Aquisição de Conhecimento Durante a Mineração de Dados: Um Estudo em Tarefas de Classificação Usando uma Abordagem Interativa

Daniela Cruzes, Manoel Mendonça e Christiane Santana

Universidade Salvador, Nuperc, Salvador, Brasil, 40171-100
{daniela, mgmn, christiane.santana}@unifacs.br)

Abstract. This paper describes an approach for interactive construction of decision trees using Treemaps and exploring the knowledge discovered and used during a classification session. The approach is user-centered. It combines the strengths of the user and the computer to build better decision trees. The user provides domain knowledge and evaluates intermediate results of the algorithm, registering interesting information. The computer automatically creates patterns satisfying user constraints and generates appropriate visualizations of the produced tree.

Keywords: Visual Data Mining, Decision Tree, Classification.

1 - Introdução

A classificação é uma das tarefas mais comuns na mineração de dados, e a construção de árvores de decisão é um dos métodos mais populares da classificação. As árvores de decisão são intuitivas, fácil de interpretar, relativamente rápidas para construir e tem acurácia igual ou superior a outros métodos de classificação [4].

A realização de uma tarefa de classificação usando um processo interativo é fortemente dependente do conhecimento do minerador e da forma como o processo é conduzido. Este conhecimento estar relacionado ao domínio dos dados, à forma de coleta dos dados, aos objetivos de coleta e da análise dos dados, aos algoritmos de mineração usados, ou à conhecimentos adquiridos em outras sessões de classificação, sobre o processo de classificação, etc.

O conhecimento criado e utilizado durante a condução do processo constitui um recurso que deve ser administrado eficientemente. Este processo é difícil de ser realizado, pois os mineradores têm dificuldades em exteriorizar seu conhecimento devido ao pouco ou nenhum tempo dedicado à reflexão sobre os problemas ocorridos e decisões tomadas durante a realização das atividades do processo.

Propomos neste artigo, um processo sistemático de obtenção de conhecimento para apoiar a aquisição, de conhecimento tácito e explícito de mineradores em uma atividade de classificação.

2 – Aquisição de Conhecimento

De acordo com Fayyad [3] o KDD, pode ser definido como um processo não trivial para, a partir de dados, extrair padrões válidos, compreensíveis, potencialmente úteis e previamente desconhecidos. É composto por diversas etapas: seleção, pré-processamento, mineração, e assimilação, que podem ser realizadas em vários ciclos de execução. Cada uma das etapas compartilha os resultados com os demais passos e pode ser repetida sempre que o analista de dados achar necessário, objetivando o refinamento do conhecimento descoberto. Este processo é altamente iterativo e interativo, dependendo da constante interferência do especialista, em cada etapa, para analisar as informações geradas e incorporar sua experiência ao modelo, buscando aprimorar a qualidade dos resultados obtidos.

Em cada uma das atividades do processo de classificação existem aplicações específicas para o uso do conhecimento prévio do minerador. Assim, podemos considerar como relevante alguns tipos de conhecimento que facilitam ou melhoram a execução das atividades:

- **Conhecimento sobre Domínio dos Dados** – conhecimento do minerador sobre o domínio dos dados que estão sendo explorados e sobre os quais deseja-se criar um modelo de classificação.
- **Conhecimento sobre Exploração de Dados** - conhecimento do minerador sobre mecanismos, ferramentas e técnicas de exploração de dados.
- **Conhecimento sobre Tarefa de Classificação:** conhecimento do minerador sobre algoritmos de classificação e mais especificamente sobre o algoritmo usado no processo de classificação
- **Conhecimento sobre Processo de Coleta dos Dados:** conhecimento do minerador sobre mecanismos utilizados para a coleta dos dados em questão, objetivos da coleta, usos prévios dos dados, informações sobre qualidade dos dados, etc.

Para apoiar o processo de aquisição do conhecimento, uma infra-estrutura deve ser desenvolvida permitindo que diversos tipos de conhecimento sejam capturados a partir de múltiplas fontes nas diversas etapas do processo.

3 - O Processo de Classificação

Existem diversas abordagens para a classificação[7], propomos a abordagem descrita na figura 2 que é dividida em quatro etapas principais: (1) Exploração dos dados (2) Preparação; (3) Classificação; (4) Avaliação dos resultados obtidos.

Para acompanhar a atividade, devem ser criadas sessões de trabalho de forma a documentar a execução da atividade. Em cada sessão o minerador sempre precisa anotar observações, decisões tomadas, conclusões, idéias, dúvidas, lições aprendidas, e tudo mais que se tornar necessário documentar. Ao finalizar a sessão, o usuário pode, de acordo com a sua necessidade, armazenar ou descartar o experimento e todas as anotações feitas durante o processo.

A etapa de preparação tem como objetivo promover modificações na representação dos dados de forma a adequá-los às exigências do algoritmo e estruturá-los de

maneira mais apropriada ou relevante para a mineração. As transformações necessárias dependem do algoritmo que será utilizado na fase de mineração [2] e do conhecimento prévio do especialista pois todo o seu conhecimento será usado para a tomada de decisões importantes na transformação dos dados.

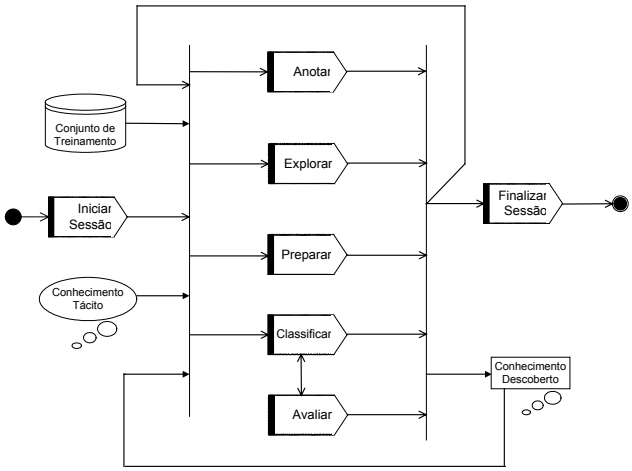

Figura 2: Processo de Classificação Interativa

A exploração dos dados tem como objetivo auxiliar o perito a se familiarizar com os dados viabilizando assim, a detecção de subconjuntos interessantes para serem submetidos ao processo de mineração, e de padrões encontrados nos dados. Nesta fase, o conhecimento de domínio do minerador é usado na exploração e ao mesmo tempo é alimentado com o conhecimento descoberto através dessa exploração. A experiência do minerador em exploração de dados pode ser um diferencial nesta fase, fazendo com que ele possa descobrir mais informações úteis para a condução do processo.

A visualização de informação pode ser utilizada nesta fase, pois visa auxiliar o processo de análise e compreensão de um conjunto de dados, através de representações gráficas manipuláveis. As técnicas de visualização de informações procuram representar graficamente dados de um determinado domínio de aplicação de modo que a representação visual gerada explore a capacidade de percepção do homem, e este a partir das relações espaciais exibidas, interprete e compreenda as informações apresentadas e, finalmente deduza novos conhecimentos.

A classificação é o núcleo do processo de descoberta de conhecimento em bases de dados; é nesta fase que ocorre a extração propriamente dita de regularidades e padrões contidos nos dados submetidos através da aplicação de algoritmos específicos capazes de extrair eficientemente conhecimento implícito e útil dos repositórios [4]. Os algoritmos podem ser totalmente automáticos ou semi-automáticos [1]. Nas duas abordagens, a escolha do algoritmo depende fundamentalmente do objetivo e das metas a serem atingidas, pois cada algoritmo possui suas particularidades e sua aplicação depende da tarefa que se deseja realizar. É importante que o minerador utilize todos os conhecimentos obtidos durantes as fases anteriores em relação ao domínio dos dados, a tarefa de classificação e sobre o processo de coleta dos dados de

forma a potencializar o uso destes conhecimentos na construção de modelos mais efetivos.

A avaliação do modelo estima a adequação de um padrão em particular (um modelo e seus parâmetros) em relação aos critérios de um processo de descoberta de conhecimento. Tanto os critérios lógicos quanto estatísticos podem ser usados na avaliação. O minerador necessita avaliar além do modelo todas as informações registradas durante todo o processo para que possa avaliar o modelo e determinar o momento certo de terminar o processo de classificação, este é o momento de empacotar e assimilar todo o conhecimento adquirido e gerado durante todo o processo de classificação.

4 – Conclusão

Propomos neste artigo, um processo sistemático de obtenção de conhecimento para apoiar a aquisição, e empacotamento de conhecimento tácito e explícito de mineradores em uma atividade de classificação. Este processo permite a exteriorização de conhecimento utilizado e assimilado durante uma sessão de classificação, mas buscando minimizar o desvio do fluxo normal de trabalho dos executantes do processo, evitando atrasos e falhas na execução de suas atividades.

Para a validação e uso do processo de classificação descrito anteriormente, uma ferramenta de mineração visual de dados foi implementada na Universidade Salvador [5]. A ferramenta possibilita a construção interativa de árvores de classificação e oferece recursos da mineração visual de dados para a criação de mecanismos eficientes de interação do usuário com o sistema durante a construção da árvore de decisão e da exploração dos dados submetidos à mineração.

Além de apoiar a aquisição de conhecimento, iremos estender o processo para apoiar também a consulta e manutenção desta base, além da criação de uma base de conhecimento relevante para a comunidade, para garantir que conhecimentos úteis sejam mantidos no repositório de conhecimento e que outros mineradores possam ter acesso a esta base.

Referências

1. Ankerst, M., Ester, M., Kriegel, H.-P.: Towards an Effective Cooperation of the User and the Computer for Classification. In ACM SIGKDD 6th Int. Conf. On Knowledge Discovery and Data Mining (KDD 2000), Boston, MA, pp. 179-188 (2000).
2. Batista, G.E.A.P.A. Pré-Processamento de Dados em Aprendizado de Máquina Supervisionado. Tese de doutorado apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP (2003)
3. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. In AI Maganize, pp. 37-54 (1996).
4. Mendonça Neto, M.; Sunderhaft, N. A State of the Art Report: Mining Software Engineering Data. State of the Art Technical Report DACS-SOAR-99-3. U.S. Department of Defense (DoD) Data & Analysis Center for Software, Rome, NY, 1999. Also available in: http://www.dacs.dtic.mil/techs/datamining/datamining.pdf
5. Mendonça Neto, M.; Cruzes, D; Santana, C. Interactive Construction of Classification Trees Using Treemaps. CLEI2004 to Appear.