Uma proposta de aplicação comercial utilizando KDD através da linguagem R

Luanna Viana Araújo[1], Rodrigo Prazim[1], Olavo Nylander Brito Neto[2]

[1]Centro Universitário do Pará (CESUPA) – Área de Ciências Exatas e Tecnologia (ACET) – 66.060- 230, Belém – Pará – Brasil

[2]Programa de Pós-Graduação em Ciência da Computação – Instituto de Ciências Exatas e Naturais - Universidade Federal do Pará (UFPA)
Caixa Postal 479 – 66.075-110 – Belém – PA – Brasil

Email: luanna.ar@gmail.com, rodrigoprazim@hotmail.com, onbn@ufpa.br

Abstract. *This paper describes a proposal through the KDD process to a company providing collection services, to aid in decision making by identifying customer profiles.*

1. Introdução

O processo *Knowledge Discovery in Databases* - KDD (descoberta de conhecimento em bases de dados) segundo Thomé (2002), foi cunhado em 1989 com o objetivo de representar todo o processo de busca e extração de conhecimento que, em seu nível mais operacional, inclui a aplicação de técnicas e algoritmos de *data mining* (mineração de dados) para manipular e encontrar indícios de correlação ou de implicação em grandes volumes de dados. Utilizando este processo com foco na indústria com o intuito de obter vantagem competitiva, assim como relatam Boente (2007) e Dantas (2008).

Goldschmidt e Passos (2005), apresenta três etapas do processo *Knowledge Discovery in Databases* (KDD) , cuja as denominações são: (1) pré-processamento que compreende as funções que se relacionam a captação, à organização e ao tratamento de dados, cujo objetivo é preparar os dados para os algoritmos da etapa seguinte; (2) *Data Mining* que realiza a busca efetiva de conhecimentos úteis no contexto proposto para a aplicação do KDD; (3) pós-processamento que abrange o tratamento do conhecimento obtido na etapa anterior. O objetivo geral desta pesquisa é identificar os perfis de clientes inadimplentes e adimplentes, com o intuito de auxiliar nas tomadas de decisão no departamento de gerência na empresa abordada. A linguagem R foi utilizada para a aplicação da técnica de KDD através dos pacotes (nnet para RNA, rpart para árvores e RandomForest para classificação), pois Segundo Torgo(2006), a linguagem R fornece um ambiente para computação estatística e gráfica. Trata-se de uma linguagem de programação especializada em computação com dados, e através desta, existe a possibilidade de verificar os diferentes resultados que cada pacote gera.

2. Aplicação na Base de dados da empresa analisada

Com base nas etapas do processo KDD, pretende-se aplicar cada fase do processo em uma tabela específica da base de dados utilizada. Os seguintes atributos compõe esta tabela: *empresa_cobradora, ndg, cpf_cnpj, nome, data_nascimento, email,*

logradouro, numero, complemento, cep, bairro, cidade, estado, cod_carteira, nome_lider, nome_operador, tipo_pessoa.

A partir destas informações houve a seleção dos seguintes atributos mais relevantes que foram transformados e padronizados como: data_nascimento (dia/mês/ano), cep (logradouro, cidade, estado), cod_carteira (lotes de empresas com débitos classificados por tempo de dívidas), tipo_pessoa (1 para pessoa física e 2 para pessoa jurídica). Estes atributos passaram por um pré-processamento onde ocorre a filtragem dos dados retirando valores inúteis como nulos (*null*) e alguns caracteres como (#,?,@, entre outros). Com o resultado desta filtragem chegou-se aos campos: cod_carteira, data_nascimento, cep, tipo_pessoa. Alguns destes atributos da tabela passaram por um processo de transformação como cod_carteira que passou a utilizar apenas números, data_nascimento representado somente por ano, o cep representado somente pelos quatro primeiros dígitos (por tratarem respectivamente de região, sub-região, setor e subsetor) e tipo_pessoa identificado pela numeração (1 e 2). Os dados do código de carteira (lotes), do ano de nascimento, cep e o tipo de pessoa (pessoa física ou jurídica) que sofreram adaptações são de suma importância para a identificação dos perfis que objetivam filtrar pessoas ou empresas inadimplentes e não inadimplentes . Na Figura 1 é apresentado uma parte da base de dados atualizada com os quatro atributos.

data_nascimento	cep	cod_carteira	tipo_pessoa
1979	7820	119	1
1972	0750	112	1
1961	0997	112	1

Figura 1. Parte dos dados filtrados e padronizados extraídos do banco de dados

3. Considerações Finais

Após a etapa de pré-processamento é executada a mineração de dados. Onde ocorre a descobertas de conhecimentos a partir dos atributos selecionados, o método a ser utilizado é o de classificação, onde será empregado as técnicas de árvore de decisão e redes neurais, com o objetivo de identificar os perfis dos clientes das empresas cadastradas. Após esta etapa será executado o pós-processamento onde pretende-se, interpretar o conhecimento obtido e comparar a partir dos perfis encontrados com o modo atual de tomada de decisão na empresa.

Referências

Boente, A. N. P., Oliveira, F. S. G., Rosa, J. L. A. (2007) "Utilização de Ferramentas de KDD para Integração de Aprendizagem e Tecnologia em Busca da Gestão Estratégica do Conhecimento na Empresa", Seget, http://www.aedb.br/seget/artigos07/1219_Artigo%20SEGET%202007.pdf, Novembro.

Dantas, E. G. R., Patrício Júnior, J. C. A., Lima, D. S.;, Azevedo, R. R. (2008) "O Uso da Descoberta de Conhecimento em Base de Dados para Apoiar a Tomada de Decisões", http://www.aedb.br/seget/artigos08/331_331_Artigo_SEGET_EJDR_Versao_Final_010808.pdf, Outubro.

Goldschmidt, R., Passos, E. GOLDSCHMIDT, R.R.; PASSOS, E. Data Mining: Um Guia Prático. Rio de Janeiro: Campus, 2005.

Thomé, A. C. G. (2002) "Redes Neurais – Uma Ferramenta para KDD e DATA MINING", Material Didático http://equipe.nce.ufrj.br/thome/grad/nn/mat_didatico/apostila_kdd_mbi.pdf, Outubro.

Torgo, L. (2006) "Introdução à programação em R", Universidade do Porto,

http://cran.r-project.org/doc/contrib/Torgo-ProgrammingIntro.pdf, Novembro.