Avaliação de Aspectos de Usabilidade em Ferramentas para Mineração de Dados

Clodis Boscarioli[1], José Viterbo[2], Mateus Felipe Teixeira [2]

[1]Curso de Ciência da Computação – Universidade Estadual do Oeste do Paraná (UNIOESTE) - Cascavel, Paraná, Brasil

[2]Instituto de Computação - Universidade Federal Fluminense (UFF)- Niterói, Rio de Janeiro, Brasil

clodis.boscarioli@unioeste.br, {viterbo,mateus.teixeira}@ic.uff.br

Abstract. *Currently various techniques and tools are proposed to allow the end user to interpret large volumes of data stored in organizational databases for a given decision taking. In this paper, we discuss the ways to interact with the tools of cluster analysis, focusing on both the grouping as the stages of interpretation. Investigated how the usability and user experience in such tools can improve the understanding of the discovered knowledge and thus improved decision making made on a database. We evaluated four different cluster analysis tools: Knime, Orange Canvas, Rapidminer Studio and Weka data mining tools.*

Resumo. *Atualmente várias técnicas e ferramentas são propostas para permitir que o usuário final possa interpretar grandes volumes de dados armazenados em bancos de dados organizacionais para uma determinada toma de decisão. Neste trabalho, discutimos as formas de interagir com as ferramentas de análise de cluster, tendo em conta tanto o agrupamento quanto as etapas de interpretação. Investigamos como a usabilidade e a experiência do usuário em tais ferramentas pode melhorar a compreensão do conhecimento descoberto e assim melhoras a tomada de decisão feita sobre uma base de dados. Foram avaliados quatro ferramentas de análise de agrupamento diferentes: Knime, Orange Canvas, Rapidminer Studio e Weka ferramentas de mineração de dados.*

1. Introdução

Devido ao rápido crescimento do volume de dados armazenados em bancos de dados organizacionais e às limitações humanas na análise e interpretação de dados, técnicas apropriadas são necessárias para permitir a identificação de uma grande quantidade de informação e conhecimento em tais bancos de dados. Esse processo analítico é conhecido como *Knowledge Discovery in Databases* (KDD), um processo não trivial que visa descobrir novos padrões úteis e acessíveis em bancos de dados [Fayyad, 1996], e compreende três etapas principais: o pré-processamento para a preparação de dados, a mineração de dados e o pós-processamento, que inclui a depuração e/ou síntese dos padrões descobertos.

A mineração de dados (*data mining*) é o núcleo do processo de KDD e se baseia em um conjunto de diferentes algoritmos para extrair padrões ocultos em bases de dados, que variam de acordo com o objetivo da análise, que pode ser, entre outros, a

definição de modelos de regressão, classificação ou agrupamento de dados. Agrupamento de dados, de interesse particular neste estudo, pode ser definido como a identificação de grupos ou subconjuntos de dados em que há uma coesão interna elevada entre os objetos que pertencem a um grupo, mas também um grande isolamento externo entre os grupos.

Assume-se aqui que processo de agrupamento de dados é composto não apenas pela aplicação parametrizada de algoritmos para identificação de grupos, mas também por uma segunda etapa, que consiste em realizar a interpretação dos resultados gerados pelos algoritmos de aprendizado não supervisionado, o que pode ser feito pela aplicação de métodos de visualização de dados. O principal objetivo da visualização de dados é integrar o usuário final no processo de descoberta de conhecimento, proporcionando uma representação gráfica dos dados originais, ou do resultado de agrupamento de dados.

Neste trabalho, investigamos como os aspectos de usabilidade e a experiência do usuário na interação com ferramentas de análise de agrupamento podem melhorar a compreensão do conhecimento descoberto. Para este fim, avaliamos quatro ferramentas de análise de agrupamento gratuitas e amplamente utilizadas: Knime, Orange Canvas, Rapidminer Studio e Weka.

Na próxima seção, apresentamos alguns conceitos básicos sobre o agrupamento de dados e visualização de dados. Na Seção 3, descrevemos as ferramentas selecionadas para o nosso estudo. Na Seção 4, descrevemos a metodologia de avaliação de usabilidade de adotada. Na Seção 5, discutimos os resultados observados. E, finalmente na Seção 6, apresentamos algumas conclusões e perspectivas da pesquisa.

2. Mineração e Visualização de Dados: Conceitos Introdutórios

Agrupamento de dados é uma denominação geral para métodos computacionais que analisam dados visando descoberta de conjuntos de observações homogêneas [Everitt, 2001]. Considerando uma base de dados com n padrões, cada um destes medido segundo p variáveis, o objetivo é encontrar uma relação que os agrupe estes padrões em k diferentes grupos. Com isto espera-se a visualização da relação entre os dados que anteriormente à operação do agrupamento de dados não era explícita.

Existem vários algoritmos para agrupamento de dados, que utilizam de maneiras diferentes para a identificação e a representação dos resultados do algoritmo. A escolha de cada algoritmo depende do tipo de dado que se pretende explorar, da aplicação e objetivos de análise. Neste trabalho, optamos por utilizar apenas o algoritmo *k-means* [Macqueen, 1967], por estar disponível em todas as ferramentas analisadas e ser um dos mais simples e mais utilizados para a tarefa de agrupamento.

No *k-means*, inicialmente define-se o número de grupos k a serem identificados. Em seguida, são definidos os k centroides iniciais, podendo-se utilizar diversas heurísticas. Para a etapa seguinte cada item da base de dados é associado ao centroide mais próximo, de acordo com uma medida de distância pré-estabelecida e ao final, recalculam-se os k centroides calculando a média para cada atributo entre os itens associados ao centroide analisado no momento. Este processo é repetido até que não haja mais mudança nos grupos formados.

A ideia principal da visualização de dados é integrar o usuário final ao processo, oferecendo-lhe uma representação gráfica dos dados. Técnicas de visualização são utilizadas após a mineração de dados para permitir ao usuário final interpretar os resultados de um determinado algoritmo. No caso de agrupamento de dados, por exemplo, atua identificando características de um determinado grupo, relação entre padrões e distinções entre grupos, distribuição espacial de padrões, entre outros. O usuário também poderá interagir com a representação gráfica para interpretar de uma melhor maneira o resultado gerado pelo método aplicado [Wong, 1999].

Há diferentes técnicas de visualização de dados, porém, a grande maioria é limitada pela dimensionalidade da base de dados a ser explorada. Com o passar do tempo, várias técnicas foram desenvolvidas para diferentes tipos de dados e também para diferentes dimensionalidades.

3. As Ferramentas Analisadas

As ferramentas selecionadas para avaliação neste trabalho são brevemente descritas a seguir. Para cada uma, primeiramente foram identificadas as técnicas de agrupamento de dados disponíveis e também as técnicas de visualização aplicáveis a agrupamento de dados.

O Knime [Knime, 2013] é uma ferramenta proposta para o uso na mineração de dados, estatística e outras áreas. Possui diversos métodos que possibilitam uma extração de conhecimento completa de uma determinada base de dados. O funcionamento da Knime é todo baseado na ideia de adição de nodos de métodos a um fluxo de execução. Porém a ferramenta apresenta algumas funcionalidades ainda não explicitadas, o que pode dificultar o seu uso.

Para o *k-means*, traz como resultado textual somente a formação dos centroides, apresentando os valores que cada atributo foi composto, e por quantos padrões este centroide foi constituído.

O Orange Canvas [Canvas, 2013] é uma ferramenta open source de mineração de dados com enfoque para classificação de dados, regressão de dados e mineração visual de dados, mas também possui métodos de associação de dados. O fluxo de execução da ferramenta é simples, e como a Knime apresenta a estrutura de nodos em que cada nodo adicionado ao campo de fluxo irá executar uma determinada tarefa. Também apresenta um sistema de *feedback* para cada método, que retorna ao usuário os dados de entrada e de saída de cada método.

O Rapidminer Studio [Rapidminer, 2013] é uma ferramenta proprietária, que possui uma versão de testes, de mineração de dados também voltada à estatística, banco de dados e processos de análises em dados. Tem como o principal foco disponibilizar um ambiente de trabalho totalmente gráfico, com a presença de elementos gráficos que significam uma operação em questão, por exemplo, um método de mineração de dados. A ferramenta possui um fluxo de execução intuitivo, seguindo a ideia de fluxo de execução baseada em nodos que representam um determinado processo. O retorno é textual, apresentando o número de grupos formados e por quantos padrões estes são formados, e também uma tabela de dados.

O Weka [Weka, 2013] é a uma ferramenta de aprendizado de máquina aberta usada para tarefas de mineração de dados, contém técnicas para pré-processamento de

dados, classificação, regressão, agrupamento, regras de associação e visualização. Como resultado da execução do método *k-means*, Weka apresenta o número de iteração, a soma do quadrado do erro dentro dos grupos e os centroides formados. A técnica de gráfico de dispersão da Weka não mostra a dispersão dos dados de um determinado grupo, e sim a dispersão dos dados de acordo com um de seus atributos em função do grupo que esta pertence.

4. Metodologia de Avaliação

Na avaliação das ferramentas realizou-se a inspeção por meio do Percurso Cognitivo, um método de avaliação de IHC por inspeção cujo principal objetivo é avaliar a facilidade de aprendizado de um sistema interativo, através da exploração de sua interface [Whartonm, 1994]. Além disso, testes de usabilidade foram realizados. Para este fim, foram definidos dois grupos de usuários, profissionais e estudantes da área das ciências exatas com algum contato prévio com a área de mineração de dados, que após a realização de uma atividade proposta em um cenário real, com duração de tempo necessária para cada um percorrer o percurso e as tarefas dadas, responderam um questionário contendo questões sobre o perfil do usuário e da usabilidade da ferramenta.

As seguintes tarefas foram definidas para serem executadas em todas as ferramentas:

1. Carregar arquivo de dados "Iris.arff";
2. Escolher a Tarefa de Agrupamento de Dados – método *k-means* com $k = 3$. Para os demais parâmetros foram utilizados os valores default;
3. Executar o algoritmo;
4. Visualizar os resultados.

4.1 Percurso Cognitivo

Embora os percursos para realizar as ações sejam diferentes em cada ferramenta, pode-se concluir que o usuário vai tentar atingir o resultado correto desde que ele saiba o funcionamento teórico dos métodos implementados para a realização de ajustes dos parâmetros, bem como o conteúdo da base de dados a ser analisada para verificação semântica dos resultados obtidos.

Para tal, as seguintes ações deveriam também ser realizadas:

- Identificar botões que possibilitem carregar a base de dados e especificar o caminho;
- Identificar botões que possibilitem inserir o método *k-means* e configurar os parâmetros de entrada;
- Identificar botões que possibilitem inserir métodos de visualização e interpretar resultado.

Ademais, há particularidades de interação entre elas. O usuário necessita saber em que categoria um determinado item é classificado para poder encontrá-lo nas ferramentas. No Weka, para a ação de abrir/carregar a base de dados pode acontecer

erros na identificação do componente da ferramenta. Por exemplo, em alguns pontos o componente chama-se “ARFF Reader” e noutros, é disponibilizado como “FILE”, ficando implícito que este componente também poderá abrir/carregar um arquivo de extensão ".arff". Caso não saiba a priori, o usuário deverá aprender enquanto realiza a tarefa qual é o fluxo necessário para realização das atividades de mineração de dados almejadas, em cada uma das ferramentas.

5. Análise de Resultados

Conforme ilustrado na Figura 1, pôde-se observar que o grupo de professores, num total de quatro, possuía maior experiência com as ferramentas que os oito alunos que participaram do experimento. Além disso, os gráficos mostram que a maioria dos usuários está bem familiarizada com Weka, enquanto muitos nunca usaram Knime ou Orange Canvas.

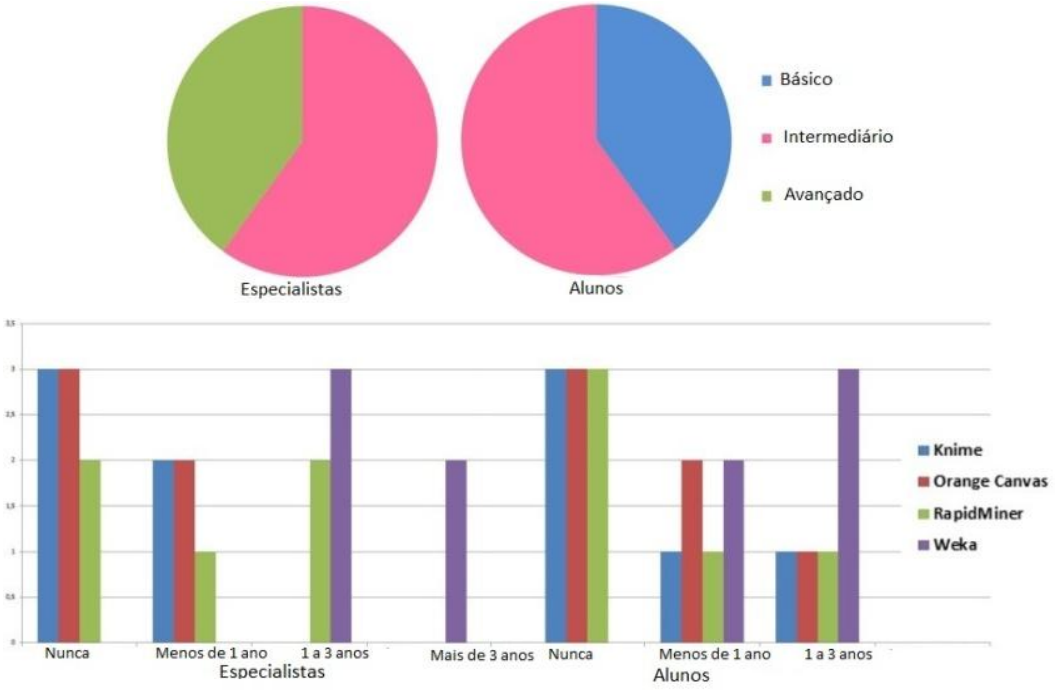

Figura 1. Nível de expertise (acima) e familiaridade com as ferramentas (abaixo) de dois grupos diferentes de usuários.

A Figura 2 apresenta os resultados da avaliação das ferramentas pelos usuários. As ferramentas Knime, Rapidminer Studio e Orange Canvas deixam o usuário interagir mesmo quando uma ação está sendo feita de forma errada, avisando-o do erro somente quando se tenta executar o fluxo do processo de mineração de dados. Este comportamento faz com que o usuário perca tempo na sua interação, podendo fazer com que o fluxo seja refeito, por não saber onde exatamente ocorreu o erro. Uma solução para isto seria o monitoramento das ações do usuário e a notificação imediata de um possível erro.

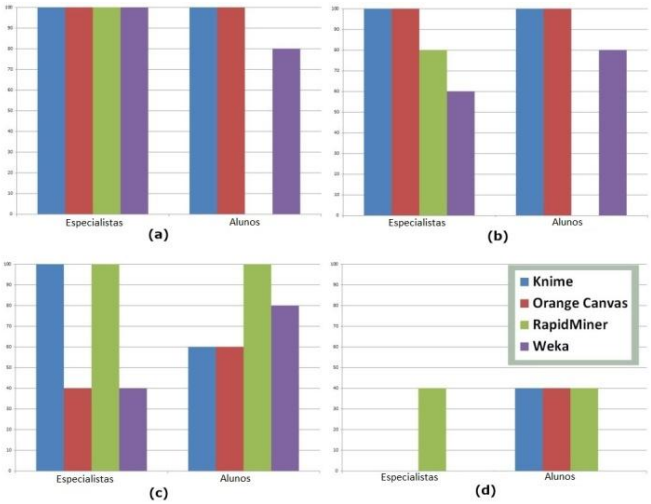

Figura 2. Proporção de usuários que avaliaram positivamente (a) a distribuição de informações, (b) os ícones e comandos na interface, (c) a descrição das tarefas disponíveis nas ferramentas e (d) a disponibilidade de mensagens de erros.

A respeito da interface, o grupo dos alunos indicou as telas carregadas de informações, já o grupo de professores não. Foi citado que Weka tem funções que não são acessadas de forma intuitiva. De acordo com os usuários, as ferramentas dispõem de informações bem distribuídas e destacadas, o que contribui para o aprendizado. Somente Rapidminer Studio, quando avaliada pelos alunos, não obteve boa avaliação. Isso se dá quando o conjunto de objetos necessário para execução de uma tarefa é grande e complexo, fazendo com que seja necessário um conhecimento prévio das suas funcionalidades, conhecimento já presente no grupo dos professores.

As ferramentas foram mal avaliadas no que diz respeito à objetividade dos ícones e botões disponíveis nas telas, ou seja, há elementos que atrapalham e/ou dificultam a escolha de uma determinada ação. Obtiveram má avalição também no que diz respeito das descrições dos elementos contidos na interface e nas opções de ajuda *online*.

Os resultados em relação à facilidade da utilização de cada ferramenta podem ser observados na Figura 3. De acordo com os estudantes, a Orange Canvas é que apresenta a interface onde é mais fácil encontrar as informações e elementos desejados e a Rapidminer Studio foi a pior avaliada. Já os professores, mostraram um resultado mais homogêneo entre as ferramentas, pois já houve um contato com uma ou mais ferramentas.

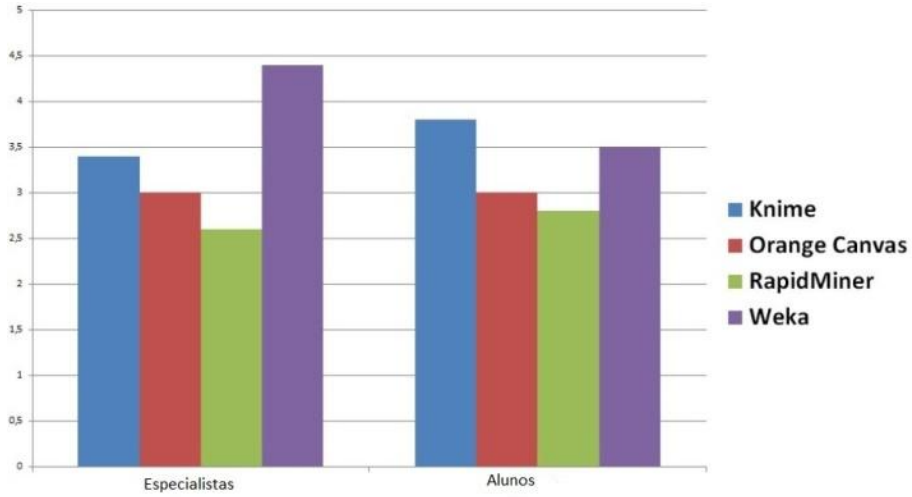

Figura 3. Avaliação dos usuários em relação a facilidade de uso de cada ferramenta.

Os resultados da avaliação da importância da visualização para interpretação dos resultados por parte dos usuários podem ser observados na Figura 4. Observa-se que os resultados foram semelhantes tanto de acordo com os alunos, quanto pelos professores. O Orange Canvas obteve a melhor avaliação, enquanto o Rapidminer Studio teve o pior resultado. As outras ferramentas tiveram desempenho regular pela avaliação realizada.

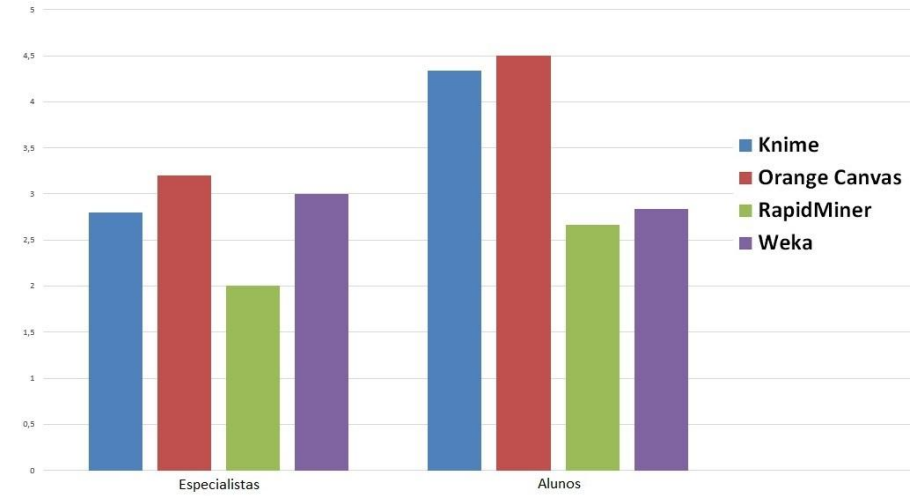

Figura 4. Avaliação dos usuários em relação à importância da visualização para interpretação dos resultados.

6. Conclusões e Trabalhos Futuros

Pôde-se perceber que, em geral, as ferramentas analisadas dão uma forte ênfase à visualização de dados. Muitas das ferramentas discutidas neste trabalho permitem ao usuário visualizar os grupos gerados pelos métodos de agrupamento, favorecendo a compreensão, interpretação e extração de conhecimento sobre o conjunto de padrões e agrupamentos.

De modo geral, a usabilidade e a facilidade de operação dos métodos de cada ferramenta são boas, e pode abranger públicos com qualquer nível de conhecimento sobre a ferramenta, ou até mesmo sobre os métodos que as ferramentas disponibilizam,

já que para a execução todas as ferramentas apresentam um fluxo de execução bastante claro, que tem um início, meio e fim.

Acreditamos que a integração interdisciplinar da KDD e IHC pode trazer significantes contribuições, tal como aumentar o potencial das ferramentas utilizando-se do conhecimento do usuário e ajudando os usuários a obter resultados cada vez mais expressivos das bases de dados através das ferramentas de mineração. Entretanto, as ferramentas analisadas mostram que outros aspectos de IHC necessitam ser abordados mais cuidadosamente.

Como trabalhos futuros tem-se a ampliação do percurso cognitivo que o usuário terá que seguir para ao término responder um novo questionário de avaliação das ferramentas selecionadas. Para o percurso serão adicionadas novas tarefas de mineração de dados, abrangendo um método de cada uma das tarefas que constituem a mineração de dados (Classificação, Associação, Agrupamento e Regressão). Com isso, tem-se a intenção de avaliar mais profundamente as ferramentas e quão experiente o usuário tem que ser para poder usar estes métodos e ao fim propor uma ontologia que disponha todos os métodos de mineração de dados de forma que melhore a interação do usuário com a ferramenta.

Referências

CANVAS. Site Oficial da Ferramenta ORANGE CANVAS. Disponível em: http://orange.biolab.si/. Acesso em 27 de outubro de 2014.

EVERITT, B. S.; LANDAU, S.; MORVEN, L. Cluster Analysis. 4a ed. Londres: Hodder Arnold Publishers, 2001.

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

KNIME. Site Oficial da Ferramenta KNIME. Disponível em: http://www.knime.org/. Acesso em 27 de outubro de 2014.

MACQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297. 1967.

RAPIDMINER. Site Oficial da Ferramenta RAPIDMINER STUDIO. Disponível em: http://rapidminer.com/products/rapidminer-studio/. Acesso em 27 de outubro de 2014.

WEKA. Waikato Environment for Knowledge Analysis. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/. Acesso em 27 de outubro de 2014.

WHARTON, C., RIEMAN, J., LEWIS, C., and POISON, P. The cognitive walkthrough method: A practitioner's guide.1994.

WONG, P. C. Visual Data Mining. IEEE Computer Graphics and Applications, Los Alamitos, v.19, no. 5, p. 20-21, 1999.