CHAPTER-7

DATA MINING TECHNIQUES
AND ALGORITHMS
CHAPTER-7

DATA MINING TECHNIQUES AND ALGORITHMS

7.1 INTRODUCTION

The purpose of this study is not to delve deeply into the technical aspects of data mining but to provide a general overview of the subject. However, in order to be able to decide when one techniques is called for, or when another would be more suitable, at least a basic understanding of not only the different data mining tasks are required, but also of the most important data mining techniques and algorithms. Highlight three basic data mining techniques as being very important since they are implemented in the majority of commercial software applications. They also cover a wide range of data mining situations. The most commonly used undirected data mining technique that will be discussed in the following section is:

- Automatic cluster detection. Two of the most commonly used directed data mining techniques that will be discussed in the following sections are:
  - Decision trees.
  - Neural networks.

- **Automatic Cluster Detection**

  Clustering is a method of grouping together similar records in a data set. The most commonly used algorithm for automatic cluster detection is K-means. This algorithm works by dividing a data set into a predetermined number of clusters. The number of clusters is represented by the phrase “k” in K-means. A mean is an average ad in
this case it refers to the average location of all the members of a cluster. Automatic cluster detection is an undirected data mining technique. As a result of this it can be applied without prior knowledge of the structure to be discovered. This is also its weakness in that if you do not know what you are looking for, it is difficult to recognize it when you find it. According to Berry and Linoff (200:109), automatic cluster detection is most useful in the following circumstances: if it is suspected that the data set contains natural groupings that may represent customers or products that have a lot in common with each other. It may turn out that these are naturally occurring customer segments that can be singled out for customized marketing approaches.

When there are many competing patterns in the data set making it hard to identify a single pattern. In this case automatic cluster detection can be used to create cluster of similar records thereby reducing the complexity of the data set so that other data mining techniques are more likely to succeed.

- **Decision Trees**

Mena (1999: 357) [20] defines a decision tree as a graphical representation of the relationship between a dependent variable (output) and a set of independent variables (input), usually in the form of a tree-shaped structure that represent a set of decisions with each node representing a test of decision. DT work by allowing records from a data set to flow through a series of tests such as “is the field 3 greater than 27?” until the record research a leaf or terminal node where it is given a class label based on the class of the records that
reached that node in the training set when the model was originally set up. Define following two types of decision trees:

- Classification trees that label records in a data set and assign them to the a proper class.
- Regression trees that estimate the value of a target variable that takes on numeric values. An example of a regression tree algorithm analysis is to calculate the expected size of claims that will be made by a insured person. Decision tree algorithms are a good choice for use in the following circumstances.
  - When the data mining tasks involves the classification of records in a data set or the prediction of outcomes.
  - When the goal is to assign each data set record to one of a few broad categories. The main advantage of using decision trees is found in its ability to generate understandable business rules in a decision support environment and the ability to model nonlinear with logical [4]. Unfortunately, decision trees also have some disadvantages associated with their use in data mining.
  - The data used by decision trees must be categorical or interval, ad data not received in this format will have to be recorded to this format in order to be used.
  - Decision trees normally represent finite number of classes or possibilities, and it becomes difficult for decision makers to quantify a finite number of variables, the accuracy of the results obtained will be limited to the number of classes selected.
  - Decision trees are suited for problems involving time series data unless a lot of efforts is put into presenting the date in such a way that trends are made visible.
• **Neural Networks**

From a data mining perspective, neural networks are simply another way of fitting a model to observed historical data in order to be able to make classifications or predictions. Kantardzic (2003:196) [21] defines a neural network as a massive parallel distributed processor made up of simple processing units with the ability to learn from experiential knowledge expressed through inter-unit connection strengths, and that can make such knowledge available for use. Highlights the following useful properties and capabilities of neural networks for use in data mining application:

- Nonlinearity – ANT nonlinearity models the inherently nonlinear real world mechanisms responsible for generating data for learning.
- Learning from example – A neural networks has the ability to tune its parameters in order to facilitate the process of learning from experience.
- Adaptability – A neural network is able to adapt to changes in its operating environment by changing its interconnection weights.
- Evidential response – A neural network can be designed not only provide information about which particular class to select for a given sample, but also about confidence in the decision made.
- Fault tolerance – A neural network has the potential to be inherently fault tolerant and capable of robust computation, meaning that it can handle missing or incorrect data more effectively than other data mining techniques.
- Uniformity of analysis and design– The same principles, notation and steps in methodology are used in all domain involving the application of neural networks.
• Neural networks are a good choice for most classification and prediction tasks when the results obtained from the model are more important than understanding how the model works.

• Neural networks are not a good idea if there are many hundreds or thousands of input features. Such large numbers of input parameters makes it difficult for the network to find patterns and can results in very long training phases that never coverage into a good solution.

• Training the neural network can take up a considerable amount of time and resources, and it is therefore recommended to limit its application to small to medium datasets.

• Neural networks lack explicitness due to its black box nature, and it is therefore difficult to understand how the neural network came to its conclusion.

• The neural network will never be able to solve a problem that a human could not given enough time.

• There is a very limited tradition of experience on which to draw when choosing between the neural networks on offer, making it difficult or impossible for an organization to confidently choose an appropriate solution.