ciência da computação

V EPAC - Encontro Paranaense de Computação ISSN:1981-8653
18 a 20 de setembro de 2013
Unioeste – Cascavel – Paraná

O MÓDULO DE AGRUPAMENTO DE DADOS DA FERRAMENTA YADMT

Clodis Boscarioli[1]; Mateus Felipe Teixeira[1]; Rosangela Villwock[1]; Thiago Magalhães Faino[1]
[1]Universidade Estadual do Oeste do Paraná – Unioeste – Cascavel – Paraná
clodis.boscarioli@unioeste.br;
{mateusteixeira_,rosangelamat,mf_thiago}@hotmail.com

Resumo
Com crescimento do volume de dados armazenados, muitas informações e conhecimento podem estar sendo perdidos devido às limitações humanas em analisar e interpretá-lo. Desse modo, surgem ferramentas e técnicas para o auxílio na extração de conhecimento em um processo chamado de Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases - KDD), que pode ser dividida em três principais etapas distintas: Pré-Processamento, Mineração de Dados e Pós-Processamento. A YADMT (Yet Another Data Mining Tool) é uma ferramenta desenvolvida na Unioeste de forma modular, facilitando a inserção de novos módulos do processo KDD. A ferramenta é constituída por sete métodos de agrupamento de dados, incluindo métodos de visualização.

Palavras-chave: Descoberta de Conhecimento em Banco de Dados, Mineração de Dados, Agrupamento de Dados.

1. Introdução
O volume de dados armazenados e manipulados pela maioria das organizações cresce diariamente a uma taxa que ultrapassa a nossa capacidade de analisar, sintetizar e extrair conhecimento a partir desses dados. Muitas vezes, esse grande volume de dados contém informações úteis, as quais podem ser chamadas de "conhecimento", sendo que, em geral, esta informação não está facilmente disponível ou identificada. Analistas humanos podem gastar semanas para descobrir este conhecimento e, por este motivo, alguns bancos de dados grandes nunca recebem uma análise detalhada adequada como deveriam (TAN, STEINBACH; KUMAR, 2005).

Este contexto justifica a existência de uma área de investigação, chamada de *KDD*, cujo principal objetivo é extrair o conhecimento a partir dos dados que sejam úteis nas tomadas de decisões, utilizando métodos de diferentes áreas, que incluem Estatística, Inteligência Artificial, Aprendizagem de Máquinas e Reconhecimento de Padrões (TAN, STEINBACH; KUMAR, 2005). Segundo Fayyad et al. (1996), o processo KDD é um processo não trivial de descobertas de padrões válidos, novos, úteis e acessíveis.

Emergem então ferramentas que contemplam, integralmente ou em partes, todo o processo de *KDD*. Algumas dessas ferramentas são encontradas diretamente acopladas aos Sistemas Gerenciadores de Banco de Dados (SGBD), que por serem proprietárias não habilitam outros formatos de dados que fujam ao padrão definido pelo SGBD. Existem outras, ditas independentes, mas que impõe a necessidade da transformação dos dados para

o formato de entrada específico destas ferramentas e quando apresentam acesso direto ao SGBD, este é de difícil parametrização.

De forma alternativa, foi proposta a partir de Benfatti *et al*. (2010) a YADMT - *Yet Another Data Mining Tool*, uma ferramenta de *KDD* em desenvolvimento na UNIOESTE, no âmbito do Grupo de Inteligência Aplicada (GIA). É uma ferramenta livre e modular, que permite o desenvolvimento em um ambiente acadêmico colaborativo e facilita evoluções. Atualmente conta com módulos de Pré-Processamento, Classificação e Agrupamento de Dados, sendo este último foco deste trabalho. A seguir, são apresentadas as implementações da YADMT para a construção do módulo de agrupamento de dados utilizando a linguagem de programação Java e os padrões de projeto já especificados.

2. O Módulo de Agrupamento

O módulo de Agrupamento de dados é atualmente constituído por sete métodos: quatro métodos hierárquicos, um método de particionamento, um método baseado em Colônia de Formigas, e um método de agrupamento a partir de Mapas Auto-Organizáveis (*Self Organizing Maps* – SOM). Os algoritmos baseados em Particionamento, também conhecidos por Métodos Não Hierárquicos, procuram pela formação de um grupo sem a necessidade da associação hierárquica. A partir de *n* objetos procura-se formar *k* grupos otimizando algum critério de particionamento (BOSCARIOLI, 2008). O método baseado em particionamento mais conhecido, e implementado no módulo, é o *k-means* (*k*-médias), descrito em (HERNÁNDEZ et. al., 2012). Também implementado neste módulo tem-se o Algoritmo de Agrupamento baseado em Formigas formulado inicialmente por Deneubourg et. al.(1991), porém, com modificações propostas por Villwock (2009).

Os métodos de agrupamento de dados hierárquicos formam grupos de forma hierárquica, admitindo vários níveis de agrupamento. Estes níveis podem ser representados por árvores, formadas durante o processo de agrupamento. De acordo com Jain e Dubes (1988) um dendrograma pode representar os níveis de agrupamento, assim como os níveis de similaridade. De acordo com Jain, Murty e Flynn (1999), a maioria dos métodos de agrupamento de dados hierárquicos é variante dos algoritmos *single-link*, *complete-link* e *minimum-variance*, métodos descritos por Johnson e Wichern (1998), bastante populares e, por isso, implementados no módulo de agrupamento de dados da YADMT.

Técnicas baseadas em Redes Neurais Artificiais (RNA) vêm se destacando quando utilizadas para a tarefa de agrupamento de dados, em particular *SOM*, também implementado na ferramenta. *SOM* é um modelo de RNA em que a aprendizagem é não supervisionada e que preserva a topologia dos dados de entrada. Esta propriedade é observada no cérebro, mas não é encontrada em outras RNA. Segundo Boscarioli (2008), a análise de dados a partir do *SOM* pode ser realizada por uma variedade de técnicas de visualização e análise de agrupamentos para a busca de padrões em uma base de dados. Para a análise do *SOM*, a ferramenta conta com os métodos de visualização Matriz-U (ULTSCH, 1992) e Matriz Densidade (ZHANG, 1993), e para a análise de agrupamentos foi implementada a metodologia de agrupamento por Matriz Densidade (XU, LI, 2002).

Também foram implementados métodos de visualização de dados, que procuram trazer uma melhor interpretação do resultado de um determinado método de agrupamento

de dados. Estes métodos utilizam-se da própria base de dados, e do grupo formado por um dado método de agrupamento, para mostrar graficamente uma representação que leve o usuário do módulo a uma melhor interpretação dos resultados obtidos.

Dentre as classes de técnicas de visualização de dados presentes na literatura, e implementadas no módulo, tem-se as técnicas de visualização de dados em 2D e 3D, que procuram representar em tela os dados conforme seus próprios atributos considerando as coordenadas *x* e *y* do plano cartesiano, proporcionando assim a dispersão dos dados.

A Figura 1 apresenta algumas telas da YADMT, da seleção dos métodos disponíveis, de execução de K-means e de visualização de dados.

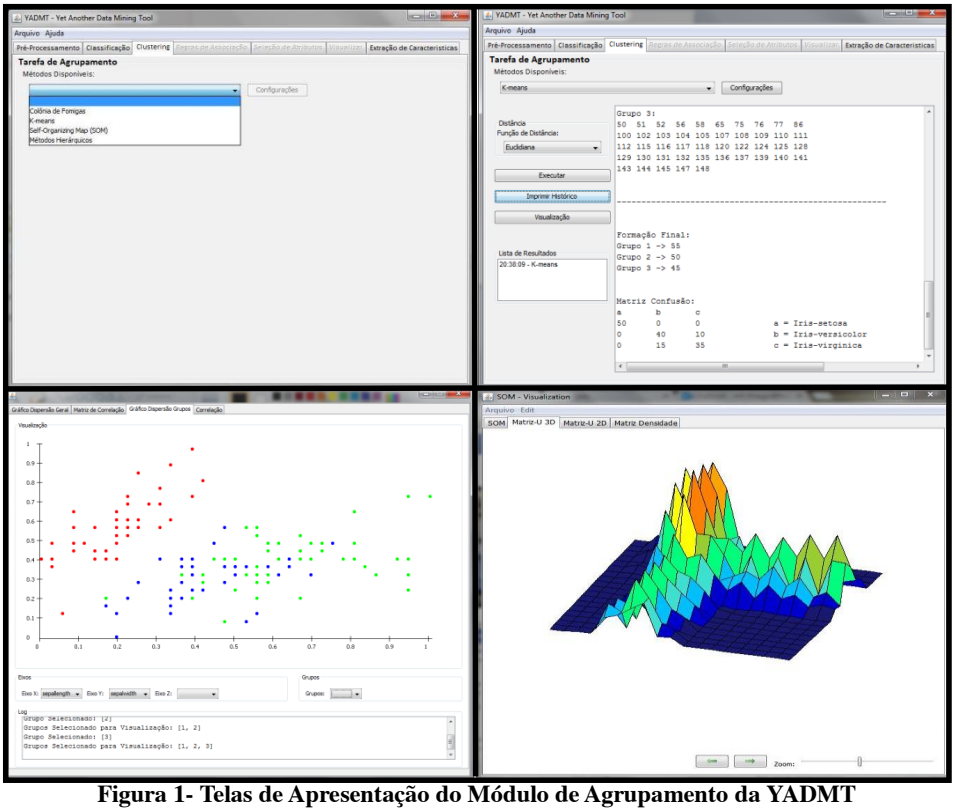

Figura 1- Telas de Apresentação do Módulo de Agrupamento da YADMT

3. Considerações Finais

O módulo de agrupamento de dados da ferramenta YADMT tem como intenção facilitar a retirada de conhecimento de uma determinada base de dados para posterior tomada de decisão. Entre os métodos de agrupamento de dados implementados no módulo estão os mais conhecidos da literatura.

Para auxiliar a interpretação dos resultados dos métodos de agrupamento de dados e, na tomada de decisão, por conseguinte, também estão disponíveis métodos de visualização de dados, cujo objetivo é facilitar o entendimento e interpretação de uma determinada base de dados e também do resultado de um método de agrupamento de dados.

Para a validação do módulo de agrupamento de dados será formada uma base de dados com dados públicos, com a intenção de aplicar os métodos implementados e avaliar sua efetividade, tanto em termos de qualidade de agrupamento quanto qualidade e efetividade dos métodos de visualização implementados. O módulo continua em desenvolvimento, da mesma forma que a ferramenta em si está em evolução.

Referências

BENFATTI, E. W.; BONIFACIO, F. N.; GIRARDELLO, A. D.; BOSCARIOLI, C. **Descrição da Arquitetura e Projeto da Ferramenta YADMT - Yet Another Data Mining Tool**. Relatório Técnico nº 01 do Curso de Ciência da Computação, UNIOESTE, Campus de Cascavel, 2010.

BOSCARIOLI, C. **Análise de agrupamentos baseada na topologia dos dados e em mapas auto-organizáveis**. 2008. Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2008.

DENEUBOURG, J.-L., GOSS, S., FRANKS, N., SENDOVA-FRANKS, A., DETRAIN, C. CHRÉTIEN, L. **The dynamics of collective sorting: Robot-like ants and ant-like robots**. In Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animals 1 (pp. 356–365). Cambridge, MA: MIT Press, 1991.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHRUSAMY, R. **Advances in knowledge Discovery & Data Mining**. California: AAAI/MIT, 1996.

HERNÁNDEZ, L.; BALADRÓN, C.; AGUIAR, J. M.; CARRO, B. SÁNCHES-ESGUEVILLAS, A. **Classification and Clustering of Electricity Demand Patterns in Industrial Parks**. 2012.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. Nova Jersey, USA: Prentice Hall, 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. **Data Clustering: A Review**. ACM Computing Surveys. v. 31, n. 3, 1999.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis**. Fourth Edition. New Jersey: Prentice Hall, 1998.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Inc. Boston, MA, USA: Addison-Wesley Longman Publishing Co. 2005.

ULTSCH, A. **Self-organizing neurak networks for visualization and classification**. *Information and Classification*, Springer-Verlag, Dortmund, Alemanha, 1992.

VILLWOCK, R. **Técnicas de Agrupamento e de Hierarquização no Contexto de Kdd – Aplicação a Dados Temporais de Instrumentação Geotécnica-Estrutural da Usina Hidrelétrica de Itaipu**. 125 f. Tese (Doutorado em Métodos Numéricos em Engenharia) – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2009.

XU, B.; LI, S., **Automatic Color Identification in Printed Fabric Images by a Fuzzy Neural Network**. AATCC Review, v. 2, n. 9, p. 42-45, 2002.

ZHANG, X.; LI, Y., **Self-Organizing Map as a new method for clustering and data analysis**. In: International Joint Conference on Neural Networks, 1993, Nagoya: Proceedings of International Joint Conference on Neural Networks – IJCNN'93, p. 2448 -2451.