Chapter 1
Advanced Modelling Paradigms in Data Mining

Dawn E. Holmes\textsuperscript{1}, Jeffrey Tweedale\textsuperscript{2}, and Lakhmi C. Jain\textsuperscript{2}

\textsuperscript{1} Department of Statistics and Applied Probability
University of California Santa Barbara
Santa Barbara
CA 93106-3110
USA

\textsuperscript{2} School of Electrical and Information Engineering
University of South Australia
Adelaide
Mawson Lakes Campus
South Australia SA 5095
Australia

1 Introduction

As discussed in the previous volume, the term \textit{Data Mining} grew from the relentless growth of techniques used to interrogation masses of data. As a myriad of databases emanated from disparate industries, enterprise management insisted their information officers develop methodology to exploit the knowledge held in their repositories. Industry has invested heavily to gain knowledge they can exploit to gain a market advantage. This includes extracting hidden data, trends or pattern from what was traditionally considered noise. For instance most corporations track sales, stock, pay role and other operational information. Acquiring and maintaing these repositories relies on mainstream techniques, technology and methodologies. In this book we discuss a number of finding techniques and expand into intelligent paradigms.

2 Foundations

Management relies heavily of information systems to gain market advantage. For this reason they invest heavily in Information Technology (IT) systems that enable them to acquire, retain and manipulate industry related facts. Payroll and accounting systems were traditionally based on statistical manipulation, however have evolved to include machine learning and artificial intelligence. A non-exhaustive list of existing techniques would include:

- Artificial Intelligence (AI) Class introduction;
- Bayesian Networks;
- Biosurveillance;
- Cross-Validation;
- Decision Trees;
• Eight Regression Algorithms;
• Elementary probability;
• Game Tree Search Algorithms;
• Gaussian Bayes Classifiers and Mixture Models;
• Genetic Algorithms;
• K-means and Hierarchical Clustering;
• Markov Decision Processes and Hidden Markov Models;
• Maximum Likelihood Estimation;
• Neural Networks;
• Predicting Real-valued Outputs;
• Probability Density Functions;
• Probably Approximately Correct Learning;
• Reinforcement Learning;
• Robot Motion Planning;
• Search - Hill Climbing, Simulated Annealing and A-star Heuristic Search;
• Spatial Surveillance;
• Support Vector Machines;
• Time Series Methods;
• Time-series-based anomaly detection;
• Visual Constraint Satisfaction Algorithms; and
• Zero and non-zero-Sum Game Theory.

2.1 Statistical Modelling

Using statistics we are able to gain useful information from raw data. Based on a found-
ing knowledge of probability theory, statistical data analysis provides historical mea-
sures from empirical data. Based on this premise, there has been an evolutionary ap-
proach in Statistical Modelling techniques [2]. A recent example is Exceptional Model
Mining (EMM). This is a framework that allows for more complicated target concepts.
Rather than finding subgroups based on the distribution of a single target attribute,
EMM finds subgroups where a model fitted to that subgroup is somehow exceptional.
These models enable experts to discover historical results, but work has also been done
on prediction using analytical techniques.

2.2 Predictions Analysis

In order to gain a market advantage, industry continues to seek, forecast or predict future
trends [3]. Many algorithms have been developed to enable us to perform prediction and
forecasting. Many of these focus on improving performance by altering the means of
interacting with data. For example, Time Series Predictions is widely applied across var-
ious domains. There is a growing trend for industry to automate this process. Many now
produce annual lists that indexes or rates their competitors based on a series of business
parameters. Focuses on a series of observations that are statistically analyzed to generate
a prediction based on a predefined number of previous values. A recent example in this
book uses the average sum of nth-order difference of series terms with limited range
margins. The algorithm performances are evaluated using measurement data-sets of