Rhodes University
COMPUTER SCIENCE HONOURS PROJECT

Literature Review

Data Mining with Oracle 10g using Clustering and Classification Algorithms

***By**: Nhamo Mdzingwa*
***Supervisor:** John Ebden*
Date: 30 May 2005

Abstract

The field of data mining is concerned with learning from data or rather turning data into information. It is a creative process requiring a unique combination of tools for each application. However, the commercial world is fast reacting to the growth and potential in this area as a wide range of tools are marketed under the label of data mining. This literature survey will explore some of the ad hoc methodology generally used for data mining in the commercial world mainly focusing on the data mining process and data mining algorithms used. It will also include a brief description of the Oracle data mining tool.

1. Introduction

Data mining has been defined by [Han et al, 2001], [Roiger et al, 2003] and [Mannila et al, 2001] as a process of extracting or mining knowledge from large amounts of data, or simply knowledge discovery in databases. It has become useful over the past decade in business to gain more information, to have a better understanding of running a business, and to find new ways and ideas to extrapolate a business to other markets [Verhees, 2002]. Below [Palace, 1996] gives a fundamental example were data mining was used:

- One Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays.

It is however necessary to examine the algorithms that are put into practice when conducting the data mining process with more emphasis on accuracy and efficiency.

1.2 Classification of Data Mining

There is a wide range of sources available on data mining and most of these have various ways of implementing the data mining process, with most sources classifying data mining into categories. Authors find it convenient to categorise data mining corresponding to different objectives for the person analysing the data. According to [Roiger et al, 2003], data mining is classified into supervised and unsupervised concept learning methods. Supervised learning builds classification models by forming concept definitions from sets of data containing predefined classes while unsupervised clustering builds models from data without the aid of predefined classes were data instances are grouped together based on a similarity scheme defined by the clustering system. The authors also mention that supervised learning builds models by using input attributes to predict output attribute values while in unsupervised learning no target attributes are produced but rather giving a descriptive relationship by using an objective function to extract clusters in the input data or particular features which are useful for describing the data.

[Berry et al, 2000] also categorises data mining into directed data mining and undirected data mining as the two main styles of data mining. According to [Berry et al, 2000] the goal in directed data mining is to use the available data set to build a model that describes one particular variable of interest in terms of the rest of the available data. The authors also point out that directed data mining often takes the form of predictive modelling, where one knows what he want to predict. Classification, prediction and estimation are the techniques used in directed data mining. In undirected data mining, no variable is singled out as the target. The goal is to establish some relationship among all the variables. Examples of this type include clustering, association rules, description and visualisation. [Berry et al, 2000]

[Mannila et al, 2001] gives three general categorises of data mining namely; Exploratory Data Analysis (EDA), Descriptive Modelling and Predictive Modelling. With EDA the goal is to explore the data without any idea of what one is looking for and typical techniques are interaction and visualisation. A descriptive model presents the main features of the data. It is essentially a summary of the data permitting the study of the important aspects of the data. Clustering techniques are used in this category. In contrast, a predictive model has the specific objective of allowing one to predict the value of some target characteristic of an object on the basis of observed values of other characteristics of the object. This category includes techniques such as Classification and regression.

From the above, it is clear that the classified categories described by the different authors all involve similar techniques. We can then say that directed data mining, supervised learning and predictive modelling of data mining describe similar techniques that can be referred to as supervised learning. Unsupervised learning, undirected data mining and descriptive modelling are techniques in the same category and will be referred to as unsupervised learning.

2 Data mining algorithms

Most data mining tools are based on the use of algorithms to implement the categories described above. Supervised learning covers techniques that include classification, estimation, prediction, decision trees and association rules. Unsupervised learning covers techniques that include clustering, association rule induction, neural networks and association discovery or market basket analysis.

2.1 Supervised learning

As stated in **1.2**, supervised learning requires target attributes to be identified. The supervised learning technique then sifts through the data trying to find patterns between independent attributes (predictors) and the dependent attribute, then builds a model that best represent the functional relationships. [Pyle, 2000] Typically, for the data mining process, the data is separated into two parts; one for training and another for testing. The initial model is built using the first sample of the data and then the model is applied to the second sample to evaluate the accuracy of the model's predictions.

2.1.1 Classification Algorithms

[Han et al, 2001] describes classification as a model built describing a predetermined set of data classes or concepts. [Roiger et al, 2003] also describes classification as a technique where the dependent or output variable is categorical. The emphasis is on building a model able to assign new instances of data to categorical classes. Classification Algorithms are comprised of the Naive Bayes, Adaptive Bayes Network supporting decision trees, Model Seeker

2.1.1.1 Naïve Bayes

According to [Berger, 2004], the Naïve Bayes algorithm builds models that predict the probability of specific outcomes. Naïve Bayes algorithm achieves this by finding patterns and relationships in the data by counting the number of times various conditions are observed. It then builds a data mining model to represent those patterns and relationships. The data mining model represents these relationships and can be applied to new data to make predictions. Naïve Bayes algorithm makes predictions using Bayes' Theorem, a statistical theorem in nature. It assumes that the effect of an attribute value on a given class is independent of the values of other attributes (class conditionally independence) [Han et al, 2001]

[Berger, 2004] also emphasises that the algorithm provides quicker model building and faster application to new data than the Adaptive Bayes Network algorithm. [Han et al, 2001] points out that Basian classifiers also are known as the naïve Basian classifier have exhibited high accuracy and speed when applied to large databases. Naïve Bayes can also be used to make predictions of categorical classes that consist of binary-type outcomes or multiple categories of outcomes [Berger, 2004]. In attempting to answer the question: "how effective are Bayesian classifies" [Han et al, 2001] indicates that in theory they have minimum error in comparison to other techniques. The authors further indicate that

in practice this is not always the case owing to inaccuracies in the assumptions made for its use, such as conditional independence and the lack of availability of probability data.

2.1.1.2 Adaptive Bayes Network

According to [Berger, 2004] Adaptive Bayes Network (ABN) algorithm is similar to Naïve Bayes and, depending on the data being analyzed, can possibly produce better models. They can also be used to generate rules or decision tree-like outcomes when built and again to make predictions when applied to new data. The rules that are generated are easy to interpret in the form of "if…..then" statements but [Berger 2004] states that it does involve a larger number of parameters to be set and it tends to take a longer time to build such a model. An additional benefit of ABN models is that they are able to produce simple "rules" that may provide insight as to why the prediction was made. A typical "prediction" and "rule" might be:
[Berger, 2004]

Prediction: BMW = "YES"

ABN Rule: >30 AGE >40 and INCOME = High
Confidence: = 85% (634 cases fit this profile, 539 purchased BMW autos)
Support = .00543 (539 cases out of 99,263 records)

2.1.1.3 Tree algorithms

A decision tree is a flow-chart-like tree structure, with internal nodes representing an attribute. In decision tree data mining, a record flows through the tree along a path determined by a series of tests until a terminal node is reached and it is then given a class label. Decision trees are useful for classification and predictions as they assign records to broad categories and output rules that can be easily translated. Different criteria are used to determine when splits in the tree occur [Han et al, 2001]. The CART (classification and regression trees) incorporates the machine learning algorithms which generates binary trees as it rates high on statistical prediction. CART divides a data set on the basis of variety to determine the best separators [Mannila et al, 2001]. The efficiency of exiting decision tree algorithms has been established for relatively small data sets. [Han et al, 2001] points out that efficiency and scalability become issues of concern when these algorithms are applied to the mining of large databases.

2.1.1.4 Association rules

Association rule mining searches for interesting relationships among items in a given data set. Market basket analysis is just one form of association rule mining [Han et al, 2001]. According to [*Al-Attar, 2004*], association rules are similar to decision trees and association rule induction is the most established and effective of the current data mining technologies. This technique involves the definition of a business goal and the use of rule induction to generate patterns relating this goal to other data fields. The patterns are generated as trees with splits on data fields. This technique allows the user to add their domain knowledge to the process and decide on attributes for generating splits [Han et al, 2001].

2.2 Unsupervised learning

With unsupervised learning, the user does not specify a target attribute for the data mining algorithm. Unsupervised learning techniques such as associations and clustering algorithms make no assumptions about a target field. Instead, they allow the data mining algorithm to find associations and clusters in the data independent of any a priori defined business objective. [Berger, 2004]

2.2.1 Clustering

[Berry et al, 2000] defines clustering as the task of segmenting a diverse group of attributes into a number of more similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. [Roiger et al, 2003] say clustering is useful for determining if meaningful relationships exist in the data, evaluating the performance of supervised learning models, detecting outliers in the data and even determining input attributes for supervised learning.

2.2.1.1 Clustering algorithms

Enhanced k-Means and Orthogonal Partitioning Clustering (O-Cluster)

Enhanced k-Means (EKM) and O-Cluster algorithms support identifying naturally occurring groupings within the data population EKM algorithm supports hierarchical clusters, handles numeric attributes and will cut the population into the user specified number of clusters. The algorithm divides the data set into k number of clusters according to the location of all members of a particular cluster in the data. Clustering makes use of the Euclidean distance formula to determine the location of data instances and their position in clusters and so requires numerical values that have been properly scaled. When choosing the number of clusters to create it is possible to choose a number that doesn't match the natural structure of the data which leads to poor results. For this reason [Berry et al, 2000] say it is often necessary to experiment with the number of clusters to be used. O-cluster algorithm handles both numeric and categorical attributes and will automatically select the best cluster definitions. [Berger, 2004]

2.2.2 Neural network

Neural network algorithm is also enveloped as unsupervised learning technique. According to [Berry et al, 2000] neural networks are the most widely used known and the least understood of the major data mining techniques. [Pyle, 2000] describes neural networks as network construction network system of interconnected interacting weights at each node acting as input and output stations. Each input to the network gets its own node which consists of a transformation of input variables fed in. The input unit is connected to the output unit with a weighting and the input is combined in the output unit with a combination function. The activation function is the passed transfer function. [Berry et al, 2000] says training a neural network is a process that involves setting weights on inputs to best approximate a target variable. This is important for optimizing

the neural network. Three steps are involved in training. Training instance variables, calculating outputs using existing weights and calculating errors and accordingly adjusting weights.

[Berry et al, 2000] further elaborates that neural networks, neural networks are not easy to use and understand but they produce very good results. The authors continue to say that neural networks require extensive data preparation as inputs must be scaled, categorical data must be converted to numerical data without introducing any ordering and missing values must be dealt with. The authors suggest that a problem with neural networks is that the results cannot be explained and so they should be used when results are more useful than understanding and not when there are a high number of inputs.

The figure below by [Elder, 1998] shows some useful properties employed by some of the algorithms described:

Algorithm	Accurate	Scalable	Interpret-able	Useable	Robust	Versatile	Fast	Hot
Classical (LR, LDA)	–	good	good –	good	–	–	good	bad
Neural Networks	good	bad	bad	bad	–	bad	bad bad	good
Visualization	good	bad bad	good	good	good good	bad	bad bad bad	good –
Decision Trees	bad	good	good	good –	good	bad	good –	good –
Polynomial Networks	good	–	bad	good –	– bad	–	– bad	–
K-Nearest Neighbors	bad	bad bad	good –	–	– bad	bad	good	bad
Kernels	good	bad bad	bad	– bad	bad	bad	good	bad

Key
good
– neutral
bad

3 Data Mining Process

There is increased interest in a process or methodology for data mining. This process is another important aspect that needs to be examined as it layouts clear steps that can be followed in the data mining process. It is argued that such a formalised process will widen the exploitation of data mining as an enabling technology for solving business

problems. It will allow people with varying expertise in data mining and from different business sectors to carry out successful data mining projects with a high degree of consistency. [*Al-Attar, 2004*]

[Berger, 2004] believes that to be effective in data mining, successful data analysts generally following a four step process:

1) **Problem definition -**This is the most important step and is where the domain expert decides the specifics of translating an abstract business objective e.g. “How can I sell more of my product to customers?” into a more tangible and useful data mining problem statement e.g. “Which customers are most likely to purchase product A?” To build a predictive model that predicts who is most likely to buy product A, we first must have data that describes the customers who have purchased product A in the past. Then we can begin to prepare the data for mining.

2) **Data gathering and preparation** In this step, we take a closer look at our available data and determine what additional data we will need to address our business problem. We often begin by working with a reasonable sample of the data, e.g., hundred of records (rare, except in some life sciences cases) to many thousands or millions of cases (more typical for business-to-consumer cases). Some processing of the data to transform for example a “Date_of_Birth” field into “AGE” and to derive fields such as “Number_of_times_Amount_Exceeds_100” is performed to attempt the “tease the hidden information closer to the surface of the data” for easier mining.

3) **Model building and evaluation** Once steps 1 and 2 have been properly completed, this step is where the data mining algorithms sift through the data to find patterns and to build predictive models. Generally, a data analyst will build several models and change mining parameters in an attempt to build the best or most useful models.

4) **Knowledge deployment** Once a useful model that adequately models the data has been found, you want to distribute the new insights and predictions to others—managers, call center representatives, and executives.

[Roiger et al, 2003] also elaborate a data mining process where emphasis is placed on data preparation for model building.

1) **Goal identification**. Properly identifying goals to be accomplished by the data mining project help with domain understanding and determining what is to be accomplished.
2) **Creating the target data**. It is emphasized that at this stage a human expert is required to choose the initial data for the project.

3) **Data preprocessing in order to deal with noisy data**. This stage involves locating duplicate records in the data set, locating incorrect attributes, smoothing the data and dealing with outliers in the data set. It includes data transformation which involves the addition or removal of attributes and instances, normalizing of data and type conversions.
4) **The actual data mining** -At this stage the model is built from training and test data sets. The resulting model is then interpreted to determine if the results it presents are useful or interesting. The model or acquired knowledge is then applied to the problem.

[Verhees, 2002] gives a brief methodology for conducting data mining also identifying problems normally associated with the process. The problems include the nature of data in the database as it is often incomplete, noisy or very large, inadequate or irrelevant. Also included are the errors in the stored data. The steps should include:

1) **Problem analysis**- here that is when it will be determined whether the problem is suitable for data mining and what data and technologies are available. Also at this stage it will be important to determine what will be done with the results of the data mining to put the problem in context.

2) **Data preparation**- should be part of the methodology and data processing. Processing involves pre-processing or cleansing of the data, data integration, variable transformation and splitting or sampling from the database.
3) **Data exploration**-. This allows the analyst or data miner to discover the unexpected in the data as well as to confirm the expected.
4) **Pattern generation**-should follow which involves applying the algorithms and validating and interpreting the patterns that result.
5) **Model validation**-is required in order to confirm the usability of the developed model. Validation can be conducted using a validation data set and assesses the quality of the model fit to the data as well as protecting the model from over- or under-fitting.
6) **Interpretation and decision making**-conclude the methodology and attempt to transform the patterns discovered during data mining into knowledge.

There are a number of initiatives for the development of a formal/documented data mining process in the world. It is reassuring to the data mining community that the processes emerging from all of these initiatives reveal a large degree of similarity. There is widespread agreement on the main steps or stages involved in such a process and any differences relate only to the detailed tasks within each stage. [*Al-Attar, 2004*] gives a summary of the major stages of a data mining process is:

1) **Goal definition-**This involves defining the goal or objective for the data mining project. This should be a business goal or objective which normally relates to a business event such as arrears in mortgage repayment, customer attrition (churn), energy consumption in a process, etc. This stage also involves the design of how the discovered patterns would be utilised as part of the overall business solution.
2) **Data selection-** This is the process of identifying the data needed for the data mining project and the sources of this data.
3) **Data preparation-** This involves cleansing the data, joining/merging data sources and the derivation of new columns (fields) in the data through aggregation, calculations or text manipulation of existing data fields. The end result is normally a flat table ready for the application of the data mining itself (i.e. the discovery algorithms to generate patterns). Such a table is normally split into two data sets; one set for pattern discovery and one set for pattern verification.
4) **Data exploration-** This involves the exploration of the prepared data to get a better feel prior to pattern discovery and also to validate the results of the data preparation. Typically, this involves examining the statistics (minimum, maximum, average, etc.) and the frequency distribution of individual data fields. It also involves field versus field graphs to understand the dependency between fields.
5) **Pattern Discovery-** This is the stage of applying the pattern discovery algorithm to generate patterns. The process of pattern discovery is most effective when applied as an exploration process assisted by the discovery algorithm. This allows business users to interact with and to impart their business knowledge to the discovery process. In the case of inducing a tree, users can at any point in the tree construction examine / explore the data filtering to that path, examine the recommendation of the algorithm regarding the next data field to use for the next branch then use their business judgement to decide on the data field for branching. The pattern discovery stage also involves analysing the ability of the discovered patterns to predict the propensity of the business event, and for verification against an independent data set.
6) **Pattern deployment-** This stage involves the application of the discovered patterns to solve the business goal of the data mining project. This can take many forms:
    - **Patterns presentation:** The description of the patterns (or the graphical tree display) and their associated data statistics are included in a document or presentation.
    - **Business intelligence:** The discovered patterns are used as queries against a data base to derive business intelligence reports. This requires the data mining tool to generate SQL representations of the decision tree.

    - **Data Scoring & Labelling:** The discovered patterns are used to score and/or label each data record in the database with the propensity and the label of the pattern it belongs to. This can be done directly by the data mining tool or through generation of SQL or C representation of the decision tree

- **Alarm monitoring:** The discovered patterns are used as 'norms' for a business process. Monitoring these patterns will enable deviations from normal conditions to be detected at the earliest possible time. This can be achieved by embedding the data mining tool as a monitoring component, or through using SQL generated by the data mining tool.

7) **Pattern Validity monitoring-** As a business process changes over time, the validity of patterns discovered from historic data will deteriorate. It is therefore important to detect these changes at the earliest possible time by monitoring patterns with new data. Significant changes to the patterns will point to the need to discover new patterns from more recent data.

4 Oracle Data Mining Tool

Oracle Data Mining is a powerful data mining software embedded in the 10g Database Enterprise Edition (EE) that enables you to discover new insights hidden in your data [Berger, 2004]. The Oracle Data Mining suite is made up of two components, the data mining Java API and the Data Mining Server (DMS). The DMS is a server-side, in-database component that performs data mining that is easily available and scalable. The DMS also provides a repository of metadata of the input and result objects of data mining.

As stated by [Berger, 2004] Oracle Data Mining supports supervised learning techniques (classification, regression, and prediction problems), unsupervised learning techniques (clustering, associations, and feature selection problems), attribute importance techniques (find the key variables), text mining, and has a special algorithm for life sciences sequence searching and alignment problems.

Oracle Data Mining can scale to the size of the problem by adding hardware or switching to more powerful platforms. Oracle Data Mining takes advantage of Oracle's parallelism for faster computing by leveraging Oracle's Real Application Clusters (RAC) technology. [Oracle, 2005]

Application developers access Oracle Data Mining's functionality through a Java-based or PL/SQL interface. Programmatic control of all data mining functions enables automation of data preparation, model-building, and model-scoring operations in production applications. [Oracle, 2005]

Choice of algorithm for Oracle Data Mining depends on the data available for mining as well as the types of results and conclusions required. [Berger, 2004] continues to give a discussion on when it is applicable to use the various algorithms; the states that if there is a set of input data and output data, then it is more likely that supervised learning will be used since input and output attributes exist. [Hand, Mannila and Smyth, 2001] also states that if the input and output data is numerical or categorical and have interesting

interactions association rules are recommended. On the other hand, [Trueblood and Lovett, Jnr] say that if the data sets have missing values neural networks may be a good choice. [*Al-Attar, 2004*] further states that when maximum accuracy is required of a model, it is helpful to create multiple models using the same data mining technique until the best model is created.

5 Conclusion

Data mining is becoming a strategically important area for many business organisations, and due to its applied importance, however, the field emerges as a rapidly growing and major area [Verhees, 2002]. It can thus be concluded that data mining is a step wise process that requires the insight and experience of the data miner. The process is also supported by the use of various software tools available for data mining

References:

[*Al-Attar, 2004*] White Paper: Data Mining - Beyond Algorithms, *Dr Akeel Al- Attar, 2004*, <http://www.attar.com/tutor/mining.htm> Accessed: 10 April 2005

[Berger, 2004] Berger, C., 09/2004, *Oracle Data Mining, Know More, Do More, Spend Less - An Oracle White Paper*, URL: <http://www.oracle.com/technology/products/bi/odm/pdf/bwp_db_odm_10gr1_0904.pdf>, Accessed: 14 April 2005

[Berry et al, 2000] *Mastering Data Mining: The Art and Science of Customer Relationship Management*, Michael J.A. Berry and Gordon S. Linoff, USA, Wiley Computer Publishing, 2000

[Han et al, 2001] *Data mining: concepts and techniques* by Jiawei Han and Micheline Kamber, San Francisco, California, Morgan Kauffmann, 2001.

[Mannila et al, 2001] David Hand, Heikki Mannila and Padhraic Smyth, *Principles of data mining*, Cambridge Massachusetts, MIT Press, 2001.

[Oracle, 2005] *The Oracle Home Page.* Revised February 2005. **Oracle 10g Data Mining FAQ** http://www.oracle.com/technology/products/bi/odm/odm_10g_faq.html#api

Accessed: 17 May 2005

[Palace, 1996] Bill Palace , Spring 1996. "Data Mining: What is Data Mining?" Anderson Graduate School of Management at UCLA: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm Accessed: 15 April 2005

[Paul et al, 2002] *Preparing and Mining Data in Microsoft SQL Server 2000 and Analysis Services*, Seth Paul, Nitin Gautam, Raymond Ballint, **Published:** September 2002, **Updated:** January 2003

[Pyle, 2000] *Data Preparation for Data Mining*: Dorian Pyle, San Francisco, California, Morgan Kauffman, 2000.

[Roiger et al, 2003] *Data mining: a tutorial- based primer* by Richard J. Roiger and Michael W. Geatz, Boston, Massachusetts, Addison Wesley, 2003..

[Trueblood et al, Jnr] Robert P. Trueblood and John N. Lovett, Jnr. *Data Mining and Statistical Analysis Using SQL*, USA, Apress, 2001

[Verhees, 2002] *Enhance your Application – Simple Integration of Advanced Data Mining Functions*, Corinne Baragoin, Ronnie Chan, Helena Gottschalk, Gregor Meyer, Paulo Pereira, Jaap Verhees, 2002, <http://www.redbooks.ibm.com/redbooks/SG246879.html> Accessed: 12 April 2005

[Elder, 1998] A Comparison of Leading Data Mining Tools-Elder Research. John F. Elder IV & Dean W. Abbott, last updated August 28, 1998 http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf Accessed: 15 May 2005