Lupa: Um Ambiente Facilitador do Desenvolvimento de Aplicações Data Mining

Stéfani Pires, Rafael Silva, Giuseppe Mongiovi

Departamento de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

{stefani.pires,rafael.silva}@sun.com, gmongiovi@uol.com.br

Abstract. *The search for information through data analysis implies the use of automated data mining techniques. There are many tools avaliable that can apply combinations of such techniques, but few have the ability to guide the user during the knowledge acquisition process. In this paper we present Lupa, a data mining environment that not only provides common KDD functionalities, but also includes means of assisting, in an "intelligent" way, the user's decision making. Assistance is provided with regards to domain modeling, algorithm selection, and in the subjective evaluation (semantics) of the results.*

Resumo. *A busca por informação através da análise de dados implicou na utilização de técnicas automatizadas de mineração de dados. Existem diversas ferramentas que integram essas técnicas, mas poucas foram projetadas buscando orientar o usuário durante o processo de aquisição de conhecimento. Neste trabalho apresentamos um ambiente de mineração de dados, denominado Lupa, que, além de possuir as facilidades dos ambientes KDD tradicionais, incorpora facilidades "inteligentes" objetivando auxiliar o usuário na modelagem do domínio, na escolha dos algoritmos mais adequados e na avaliação subjetiva (semântica) do conhecimento obtido.*

1. Introdução

A tecnologia atual nos permite capturar e armazenar uma vasta quantidade de dados. Encontrar padrões, tendências e anomalias nesses conjuntos de dados, e sumarizá-los com modelos quantitativos simples, é um dos grandes desafios da era da informação [Witten & Frank 2005]. Mineração de dados nos dias de hoje é um fato e sua aceitação e aplicabilidade se expande nas mais diversas áreas, basta que se tenha uma fonte de dados – seja um banco de dados de uma empresa, imagens, *streams* ou até mesmo a grande rede *(Internet)* – que é possível extrair a informação e o conhecimento que estão neles embutidos. O valor agregado aos dados passa a gerar uma mina muito mais valiosa de informação, que é o grande diferencial de quem utiliza as técnicas de mineração de dados. O objetivo é extrair o máximo de conhecimento relevante, pois estas informações ditam o rumo dos negócios ou das pesquisas.

A busca por informação através da análise de dados implicou na utilização de técnicas automatizadas de mineração de dados para extrair o conhecimento em áreas como *marketing*, economia, médica, telecomunicações, governamental, processamento

de imagens, etc. Extrair informação relevante não é uma tarefa trivial, é o produto final de todo um processo de extração de conhecimento denominado *Knowledge Discovery in Databases (KDD)*, onde são utilizadas técnicas específicas de acordo com o domínio em análise.

Atualmente existe uma grande quantidade de ferramentas que integram as técnicas do processo KDD visando facilitar sua utilização e atender a crescente demanda de aplicações. São inúmeros os ambientes de mineração que proporcionam integração com sistemas de banco de dados ou *data warehouses*, ferramentas sofisticadas para pré-processamento, integração e transformação, mineração, avaliação e visualização dos dados, tais como o Weka, YALE, Projeto R, Tanagra, Orange, SAS E-miner, Oracle DM, Insightful Miner e Clementine SPSS [Kdnuggets 2007], entre outros.

A convergência para a utilização de ferramentas genéricas ou específicas incentiva o estudo das mesmas e o desenvolvimento de novas ferramentas, para suprir deficiências encontradas. Embora as alternativas computacionais existentes apresentem muitas características úteis e desejáveis, poucas foram projetadas objetivando orientar o usuário ao longo de todo o processo KDD [Goldschmidt 2003]. Por exemplo, nos ambientes analisados percebe-se claramente a deficiência de facilidades, tais como:

1. Um sistema inteligente que possa auxiliar o usuário a compreender os conceitos dos dados e como extrair conhecimento dos mesmos, orientando-o na escolha dos procedimentos e dos algoritmos de pré-processamento e garimpagem mais adequados para lidar com o tipo de aplicação;
2. Obtenção, através de um especialista do domínio, de informações auxiliares e complementares, constituindo-se uma base de conhecimento preliminar (*background knowledge*).
3. Avaliação subjetiva da base de conhecimento gerada;
4. Eliciação, automática ou semi-automática, de características de objetos visando uma dada meta de aplicação.

Para lidar com a primeira deficiência, [Goldschmidt 2003] propõe uma Máquina Inteligente que possa orientar o usuário ao longo do processo KDD. [Mongiovi 1995] apresenta soluções para sanar as demais deficiências que, embora tenham sido desenvolvidas para a família dos algoritmos indutivos, suas idéias podem ser transportadas para as demais famílias de algoritmos de mineração de dados. [Sinoara 2006] propõe uma metodologia para avaliação de regras de associação. No que se refere ao agupamento, [Candillier et al 2006] propõem técnicas para avaliar a qualidade dos grupos gerados. Objetivando contornar todas essas deficiências, e baseando-se nas idéias acima mencionadas, neste trabalho apresentamos a descrição de um Ambiente de Descoberta de Conhecimento, denominado *Lupa*, que, além de possuir as facilidades dos ambientes KDD tradicionais, incorpora as idéias de uma abordagem automatizada às deficiências ressaltadas e proporciona uma alternativa de aprendizagem para o meio acadêmico no que se refere à execução de algoritmos que realizam tarefas específicas do processo de mineração de dados.

Este trabalho está organizado da seguinte forma: na seção 2 é apresentada a arquitetura do ambiente do ponto de vista conceitual, na seção 3, são relatados os resultados dos estudos e pesquisas para a minimização das deficiências apresentadas

acima. Por fim, na seção 4 apresentamos a conclusão.

2. Arquitetura do *Lupa*

O Lupa foi projetado visando facilitar o desenvolvimento de aplicações *data mining*. Assim sendo, a sua arquitetura espelha o fluxo natural de um processo KDD e se encontra dividido conceitualmente em dois grandes processos: um assistente inteligente para auxílio ao usuário e o processo KDD em si (Figura 1).

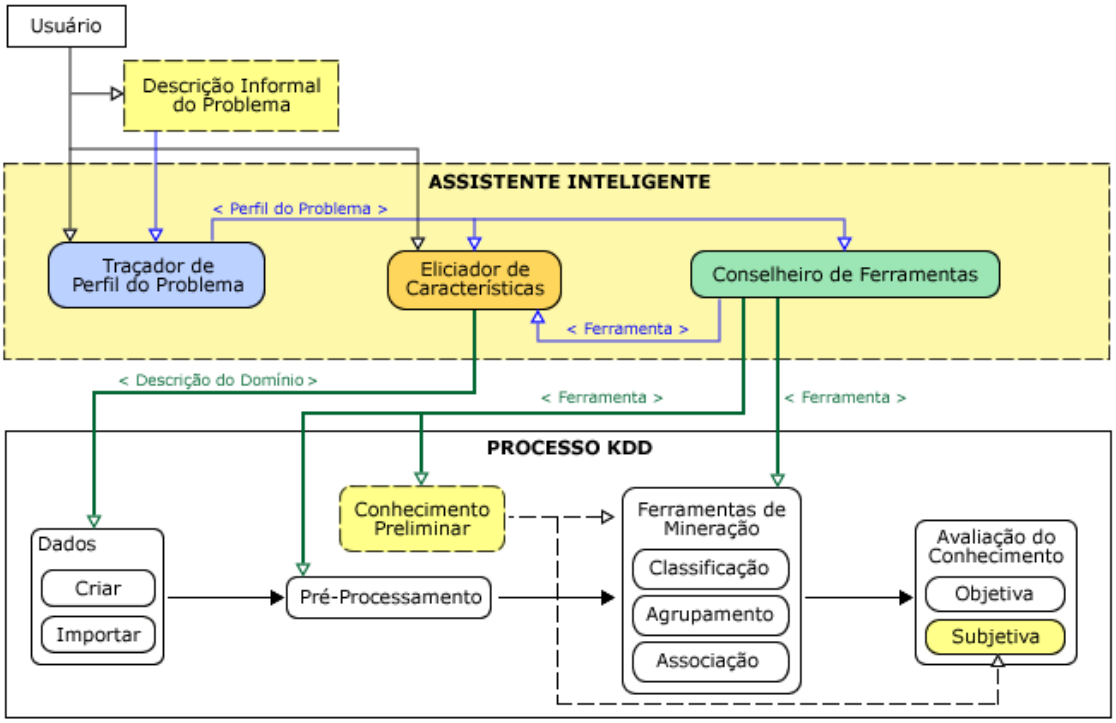

Figura 1. Arquitetura do *Lupa*

Os dois conceitos, observados na Figura 1, estão demarcados por dois grandes retângulos. Na parte superior encontra-se o sistema inteligente, onde a linha tracejada indica que todo o seu conjunto de ferramentas é opcional, as setas pretas representam a interação do usuário com o sistema, as setas azuis a troca de informações entre os módulos internos e as setas verdes representam a comunicação do assistente inteligente com o processo KDD. Nessa arquitetura, o processo KDD está representado na parte inferior, onde as linhas com setas preenchidas identificam o fluxo do processo e as linhas tracejadas, envolvendo o conhecimento preliminar, indicam que o seu uso é opcional.

As técnicas abordadas no processo KDD não diferem do proposto na literatura e por isso, neste trabalho, daremos pouca ênfase à sua descrição. O pré-processamento compreende as funções relacionadas à captação, à organização, ao tratamento e à preparação dos dados. Essa etapa possui fundamental relevância no processo de descoberta de conhecimento, compreendendo desde a correção de dados (ex.: campos obsoletos ou redundantes, falta de dados e pontos fora da curva) até o ajuste da formatação destes para os algoritmos de mineração a serem utilizados [Goldschmidt & Passos 2005]. A mineração de dados é uma tarefa do processo KDD que busca obter informações a partir de uma base de dados, extraindo informações (padrões) não óbvios,

mas úteis. Técnicas de mineração de dados podem ser divididas em duas formas – diretas e indiretas. Mineração direta tenta explicar ou classificar algum atributo alvo em particular. Mineração indireta busca encontrar padrões e similaridades entre grupos de registros sem o uso de um atributo alvo ou classes predefinidas [Berry & Linoff 2004]. O *Lupa* utiliza como mineração direta técnicas de Classificação (ex.: árvores de decisão, redes neurais e regressão linear) e Regras de Associação e como mineração indireta técnicas de Agrupamento (completo ou difuso).

O Assistente Inteligente é uma proposta bastante abrangente de uma forma de auxílio ao usuário à utilização adequada das técnicas do processo KDD. Um auxílio que cobre desde antes da importação dos dados no ambiente, possibilitando ao usuário definir exatamente que tipo de problema ele tem e que tipo de soluções deseja encontrar, para que, em seqüência, o sistema possa guiar seus passos no processo, sugerindo os melhores dados a serem usados e as técnicas mais adequadas a serem aplicadas.

3. Descrição Conceitual das Facilidades de Automação

Como mencionado na seção introdutória, o *Lupa* incorpora algumas idéias objetivando sanar, de forma automática, as principais deficiências dos ambientes KDD tradicionais. Para isso, o *Lupa* possui quatro facilidades de automação: seleção inteligente de algoritmos, eliciação de características de objetos, eliciação de conhecimento preliminar e avaliação subjetiva de conhecimento.

3.1. Seleção Inteligente de Algoritmos

Análises incorretas realizadas em dados que não foram pré-processados corretamente podem conduzir a conclusões errôneas, ou análises impróprias podem ser aplicadas a conjuntos de dados que levem a abordagens completamente diferentes. Caso estes erros perpetuem até o processo da análise, eles podem conduzir a falhas muito dispendiosas [Larose 2005]. Portanto, a escolha e execução dessas tarefas não são triviais, onde, geralmente, necessita-se de um especialista para tomar decisões – tais como a modelagem do domínio, escolha da algoritmos de preparação e garimpagem dos dados – em função do objetivo proposto.

Visando minimizar esses riscos e proporcionar uma maior confiança na escolha dos algoritmos de mineração e pré-processamento adequados, está sendo desenvolvido no *Lupa* um sistema inteligente de seleção de algoritmos, denominado Conselheiro de Ferramentas (CF). A base de conhecimento utilizada pelo sistema será adquirida através de uma análise crítica das vantagens e desvantagens no uso de ferramentas e técnicas disponíveis. A idéia geral é extrair de um ou mais especialistas em mineração de dados, toda e qualquer informação relevante para a escolha ou não de procedimentos, incluindo sua seqüência no processo. Por exemplo, saber identificar em qual situação uma rede neural deve ser utilizada e onde ela não seria uma boa solução; em se tratando de um problema de classificação, poderíamos escolher entre o paradigma indutivo ou o conexionista.

O conhecimento é hierarquicamente estruturado e representado na forma de regras de produção, com fatores de certeza associados à regra e às condições. As condições refletem as características de uma dada aplicação e a conclusão representa a seqüência mais adequada de tarefas para essa aplicação, ou simplesmente uma técnica específica. Por exemplo, para uma aplicação típica de fixação de clientes tem-se:

Se *tipo_aplicação = 'fixação_cliente'* ***&*** *população = 'heterogênea'* ***então***
seqüência_tarefa = 'classificação' ***&*** *'segmentação' (FC = 1.0)*

Se *tipo_tarefa = 'classificação'* ***&*** *tipo_dados = 'nominal'* ***&***
precisão_dados = 'irrelevante' ***então*** *técnica = 'indução_árvore_decisão' (FC = 0.8)*

A entrada do CF é a descrição do perfil do problema, fornecida pelo Traçador de Perfil. Esse traçador obtém do usuário, através de uma entrevista objetiva, as informações necessárias para definir uma descrição formal do problema. A entrevista é baseada em um conjunto pré-definido de parâmetros com o intuito de extrair do usuário aqueles parâmetros que realmente definem o problema. Por exemplo, trata-se de um problema de otimização, classificação ou correlação? A precisão da solução é relevante? As perguntas tentam identificar os parâmetros que classificam um tipo de problema na base de conhecimento do sistema inteligente, para que, por exemplo, por técnicas de analogia, o conselheiro de ferramentas possa filtrar nessa base o conjunto de parâmetros semelhantes e os algoritmos utilizados nos exemplos identificados. Uma vez identificado o conjunto de algoritmos de preparação e garimpagem, com suas respectivas relevâncias, que poderão ser utilizados para a abordagem do domínio, o usuário pode optar pelo seu uso.

3.2. Eliciação de Características de Objetos

A escolha do conjunto de variáveis que definem um domínio aparentemente parece ser uma tarefa trivial. Entretanto, a prática tem mostrado que essa escolha é um problema de uma certa complexidade, visto que, para um dado domínio, além de identificar o conjunto de características, é necessário observar cuidadosamente os aspectos de completude, relevância e de redundância.

Para auxiliar o usuário na modelagem do domínio, o *Lupa* dispõe de uma ferramenta semi-automática de eliciação de características. Essa ferramenta é baseada na Teoria das Construções Pessoais [Kelly 1955] e segue a filosofia dos sistemas de aquisição semi-automática de conhecimento, em particular do ETS e do AQUINAS [Boose 1990]. Esses sistemas, através de entrevistas curtas e objetivas, conseguem extrair do especialista as características mais representativas do domínio, eliminando assim os problemas de irrelevância e redundância.

3.3. Eliciação de Conhecimento Preliminar

O conhecimento preliminar é o conhecimento disponível sobre o domínio antes do uso de um algoritmo generalizador e tem um caráter de complementariedade ao conjunto de exemplos, podendo contribuir de duas formas na aprendizagem: melhorando a qualidade e utilidade do conhecimento gerado e facilitando o trabalho dos algoritmos de aprendizagem, através da diminuição do seu espaço de busca [Mongiovi 1995].

O *Lupa* apresenta três formas diferentes de conhecimento preliminar: relevância semântica, generalização e custo dos atributos. A relevância semântica tem por objetivo fornecer subsídios aos algoritmos indutivos para produzir bases de conhecimento de melhor qualidade e utilidade, pois, geralmente, os algoritmos indutivos são guiados apenas pelo conjunto de treinamento, ou seja, apenas por informações estatísticas [Mongiovi 1995]. O ambiente permite definir uma matriz de relevância que representa o conhecimento a respeito do relacionamento semântico entre as condições e as classes. Pode-se definir também, uma matriz de relevância nebulosa que possui o grau de

pertinência entre os elementos de classificação e os atributos do domínio.

Os atributos de um domínio podem ser generalizados com o propósito de facilitar o aprendizado a partir de exemplos [Núnez 1991]. A generalização consiste na substituição de dados por conceitos de mais alto nível, ela pode conter valores que não estão presentes na base de dados, permitindo assim que seja induzido um conhecimento ainda não observado.

O custo refere-se ao valor genérico de um atributo. Ele deve ser interpretado no mais alto nível de abstração. Custo pode ser mensurado em diversas unidades, como monetária, temporal ou unidades abstratas de utilidade [Turney 2000]. Para os algoritmos de agrupamento, o custo pode servir para medir a importância relativa das variáveis que definem o dominio.

3.4. Avaliação Subjetiva de Conhecimento

O produto final de um processo KDD é um conhecimento que, para tornar-se mais confiável e útil, deve ser avaliado por medidas que possam garantir seu grau de aplicabilidade. Essas medidas procuram quantificar a complexidade do conhecimento obtido e podem assumir aspectos objetivos e subjetivos, sendo que as medidas objetivas são exclusivamente dependentes da estrutura dos padrões e dos dados utilizados e, por outro lado, as medidas subjetivas são fundamentalmente dependentes do interesse e/ou necessidade dos usuários que irão utilizar o conhecimento extraído [Silberschatz & Tuzhilin, 1996].

As técnicas objetivas procuram avaliar o conhecimento sob o aspecto numérico, com medidas de fácil mensuração, tais como a acurácia, que utiliza um conjunto de teste (geralmente uma parcela da base de dados) para validar os conceitos obtidos. Em técnicas de mineração como associação e classificação podemos medir o número e tamanho médio das regras, que podem ser obtidos pelo número médio de condições por regra, facilitando uma compreensão superficial (aspecto sintático) da base de conhecimento gerada.

Do ponto de vista semântico, os aspectos objetivos não são suficientes para determinar os fatores de compreensibilidade e credibilidade de uma base de conhecimento. A compreensibilidade mede o grau de entendimento de uma informação extraída; em seu aspecto subjetivo, tenta mensurar quanto um conceito é mais compreensível que outro. O grau de relevância de um conhecimento está diretamente relacionado com sua credibilidade, pois se o mesmo apresentar um baixo valor, o especialista poderá desconsiderar e descartá-lo, tornando-se, conseqüentemente, um conhecimento inútil. A identificação dessas medidas é o primeiro passo para se tentar automatizar esses fatores.

Nos algoritmos do paradigma indutivo a compreensibilidade e a credibilidade podem ser medidas de forma automática observando-se o aparecimento de condições relevantes nas regras encontradas. A matriz de relevância como forma específica de representação de conhecimento preliminar é uma ferramenta simples e objetiva, podendo ser utilizada para valorar a importância da condição de uma regra, e consequentemente da premissa, para concluir o elemento de classificação presente na regra, conseguindo-se, assim, um grau de relevância para a regra [Mongiovi 1995].

Quanto às técnicas de agrupamento ou clusterização, [Candillier et al 2006]

afirmam que a avaliação dos resultados de agrupamento são de natureza subjetiva e que na prática existem quatro técnicas usadas para medir a qualidade dos *clusters* – uso de base de dados artificiais, uso de base de dados rotuladas, trabalhar com um especialista ou utilizar algum critério interno, como a separação *intercluster* – mas cada técnica possui suas limitações. Portanto, eles propõem que a técnica de agrupamento seja utilizada como um passo do pré-processamento de outra tarefa que seja capaz de ser avaliada (ex.: aprendizagem supervisionada). Inicialmente é feita uma partição do conjunto de dados separando-os em *clusters* e, posteriormente é aplicada a cada grupo uma técnica supervisionada para comparar os resultados providos de informações vindas do primeiro passo. Foi observado nos resultados que receberam uma informação prévia, um aumento significativo no grau de qualidade do conhecimento gerado.

Como mencionado na seção 1, [Sinoara 2006] propõe uma metodologia para identificação de regras de associação interessantes que combina o uso de medidas objetivas e subjetivas. No início do processo são utilizadas as medidas objetivas, filtrando o conjunto de regras para facilitar a participação do especialista no fornecimento de seu conhecimento e seus interesses para o cálculo das medidas subjetivas – conformidade, antecedente inesperado, consequente inesperado e antecedente e consequente inesperados – propostas por [Liu et al 2000] para avaliação de regras de associação.

Muitos dos algoritmos de mineração de dados produzem uma enorme quantidade de padrões, sendo que poucos são realmente interessantes ao usuário [Padmanabhan & Tuzhilin 2000]. A necessidade da avaliação do aspecto semântico no conhecimento obtido impulsiona o desenvolvimento de um módulo de pós-processamento em análise subjetiva no *Lupa*, abrangendo as técnicas apresentadas.

4. Conclusão

O uso preciso de técnicas do processo KDD permite a extração de conhecimento interessante, inesperado e útil, porém, não é uma tarefa trivial, assim sendo, foram apresentadas propostas para facilitar a sua utilização como a incorporação de um Assistente Inteligente – uma ferramenta que destina-se a orientar o usuário durante todo o planejamento e execução de uma dada aplicação auxíliando na definição do perfil do problema, eliciação de características e, como observado por [Goldschmidt 2003], a escolha das ferramentas de mineração e pré-processamento mais apropriadas para um determinado problema. Foram apresentadas também, as abordagens adotadas para a extração de conhecimento mais relevante e útil, e abordagens para uma validação semântica do conhecimento obtido com as técnicas de garimpagem.

Neste trabalho apresentamos o *Lupa,* um ambiente de mineração de dados de propósito geral, que visa compreender além das técnicas do processo KDD, as abordagens descritas acima. Estão finalizados o módulo de conhecimento preliminar, algoritmos de classificação, associação e agrupamento. Os módulos para a construção do Assistente Inteligente e técnicas de avaliação subjetiva encontram-se em desenvolvimento.

Referências

Berry, M. J. A. e Linoff, G. Data Mining (2004) Techniques: for marketing, sales, and customer relationship management. 2nd ed, Indianapolis, Indiana.

Boose, J. (1990) Knowledge acquisition for Knowledge-Based Systems, IOS press, Tokyo.

Candillier, L. Tellier, I. Torre, F. e Bousquet, O. (2006) Cascade Evaluation of Clustering Algorithms. In *17th European Conference on Machine Learning* ECML'2006, pages 574-581, Berlin, Germany.

Goldschmidt, R. (2003) Assistência Inteligente à orientação do processo de descoberta de conhecimento em bases de dados, Tese de Doutorado, PUC-RJ, Rio de Janeiro, RJ.

Goldschmidt, R. e Passos, E. (2005) Data Mining: Um Guia Prático. Editora Campus, São Paulo, SP.

Kdnuggets (2007) "Software: Suites for Data Mining and Knowledge Discovery", http://www.kdnuggets.com/software/suites.html.

Kelly, G. (1955) The Psichology of Personal Constructs. New York-Norton.

Larose, D. T. (2005) Discovering knowledge in data: an introduction to data mining, Hoboken, New Jersey.

Liu, B. Hsu, W. Chen, S. e Ma, Y. (2000) Analyzing the subjective interestingness of association rules. In *IEEE Intelligent Systems & their Applications*.

Mongiovi, G. (1995) Uso de relevância semântica na melhoria da qualidade dos resultados gerados pelos métodos indutivos de aquisição de conhecimento a partir de exemplos, Tese de Doutorado, Centro de Ciências e Tecnologia, UFPB, Campina Grande, PB.

Núnez, M. (1991) The Use of Background Knowledge in Decision Tree Induction. Kluwer Academics Publishers, Boston.

Padmanabhan, B. e Tuzhilin A. (2000) Small is beautiful: Discovering the minimal set of unexpected patterns. In *Proceedings of the International Conference in Knowledge Discover and Data Mining* (KDD 2000), pages 54-63.

Silberschatz, A. e Tuzhilin A. (1996) What makes patterns interesting in knowledge discovery systems. In *IEEE Transactions on Kwnoledge and Data Engineering*.

Sionara, R. A. (2006) Identificação de regras de associação interessantes por meio de análises com medidas objetivas e subjetivas, Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, SP.

Witten, I. H. e Frank, E. (2005) Data Mining: pratical machine learning tools and techniques. Morgan Kaufmann Publishers, San Francisco, CA.