INTRODUCTION TO DATA MINING

K. K. Kshetrapalapuram [1]

[1] Department of Computer Science & Software Engineering,
The University of Melbourne

EXTENDED TUTORIAL ABSTRACT

Data Mining seems to be a buzzword in the corporate world today. The collection of massive volumes of data, cheap hardware storage and fast multiprocessor systems combined with the need for efficiently summarising patterns in data for better decision-making has pushed data mining out of developmental research into practical industry and research use.

Statistics provides accurate (or within accurate error estimates) summaries and parameter values, but suffers from not being scalable. Artificial Intelligence compromises accuracy for speed. Data mining takes the accuracy of statistics and speed of AI to deliver a powerful new exploratory and predictive tool in any application domain.

The tutorial will begin with the role that data mining plays in knowledge discovery and Enterprise Resource Planning (ERP).

Once the data to be analysed is collected, statistical exploratory tools like correlations, and subjective reasoning can be used to reduce the dimensions of the problem. Principal Component Analysis for instance is used to determine which attributes in the data contribute to the most amount of variance in the data. The tutorial will explain how data exploration, transformation and visualisation techniques will help users understand the shape and limitations of the data being analysed, thus providing a subjective sense of the quality of the analysis.

Once the data is cleaned, a model of the data is built. Such a model summarises useful information about the data and may be used for link analysis (trying to determine time-related, frequent pattern or time-related relationships between attributes in the data) or for classification of the data. Model-building will be explained in the tutorial.

The model of the data is used to classify future data instances. Such a model can capture complex non-linear relationships between the data using algorithms inspired by Artificial Intelligence like decision trees and neural networks. This tutorial will explain various regression and classification algorithms that are suitable. The K-Means Clustering algorithm will introduce the audience to unsupervised classification techniques.

Other commonly used techniques such as Discriminant analysis, OLAP (On-line analytical Processing System) are also explained. An overview of the data mining software available today will expose the audience to the software aspect of data mining algorithms.

The tutorial assumes no prior knowledge of data mining or any other concept. Knowledge of basic statistics would be an advantage.