Utilizando processamento em cascata e agrupamento em imagens para otimizar modelos de classificação de solos

Gabriel Teston Vasconcelos[1], Kleber Xavier Sampaio de Souza[2], João Camargo Neto[3], Stanley Robson de Medeiros Oliveira[4]

[1] Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP, Brasil, gtestonvasconcelos@gmail.com

[2,3,4] Embrapa Informática Agropecuária, Campinas, SP, {kleber.sampaio, joao.camargo, stanley.oliveira}@embrapa.br

RESUMO

O presente artigo propõe duas novas abordagens para a classificação automática de solos, com o objetivo de aperfeiçoar o processo classificatório. As abordagens consistem no processamento em cascata para cada nível categórico e a organização dos horizontes do um perfil em forma de imagens possibilitando a classificação de perfis de solo como um todo. Constatou-se uma leve melhora, contudo a mesma carrega um grande custo computacional.

PALAVRAS-CHAVE: Árvores de decisão, Floresta aleatória, SVM, Mineração de dados, Atributos de solo.

ABSTRACT

This article proposes two new approaches to the automatic soil classification, with the objective to improve the classification process. Those approaches consist in the cascate processing for each classification level and the organization of the soil horizons into an image being able to classify the soil profile as a whole. It was found some improvement but it comes with a huge computational cost.

KEYWORDS: Decision Tree, Random Forest, SVM, Data Mining, Soil Attributes.

INTRODUÇÃO

Com o intuito de aperfeiçoar a classificação automática de perfis de solos em aplicações práticas (Vasconcelos et al., 2018) propõe-se neste trabalho o processamento em cascata levando em consideração a hipótese que classificadores de níveis diferentes podem não concordar durante o processo de classificação.

Também se propõe uma maneira de obter a classificação de perfis de solos levando em consideração seus *n* horizontes simultaneamente através da representação dos atributos de forma contínua.

O objetivo final desta publicação é avaliar a viabilidade do uso das técnicas propostas em aplicações práticas de classificação automática de solos.

MATERIAL E MÉTODOS

Origem dos dados

Para o treinamento dos modelos analisados foram utilizados o dados providos pelo Mapeamento de Recursos Naturais do Brasil, disponíveis no Instituto Brasileiro de Geografia e Estatística (IBGE, 2018). Dos dados obtidos, foram considerados os referentes aos atributos físicos e químicos dentro do contexto da pedologia. Resultando em 23.534 horizontes com 95 atributos cada, referentes a 5013 perfis de solo.

Tratamento dos dados

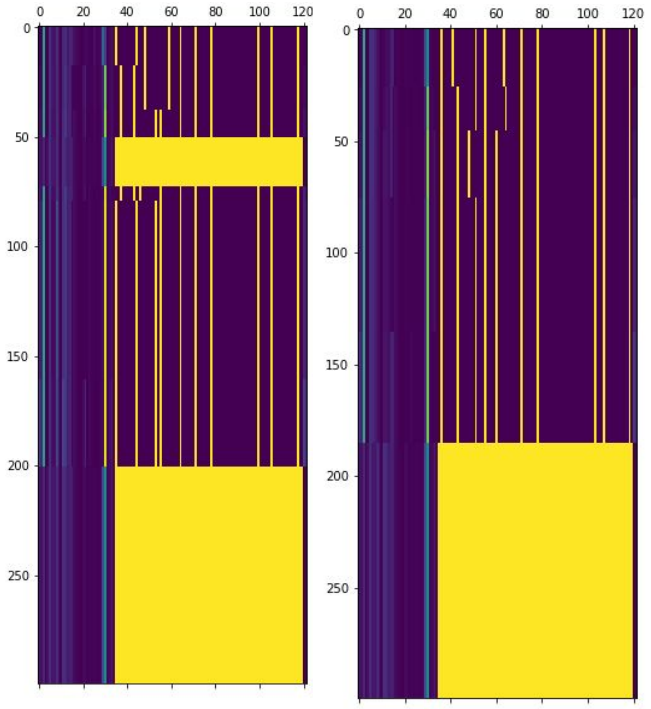
Dos 95 atributos originais foram removidos os atributos de identificação, que não são utilizados para classificação de solos e os atributos com mais de 50% de valores faltantes.

A base de dados é composta de atributos numéricos, assim como atributos categóricos. Para os atributos numéricos foi aplicada uma transformação comprimindo os valores entre 0 e 1 e valores faltantes substituídos pela média (Geron, 2017), para os valores categóricos foi aplicada a técnica de one-hot-bit encoding (Geron, 2017), transformando cada atributo categórico em n atributos binários, sendo n o número de categorias presentes no atributo original e valores faltantes representando uma das categorias. Também foram removidos os atributos categóricos com mais de 30 categorias, pois a sua presença apenas piorava a classificação final.

Por fim foram considerados os horizontes até os 3 metros de profundidade, dada que a classificação feita por pedólogos é feita considerando esse intervalo, para a criação de

“imagens” representando os perfis de maneira “contínua”. Onde cada faixa de valores descrita pelos atributos de profundidade de cada horizonte foi populada com os valores do horizonte a qual pertencia e em casos de valores faltantes completados com os valores centrais correspondentes. Cada imagem é composta de um canal, com dimensões de (300, 122) pixels como na Figura 1.

Figura 1 – Imagem representando perfis de solo de diferentes classes.



Cambissolo Latossolo

Fonte: (Autor, 2019)

Algoritmos de aprendizado de máquina

Foram utilizados três algoritmos de Aprendizado de Máquina (AM) todos contidos na biblioteca Scikit-learn (Pedregosa *et al, 2011*).

Árvore de Decisão (sklearn.tree.DecisionTreeClassifier)

Consiste na criação de um grafo direcional em formato de árvore começando da raiz, onde cada nó contém uma pergunta que leva ao próximo dependendo da resposta da mesma.

Floresta Aleatória (sklearn.ensemble.RandomForestClassifier)

Consiste na criação de um comitê de árvores de decisão, onde cada uma tem acesso a uma combinação aleatória dos atributos.

Máquinas de Vetor de Suporte (SVM) (sklearn.svm.SVC)

Consiste na divisão do espaço euclidiano gerado pelos atributos utilizando hiperplanos de separação ótima no intuito de dividir esse espaço nas classes desejadas.

Processamento dos dados em cascata

O processamento em cascata foi definido de maneira que a classificação do nível *n* depende dos atributos do perfil mais a classificação dos níveis anteriores. portanto a classificação nível 4 (Ordem, Subordem, Grande Grupo, Subgrupo), depende dos atributos do perfil mais a classificação dos níveis 1,2 e 3.

Diferentemente da abordagem anterior (Vasconcelos et al., 2018) onde se classificavam horizontes, cada modelo foi treinado para classificar um perfil, tornando o processo de classificação mais próximo ao realizado por pedólogos. O processamento em cascata possibilita que durante o processo de classificação o modelo que classifica o nível n+1 aperfeiçoe a classificação do modelo do nível anterior, uma vez que o de nível n+1 é mais especializado e capaz de verificar outras relações entre os atributos.

Validação dos modelos

A validação dos modelos foi feita em duas etapas, **Relativa** onde assumiu-se o acerto das classificações anteriores e **Absoluta** onde a classificação foi realizada a partir do processamento em cascata usando os modelos gerados, tendo como métrica de avaliação a acurácia.

Para a criação dos modelos foi usada uma divisão entre treino e teste de 70% dos perfis para treino e 30% para teste.

RESULTADOS E DISCUSSÃO

Com os modelos em casca treinados foi possível comparar o resultado na classificação e de cada algoritmo para cada nível de classificação com a classificação de um modelo treinado nos mesmos dados, porém que classifica cada nível individualmente, como pode ser visto nas Tabelas 1 a 3.

O ganho médio na acurácia utilizando o processamento em cascata foi de 4,3%, sendo possível notar que o modelo baseado em SVM foi que tirou maior proveito do processamento em cascata com um ganho médio na acurácia de 7,5%.

Tabela 1 – Acurácia obtida com Decision Tree para os diferentes métodos de classificação.

Nível	Tipo de classificação (Decision Tree)		
	Sem cascata	Com cascata (Relativa)	Com Cascata (Absoluta)
Ordem	0.532	0.532	0.532
Subordem	0.327	0.6613	0.360
Grande grupo	0.249	0.7184	0.303
Subgrupo	0.209	0.650	0.254

Tabela 2 – Acurácia obtida com Random Forest para os diferentes métodos de classificação.

Nível	Tipo de classificação (Random Forest)		
	Sem cascata	Com cascata (Relativa)	Com Cascata (Absoluta)
Ordem	0.674	0.674	0.674
Subordem	0.482	0.553	0.521
Grande grupo	0.402	0.473	0.438
Subgrupo	0.340	0.396	0.3512

Tabela 3–Acurácia obtida com SVM os diferentes métodos de classificação.

Nível	Tipo de classificação (SVM)		
	Sem cascata	Com cascata (Relativa)	Com Cascata (Absoluta)
Ordem	0.527	0.527	0.527
Subordem	0.330	0.355	0.354
Grande grupo	0.120	0.260	0.259
Subgrupo	0.094	0.231	0.231

Tabela 4 – Ganho de acurácia (Absoluta) com processamento em cascata.

Nível	Classificador		
	Decision Tree	Random Forest	SVM
Ordem	0.000	0.000	0.000
Subordem	0.033	0.039	0.024
Grande grupo	0.054	0.036	0.140
Subgrupo	0.045	0.012	0.137

Média	0.033	0.022	0.075

CONCLUSÕES

Mesmo com um possível ganho na acurácia com o processamento em cascatas não é possível constatar uma melhora significativa dessa técnica, tornando pouco viável sua utilização em na aplicação proposta em (Vasconcelos et al., 2018) uma vez que cada classificação em cascata chega a ser 4 vezes mais complexa computacionalmente.

Porém, foi possível estabelecer um método para a classificação de perfis, ao invés de horizontes, tornando a classificação automática de solos mais próxima à realizada por pedólogos.

AGRADECIMENTOS

Os autores agradecem ao programa CNPq/PIBIC pela concessão da bolsa de Iniciação Científica, processo Nº125043/2018-0 para o aluno Gabriel Teston Vasconcelos e à equipe do projeto SmartSolos da Embrapa pelo apoio oferecido durante o desenvolvimento.

REFERÊNCIAS

GERON, A. Hands-On Machine Learning with Scikit-Learn & Tensor Flow - Concepts, Tools and Techniques to Bild Intelligent Systems. O'Reilly Media Inc, Sebastopol, USA. 2017.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E., Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

VASCONCELOS, G. T.; SOUZA, K. X. S. de; OLIVEIRA, S. R. de M.; CAMARGO NETO, J. Montagem de ambiente para classificação de solos usando ScikitLearn. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 14., 2018, Campinas. Resumos expandidos... Brasília, DF: Embrapa, 2018. p. 104-110. (Embrapa Informática Agropecuária. Eventos técnicos & científicos, 1). Editores técnicos: Carla

Geovana do Nascimento Macário, Carla Cristiane Osawa, Flávia Bussaglia Fiorini, Maria Fernanda Moura, Poliana Fernanda Giachetto.