# Pré-processamento de Dados em Aprendizado de Máquina Supervisionado

*Gustavo Enrique de Almeida Prado Alves Batista*

# Pré-processamento de Dados em Aprendizado de Máquina Supervisionado

*Gustavo Enrique de Almeida Prado Alves Batista*

**Orientadora:** *Profa. Dra. Maria Carolina Monard*

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação — ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências — Ciências de Computação e Matemática Computacional.

**USP - São Carlos**
**Março/2003**

# Resumo

A qualidade de dados é uma das principais preocupações em Aprendizado de Máquina — AM — cujos algoritmos são freqüentemente utilizados para extrair conhecimento durante a fase de Mineração de Dados — MD — da nova área de pesquisa chamada Descoberta de Conhecimento de Bancos de Dados. Uma vez que a maioria dos algoritmos de aprendizado induz conhecimento estritamente a partir de dados, a qualidade do conhecimento extraído é amplamente determinada pela qualidade dos dados de entrada.

Diversos aspectos podem influenciar no desempenho de um sistema de aprendizado devido à qualidade dos dados. Em bases de dados reais, dois desses aspectos estão relacionados com (i) a presença de valores desconhecidos, os quais são tratados de uma forma bastante simplista por diversos algoritmos de AM, e; (ii) a diferença entre o número de exemplos, ou registros de um banco de dados, que pertencem a diferentes classes, uma vez que quando essa diferença é expressiva, sistemas de aprendizado podem ter dificuldades em aprender o conceito relacionado com a classe minoritária.

O problema de tratamento de valores desconhecidos é de grande interesse prático e teórico. Em diversas aplicações é importante saber como proceder quando as informações disponíveis estão incompletas ou quando as fontes de informações se tornam indisponíveis. O tratamento de valores desconhecidos deve ser cuidadosamente planejado, caso contrário, distorções podem ser introduzidas no conhecimento induzido. Neste trabalho é proposta a utilização do algoritmo K-VIZINHOS MAIS PRÓXIMOS como método de imputação. Imputação é um termo que denota um procedimento que substitui os valores desconhecidos de um conjunto de dados por valores plausíveis. As análises conduzidas neste trabalho indicam que a imputação de valores desconhecidos com base no algoritmo K-VIZINHOS MAIS PRÓXIMOS pode superar o desempenho das estratégias internas utilizadas para tratar valores desconhecidos pelos sistemas C4.5 e CN2, bem como a IMPUTAÇÃO PELA MÉDIA OU MODA, um método amplamente utilizado para tratar valores desconhecidos.

O problema de aprender a partir de conjuntos de dados com classes desbalanceadas é de crucial importância, um vez que esses conjuntos de dados podem ser encontrados em diversos domínios. Classes com distribuições desbalanceadas podem se constituir em um gargalo significante no desempenho obtido por sistemas de aprendizado que assumem uma distribuição balanceada das classes. Uma solução para o problema de aprendizado com distribuições desbalanceadas de classes é balancear artificialmente o conjunto de dados. Neste trabalho é avaliado o uso do método de seleção unilateral, o qual realiza uma remoção cuidadosa dos casos que pertencem à classe majoritária, mantendo os casos da classe minoritária. Essa remoção cuidadosa consiste em detectar e remover casos considerados menos confiáveis, por meio do uso de algumas heurísticas.

Uma vez que não existe uma análise matemática capaz de predizer se o desempenho de um método é superior aos demais, análises experimentais possuem um papel importante na avaliação de sistema de aprendizado. Neste trabalho é proposto e implementado o ambiente computacional DISCOVER LEARNING ENVIRONMNET — DLE — o qual é um *framework* para desenvolver e avaliar novos métodos de pré-processamento de dados. O ambiente DLE é integrado ao projeto DISCOVER, um projeto de pesquisa em desenvolvimento em nosso laboratório para planejamento e execução de experimentos relacionados com o uso de sistemas de aprendizado durante a fase de Mineração de dados do processo de KDD.

# Abstract

Data quality is a major concern in Machine Learning, which is frequently used to extract knowledge during the Data Mining phase of the relatively new research area called Knowledge Discovery from Databases — KDD. As most Machine Learning algorithms induce knowledge strictly from data, the quality of the knowledge extracted is largely determined by the quality of the underlying data.

Several aspects may influence the performance of a learning system due to data quality. In real world databases, two of these aspects are related to (i) the presence of missing data, which is handled in a rather naive way by many Machine Learning algorithms; (ii) the difference between the number of examples, or database records, that belong to different classes since, when this difference is large, learning systems may have difficulties to learn the concept related to the minority class.

The problem of missing data is of great practical and theoretical interest. In many applications it is important to know how to react if the available information is incomplete or if sources of information become unavailable. Missing data treatment should be carefully thought, otherwise bias might be introduced into the knowledge induced. In this work, we propose the use of the K-NEAREST NEIGHBOUR algorithm as an imputation method. Imputation is a term that denotes a procedure that replaces the missing values in a data set by some plausible values. Our analysis indicates that missing data imputation based on the K-NEAREST NEIGHBOUR algorithm can outperform the internal missing data treatment strategies used by C4.5 and CN2, and the MEAN OR MODE IMPUTATION, a widely used method for treating missing values.

The problem of learning from imbalanced data sets is of crucial importance since it is encountered in a large number of domains. Imbalanced class distributions might cause a significant bottleneck in the performance obtained by standard learning methods, which assume a balanced distribution of the classes. One solution to the problem of learning with skewed class distributions is to artificially balance the data set. In this work we propose the use of the one-sided selection method, which performs a careful removal of cases belonging to the majority class while leaving untouched all cases from the minority class. Such careful removal consists of detecting and removing cases considered less reliable, using some heuristics. An experimental application confirmed the efficiency of the proposed method.

As there is not a mathematical analysis able to predict whether the performance of a learning system is better than others, experimentation plays an important role for evaluating learning systems. In this work we propose and implement a computational environment, the DISCOVER LEARNING ENVIRONMENT — DLE — which is a framework to develop and evaluate new data pre-processing methods. The DLE is integrated into the DISCOVER project, a major research project under development in our laboratory for planning and execution of experiments related to the use of learning systems during the Data Mining phase of the KDD process.

*Aos meus pais,*
*Joselito e Margarida,*

*Às minhas irmãs,*
*Anapaula e Analúcia,*

*À Maria Carolina Monard.*