ao trabalho de classificação de solos. Em particular, foi observado que o modelo de classificação baseado no algoritmo SVM é bastante eficiente, alcançando uma precisão acima de 90% para todas as 13 classes de solo.

O uso de algoritmos de Aprendizado de Máquina 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 algoritmos (Redes Neurais Profundas, Random Forest, Boosting), em conjunto com métodos de seleção e atributos; (b) explorar a eficiência dos algoritmos para a classificação de solos até o 4º nível categórico; (c) identificar qual é o método de seleção de atributos e o algoritmo de classificação de dados mais apropriado para cada nível categórico do Sistema Brasileiro de Classificação de Solos (SiBCS).

5. AGRADECIMENTOS

Ao programa CNPq/PIBIC pela concessão da bolsa de Iniciação Científica, processo Nº 106600/2018-4 para o aluno Gabriel Teston Vasconcelos.

6. REFERÊNCIAS

AHA, D.; KIBLER, D. Instance-based learning algorithms. **Machine Learning**. 6:37-66, 1991.

FUKUNAGA, K.; NARENDRA, P. M. A branch and bound algorithm for computing k-nearest neighbors. IEEE Transactions on Computers, v. 100, n. 7, p. 750–753, 1975.

HAN, J.; KAMBER, M.; PEI. J. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 3 ed, San Francisco, CA, USA, 2011.

ISBN 978-85-7029-145-5

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed., Springer, 2016.

IBGE. Mapeamento de recursos naturais do Brasil: Escala 1:250.000. Documentação Técnica Geral. Rio de Janeiro, Fevereiro 2018. Disponível em:

http://geoftp.ibge.gov.br/informacoes_ambientais/vegetacao/vetores/escala_250_mil/recorte_milionesimo/DOCUMENTACAO_TECNICA_MRN.pdf. Acessado em 5 de maio de 2018.

MÁXIMO, F. A.; OLIVEIRA, S. R. de M.; LOPES-ASSAD, M. L. R. C. Avaliação de métodos de seleção de atributos para classificação de solos. **Anais** do Sexto Congresso Brasileiro de Agroinformática (SBIAgro 2007). São Pedro, SP, 8 a 11 de outubro de 2007.

MITCHELL, T. M. Machine Leraning. McGraw-Hill Science/Engineering/Math, 1997. 432p.

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; LUMBRERAS, J. F.; COELHO, M. R.; ALMEIDA, J. A, de; CUNHA, T. J. F. da; OLIVEIRA, J. B. de. **Sistema brasileiro de classificação de solos**, 3 ed. Rio de Janeiro: Embrapa Solos, 2013.

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

Vapnik, V. N. Statistical Learning Theory. John Wiley and Sons, New York, 1998.

Weston, J.; Watkins, C. Multi Class Support Vector Machines, in Proceedings of ESANN99, ed. M. Verleysen, D. Facto Press, Brussels, pp. 219-224, 1999.

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