Automação de experimentos científicos utilizando KnowledgeFlow

Nádia Vieira Ribeiro[1]
Luiz Manoel Silva Cunha[2]

Resumo: Este trabalho apresenta um estudo de caso referente à utilização da tecnologia KnowledgeFlow, para automação de experimento científico, em Mineração de Dados, na busca por novos conhecimento visando ao aperfeiçoamento dos processos de caracterização e classificação de solos tipo Bruno, classificados nas classes Latossolo Bruno e Nitossolo Bruno. O KnowledgeFlow, mesmo com algumas limitações pontuais, trouxe um retorno positivo. Sua inserção no Sistema Brasileiro de Classificação de Solos (SiBCS) contribuirá para tornar a classificação dos solos mais ágil e de maneira mais otimizada.

Palavras-chave: workflow científico, automação de processos, mineração de dados, weka.

Introdução

Os processos de elaboração de conhecimento científico baseiam-se em constantes repetições de experimentos e análises de grandes conjuntos de dados, que podem tornar a atividade exaustiva ou até mesmo inviável. Diante disso, a busca por metodologias e ferramentas que agilizem os procedimentos são indispensáveis para garantir melhor estruturação, flexibilidade e sucesso do trabalho em questão.

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

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

Uma das metodologias utilizada atualmente é a adoção de fluxos de trabalho, mais conhecidos como *workflows*. Segundo Cuevas-Vicentitin et al. (2012) e Yu; Buyya (2005), *workflow* científico é uma abordagem para automação de um experimento ou de um processo científico, expressa em termos das atividades a serem executadas e, principalmente, das dependências dos dados manipulados.

Este trabalho refere-se ao desenvolvimento de modelos de classificação por meio de Árvores de Decisão, utilizando a ferramenta KnowledgeFlow do software Weka, com intuito de aprimorar, automatizar e facilitar os processos de mineração de dados - como seleção de atributos, balanceamento das classes e modelagem - e suas análises, resultando em melhorias na caracterização e na classificação mais acurada dos solos Brunos no escopo do SiBCS.

Materiais e Métodos

O estudo foi dirigido no Laboratório de Inteligência Computacional (LabIC), situado na Embrapa Informática Agropecuária[3], Unidade Descentralizada da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).

Os dados usados foram extraídos do Sistema de Informação de Solos Brasileiros[4] e de documentos - como ficha para descrição morfológica dos solos no campo, teses, dissertações de mestrado e artigos científicos Inicialmente, 182 observações contendo 28 atributos, sendo o último denominado “classe” - para o estudo de caso. As observações obtidas são oriundas de levantamentos pedológicos realizados nos estados do Paraná, de Santa Catarina e do Rio Grane do Sul. Elas foram agrupadas nas classes: Latossolo Bruno (LB), Nitossolo Bruno (NB) e Outros Latossolos, Nitossolos e Cambissolos (OLNC). A Figura 1 ilustra os processos incluídos no experimento em questão e as informações de parametrização.

O KnowledgeFlow surgiu como uma alternativa para a ferramenta Explorer do software Weka (HALL; REUTEMANN, 2008), embora algumas funções sejam específicas de cada uma. A ferramenta consiste no fluxo de dados

[3] Embrapa Informática Agropecuária: https://www.embrapa.br/informatica-agropecuaria

[4] Sistema de Informação de Solos Brasileiros: http://www.sisolos.cnptia.embrapa.br

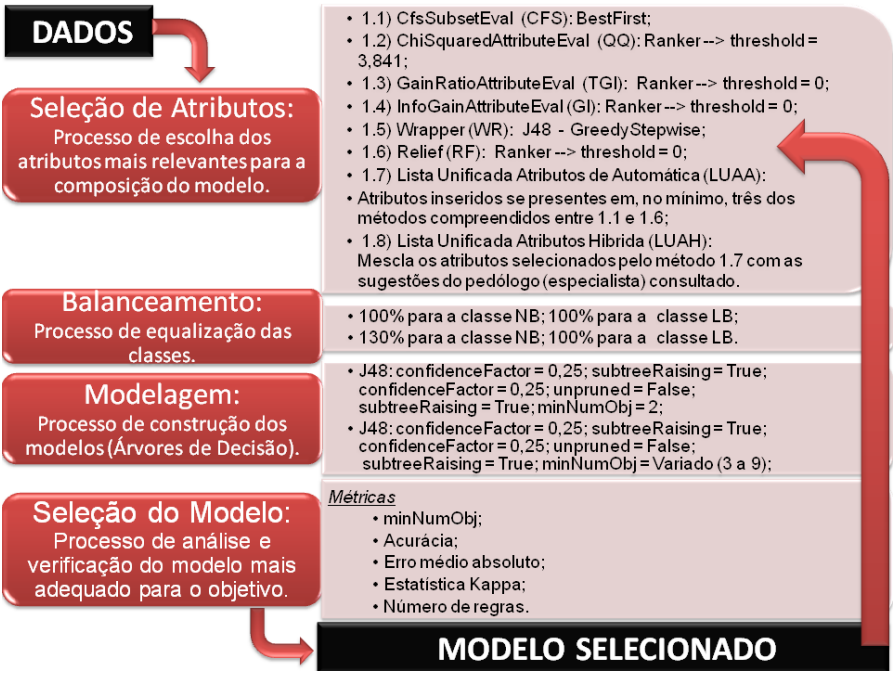

Figura 1. Processos do experimento de solos Brunos.

conectados de maneira a formar o processo desejado, onde o usuário seleciona o componente da barra de ferramentas e o dispõe da forma que lhe for conveniente. Estes são parametrizados, conectados em um grafo direcionados para processar e analisar os dados. Dessa forma, é possível visualizar em termos de como os dados fluem através do sistema.

Resultados e Discussão

Os processos metodológicos descritos acima foram estruturados e automatizados com a utilização do KnowledgeFlow, estando inseridos no âmbito da atividade “Descoberta de conhecimentos em bases de dados de solos: uma contribuição à classificação dos solos Brunos”, incorporada no projeto “Pesquisa e Inovação para o aprimoramento da taxionomia de Solos Brasileiros”, este conduzido pela Embrapa Solos. Apenas com o intuito de distinção de tarefas, o fluxo de seleção de atributos foi feito separadamente

do fluxo usado para balanceamento e modelagem, conforme representa a Figura 2.

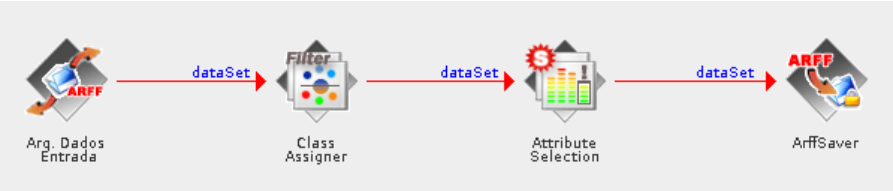

Figura 2. KnowledgeFlow utilizado para a seleção de atributos.

Para o processo de seleção de atributos, o uso da ferramenta foi relevante para a redução e a simplificação das etapas construtivas, visto que com um mesmo fluxo, apenas alterando o algoritmo de seleção e o nome do arquivo de saída, foi possível usar 6 métodos distintas (CFS, QQ, TGI, GI, WR e RF) com um despendimento de tempo mínimo.

Para determinar o melhor nível de balanceamento de classes, utilizou-se o fluxo apresentado na Figura 3, onde o ramo superior refere-se ao balanceamento de 100% para a classe NB (1º componente SMOTE) e 100% para a classe LB (2º componente SMOTE) e o ramo inferior refere-se ao balanceamento de 130% para a classe NB e de 100% para a classe LB.

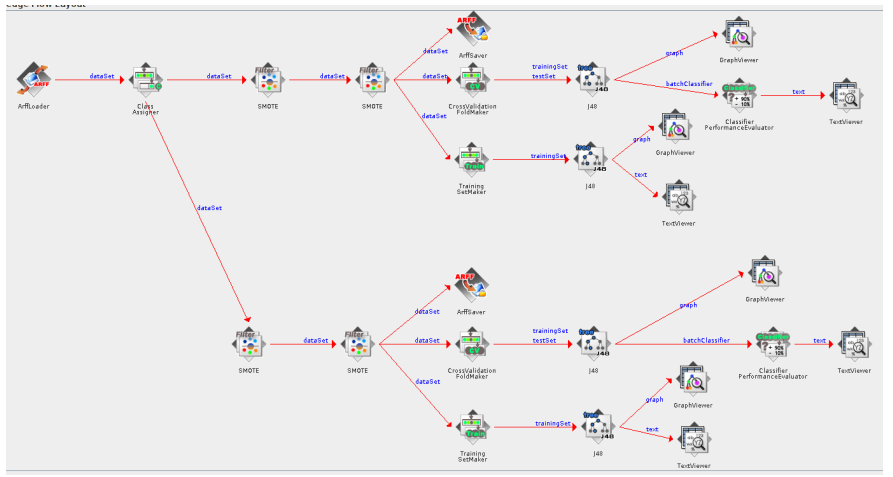

Figura 3. KnowledgeFlow utilizado para o balanceamento e modelagem com número mínimo de objetos fixo.

O KnowledgeFlow permitiu, a partir de diversas combinações, encontrar os melhores níveis de balanceamento pois, dada uma porcentagem, um modelo era gerado automaticamente; assim, selecionando o modelo de melhor desempenho, determinava-se o melhor nível de balanceamento.

Analogamente ao procedimento anterior, para o número mínimo de objetos por folha fixo variado, utilizou-se o fluxo apresentado na Figura 4, em que os ramos à direita originavam os resultados e parâmetros da melhor árvore, usando o método de amostragem de dados Cross Validation 10 folds para cada número mínimo de objetos e os ramos à esquerda originavam as árvores - impressas em formato de texto ou gráfico, para cada número mínimo de objetos correspondente.

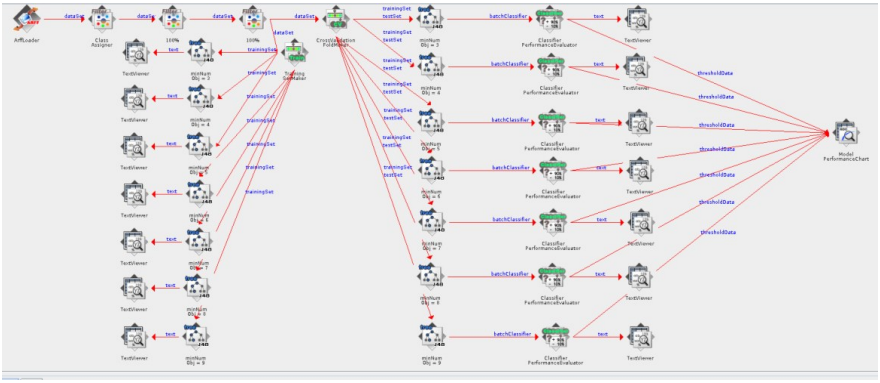

Figura 4. KnowledgeFlow utilizado para o balanceamento e modelagem com número mínimo de objetos variado.

Dentre as vantagens e benefícios agregados pelo uso da ferramenta, destacam-se a estruturação e organização dos procedimentos realizados (ILKAY, 2006) - o que implica na facilidade de exposição do trabalho para especialistas de outras áreas, a redução no tempo gasto para execução do trabalho proposto e a viabilidade de correções e adequações pontuais sem que seja necessário o retrabalho completo. KnowledgeFlow permite o projeto e a execução de configurações para processamento de dados de forma contínua, recurso este que a ferramenta Explorer, do software Weka, não suporta.

As limitações encontradas referem-se à ausência da possibilidade de salvar os resultados dos buffers de maneira automatizada, com a inserção de um ícone saver, por exemplo; a ferramenta permite armazenar os resultados de

forma individual. Na versão utilizada (3.6.10), a ferramenta não permitiu salvar as representações gráficas das árvores, fazendo-se necessário a utilização de meios alternativos como Print Screen. O arquivo do KnowledgeFlow gerado na extensão ".kf", não foi compatível com outras versões testadas (3.6.9 e 3.7.12); a solução encontrada foi salvar o arquivo utilizando a extensão ".kfml".

Considerações Finais

Concluído esse estudo, foi possível averiguar que o uso do KnowledgeFlow, apesar de ainda apresentar algumas limitações pontuais, trás um retorno positivo para a implementação da experimentação científica automatizada. A continuação desse trabalho pode se dar avaliando a utilização da ferramenta em rede.

Referências

ALTINTAS, I.; BARNEY, O.; CHENG, Z.; CRITCHLOW, T.; LUDAESCHER, B.; PARKER, S.; SHOSHANI, A.; VOUK, M. Accelerating the scientific exploration process with scientific workflows. **Journal of Physics**, p. 468-477, 2006.

BOUCKAERT, R. R.; FRANK, E.; HALL, M.; KIRKBY, R.; REUTEMANN, P.; SEEWALD, A.; SCUSE, D. **Weka manual for Version 3-6-10**. Hamilton: University of Waikato, 2013. Disponível em: <http://facweb.cs.depaul.edu/mobasher/classes/csc478/Notes/WekaManual-3-6-10-1.pdf>. Acesso em: 15 out. 2015.

CUEVAS-VICENTTÍN, V.; DEY, S.; KÖHLER, S.; RIDDLE, S.; LUDÄSCHER, B. Scientific workflows and provenance: Introduction and research opportunities. **Datenbank-Spektrum**, 12, n. 3, p. 193-203, Oct. 2012.

HALL, M.; REUTEMANN, P. **WEKA knowledgeflow tutorial for version 3-5-8**. University of Waikato, 2008. 13 p. Disponível em: <http://software.ucv.ro/~eganea/AIR/KnowledgeFlowTutorial-3-5-8.pdf>. Acesso em: 15 out. 2015.

YU, J.; BUYYA, R. A taxonomy of scientific workflow systems for grid computing. **Sigmod Record**, v. 34, n. 3, p. 44-49, 2005.