# CHAPTER 3: DATA MINING: AN OVERVIEW

## 3.1 Introduction

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods such as neural networks or decision trees. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing) [48].

The term "data mining" is primarily used by statisticians, database researchers, and the business communities. The term KDD (Knowledge Discovery in Databases) refers to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process [48, 57]. The steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful knowledge is derived from the data. Data mining is an extension of traditional data analysis and statistical approaches as it incorporates analytical techniques drawn from various disciplines like AI, machine learning, OLAP, data visualization, etc.

## 3.2 Classification of Data Mining System

Data mining systems can be categorized according to various criteria as follows [45]:

- **Classification of data mining systems according to the type of data sources mined**: This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

- **Classification of data mining systems according to the database involved**: This classification based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.

- **Classification of data mining systems according to the kind of knowledge discovered:** This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

- **Classification of data mining systems according to mining techniques used**: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

## 3.3 The Knowledge Discovery Process

Data mining is one of the tasks in the process of knowledge discovery from the database.

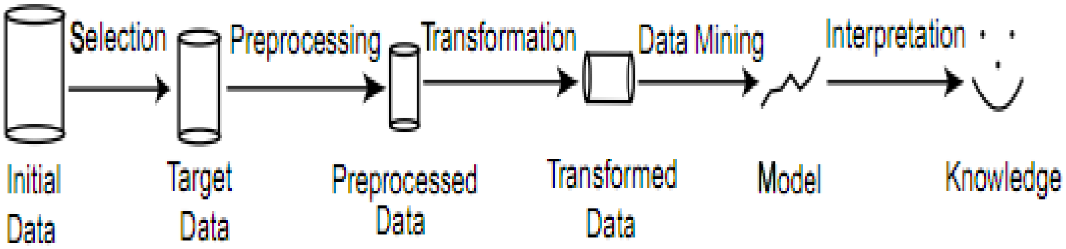

**Figure 3.1: Steps of Knowledge Discovery in Databases**

The steps in the KDD process contain [45]:

- **Data cleaning:** It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collection.

- **Data integration:** In this stage, multiple data sources, often heterogeneous, are combined in a common source.

- **Data selection:** The data relevant to the analysis is decided on and retrieved from the data collection.

- **Data transformation:** It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure.

- **Data mining:** It is the crucial step in which clever techniques are applied to extract potentially useful patterns.

- **Pattern evaluation:** In this step, interesting patterns representing knowledge are identified based on given measures.

- **Knowledge representation:** It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

## 3.4 Data Mining Life cycle

The life cycle of a data mining project consists of six phases [91, 26]. The sequence of the phases is not rigid. Moving back and forth between different phases is always required depending upon the outcome of each phase. The main phases are:

- **Business Understanding**: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

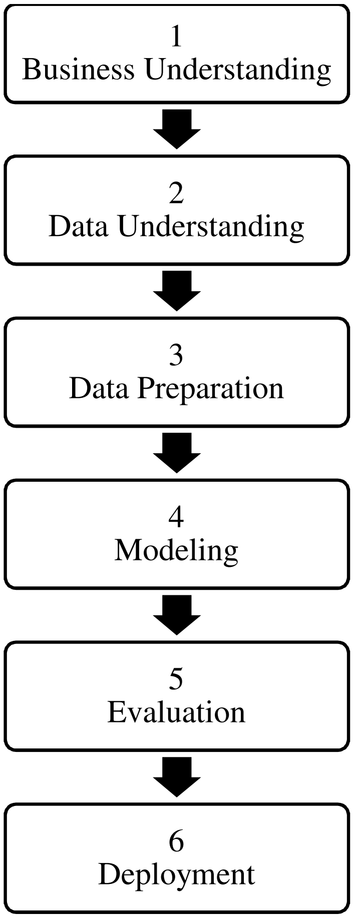

**Figure 3.2: Phases of Data Mining Life Cycle**

- **Data Understanding**: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

- **Data Preparation**: covers all activities to construct the final dataset from raw data.

- **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

- **Evaluation:** In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

- **Deployment**: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

## 3.5 Data Mining Functionalities

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data [119]. The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

- **Characterization:** It is the summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may wish to characterize the customers of a store who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute oriented induction method can be used to carry out data summarization. With a data cube containing summarization of data, simple OLAP operations fit the purpose of data characterization.

- **Discrimination**: Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may wish to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.

- **Association analysis**: Association analysis studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. This is commonly used for market basket analysis. For example, it could be useful for the manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form: P→Q [s, c], where P and Q are conjunctions of attribute value-pairs, and s (support) is the probability that P and Q appear together in a transaction and c (confidence) is the conditional probability that Q appears in a transaction when P is present. For example, RentType(X,“game”)□Age(X,“13-19”)→Buys(X,“pop”)[s=2%, =55%]

  The above rule would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a pop, and that there is a certainty of 55% that teenage customers who rent a game also buy pop.

- **Classification**: It is the organization of data in given classes. Classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the manager of a store could analyze the customers’ behavior vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels “safe”, “risky” and “very risky”. The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.

- **Prediction:** Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

- **Clustering**: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

- **Outlier analysis**: Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.

- **Evolution and deviation analysis**: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system.

## 3.6 Data Mining Models

The data mining models are of two types [146, 70]: Predictive and Descriptive.

### 3.6.1 Descriptive Models

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc. Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters. The clusters are defined by studying the behavior of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database into disjoint grouping of similar tuples. Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes. Association rule mining is a two-step process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

### 3.6.2 Predictive Models

The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc. Many of the data mining applications are aimed to predict the future state of the data.

Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes, this is a supervise learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that map data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

## 3.7 Data Mining and Statistics

The disciplines of statistics and data mining both aim to discover structure in data. So much do their aims overlap, that some people regard data mining as a subset of statistics. But that is not a realistic assessment as data mining also makes use of ideas, tools, and methods from other areas – particularly database technology and machine learning, and is not heavily concerned with some areas in which statisticians are interested. Statistical procedures do, however, play a major role in data mining, particularly in the processes of developing and assessing models. Most of the learning algorithms use statistical tests when constructing rules or trees and also for correcting models that are over fitted. Statistical tests are also used to validate machine learning models and to evaluate machine learning algorithms [70].

Some of the commonly used statistical analysis techniques are discussed below.

Descriptive and Visualization Techniques include simple descriptive statistics such as averages and measures of variation, counts and percentages, and cross-tabs and simple correlations. They are useful for understanding the structure of the data. Visualization is primarily a discovery technique and is useful for interpreting large amounts of data; visualization tools include histograms, box plots, scatter diagrams, and multi-dimensional surface plots.

- Cluster Analysis seeks to organize information about variables so that relatively homogeneous groups, or "clusters," can be formed. The clusters formed with this family of methods should be highly internally homogenous (members are similar to one another) and highly externally heterogeneous (members are not like members of other clusters).

- Correlation Analysis measures the relationship between two variables. The resulting correlation coefficient shows if changes in one variable will result in changes in the other. When comparing the correlation between two variables, the goal is to see if a change in the independent variable will result in a change in the dependent variable. This information helps in understanding an independent variable's predictive abilities. Correlation findings, just as regression findings, can be useful in analyzing causal relationships, but they do not by themselves establish causal patterns.

- Discriminant Analysis is used to predict membership in two or more mutually exclusive groups from a set of predictors, when there is no natural ordering on the groups. Discriminant analysis can be seen as the inverse of a one-way multivariate analysis of variance (MANOVA) in that the levels of the independent variable (or factor) for MANOVA become the categories of the dependent variable for discriminant analysis, and the dependent variables of the MANOVA become the predictors for discriminant analysis.

- Factor Analysis is useful for understanding the underlying reasons for the correlations among a group of variables. The main applications of factor analytic techniques are to reduce the number of variables and to detect structure in the relationships among variables; that is to classify variables. Therefore, factor analysis can be applied as a data reduction or structure detection method. In an exploratory factor analysis, the goal is to explore or search for a factor structure. Confirmatory factor analysis, on the other hand, assumes the factor structure is known a priori and the objective is to empirically verify or confirm that the assumed factor structure is correct.

- Regression Analysis is a statistical tool that uses the relation between two or more quantitative variables so that one variable (dependent variable) can be predicted from the other(s) (independent variables). But no matter how strong the statistical relations are between the variables, no cause-and-effect pattern is necessarily implied by the regression model.

Regression analysis comes in many flavors, including simple linear, multiple linear, curvilinear, and multiple curvilinear regression models, as well as logistic regression. Logistic Regression is used when the response variable is a binary or qualitative outcome. Although logistic regression finds a "best fitting" equation just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, that is, it maximizes the probability of obtaining the observed results given the fitted regression coefficients. Because logistic regression does not make any assumptions about the distribution for the independent variables, it is more robust to violations of the normality assumption. Some of the more common flavors that logistic regression comes in include simple, multiple, polychromous and Poisson logistic regression models.

## 3.8 Data Mining Techniques and Algorithms

This section provides an overview of some of the most common data mining algorithms in use today. The section has been divided into two broad categories:

- Classical Techniques: Statistics, Neighborhoods and Clustering
- Next Generation Techniques: Trees, Networks and Rules

These categories will describe a number of data mining algorithms at a high level and shall help to understand how each algorithm fits into the landscape of data mining techniques. Overall, six broad classes of data mining algorithms are covered. Although there are a number of other algorithms and many variations of the techniques described,

### 3.8.1 Classical Techniques: Statistics, Neighborhoods and Clustering

This category contains descriptions of techniques that have classically been used for decades and the next category represents techniques that have only been widely used since the early 1980s. The main techniques here are the ones that are used 99.9% of the time on existing business problems. There are certainly many other ones as well as proprietary techniques from particular vendors - but in general the industry is converging to those techniques that work consistently and are understandable and explainable.

### 3. 8.1.1 Statistics

By strict definition statistics or statistical techniques are not data mining. They were being used long before, the term data mining was coined to apply to business applications. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. This is why it is important to have the idea of how statistical techniques work and how they can be applied.

#### 3.8.1.1.1 Prediction using Statistics

The term "prediction" is used for a variety of types of analysis that may elsewhere be more precisely called regression. Regression is further explained in order to simplify some of the concepts and to emphasize the common and most important aspects of predictive modeling. Nonetheless regression is a powerful and commonly used tool in statistics.

#### 3.8.1.1.2 Linear Regression

In statistics prediction is usually synonymous with regression of some form. There are a variety of different types of regression in statistics but the basic idea is that a model is created that maps values from predictors in such a way that the lowest error occurs in making a prediction. The simplest form of regression is simple linear regression that just contains one predictor and a prediction. The relationship between the two can be mapped on a two dimensional space and the records plotted for the prediction values along the Y axis and the predictor values along the X axis. The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line (the prediction from the model). Graphically this would look as it does in figure 3.3. The simplest form of regression seeks to build a predictive model that is a line that maps between each predictor value to a prediction value. Of the many possible lines that could be drawn through the data the one that minimizes the distance between the line and the data points is the one that is chosen for the predictive model.

On average if one presumes the value on the line it should represent an acceptable compromise amongst all the data at that point giving conflicting answers. Likewise if there is no data available for a particular input value the line will provide the best guess at a reasonable answer based on similar data.

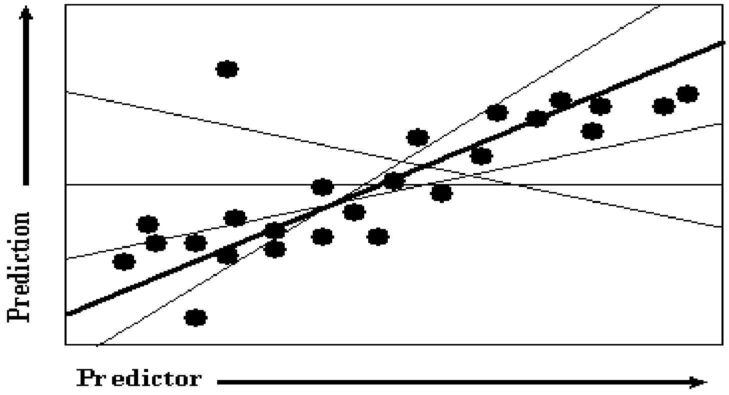

**Figure 3.3: Predictive Modeling through Linear Regression**

The predictive model is the line shown in figure 3.3. The line will take a given value for a predictor and map it into a given value for a prediction. The equation would look like: Prediction = a + (b * Predictor) which is just the equation for a line Y = a + bX.

As an example for a bank the predicted average consumer bank balance might equal 1,000 + (0.01 * customer's annual income). The trick, with predictive modeling, is to find the model that best minimizes the error. The most common way to calculate the error is the square of the difference between the predicted value and the actual value. Calculating the points that are very far from the line will have a great effect on moving the choice of line towards them in order to reduce the error. The values of a and b in the regression equation that minimize this error can be calculated directly from the data relatively quickly.

Regression can become more complicated than the simple linear regression in a variety of different ways in order to better model particular database problems. There are, however, three main modifications that can be made:

- More predictors than just one can be used.
- Transformations can be applied to the predictors.
- Predictors can be multiplied together and used as terms in the equation.
- Modifications can be made to accommodate response predictions that just have yes/no or 0/1 values.

Adding more predictors to the linear equation can produce more complicated lines that take more information into account and hence make a better prediction. This is called multiple linear regression and might have an equation like the following if 5 predictors were used (X1, X2, X3, X4, X5):

$$Y = a + b1(X1) + b2(X2) + b3(X3) + b4(X4) + b5(X5)$$

This equation still describes a line but it is now a line in a 6 dimensional space rather than the two dimensional space.

By transforming the predictors by squaring, cubing or taking their square root, it is possible to use the same general regression methodology and now create much more complex models that are no longer simple shaped like lines. This is called non-linear regression. A model of one predictor might look like this: $Y = a + b1(X1) + b2 (X2)$. In many real world cases, analysts perform a wide variety of transformations on their data. If they do not contribute to a useful model their coefficients in the equation will tend toward zero and then can be removed. The other transformation of predictor values is multiplying them together. For example a new predictor created by dividing hourly wage by the minimum wage, can be more effective predictor than hourly wage by itself.

When trying to predict a customer response that is just yes or no, the standard form of a line doesn't work. Since there are only two possible values to be predicted it is relatively easy to fit a line through them. However, that model would be the same no matter what predictors/ data were being used. Typically in these situations a transformation of the prediction values is made in order to provide a better predictive model. This type of regression is called logistic regression and because so many business problems are response problems, logistic regression is one of the most widely used statistical techniques for creating predictive models.

### 3.8.1.2 Nearest Neighbor

Clustering and the Nearest Neighbor prediction technique are among the oldest techniques used in data mining. Most people think that clustering is like records are grouped together. Nearest neighbor is a prediction technique that is quite similar to clustering. Its essence is that in order to predict what a prediction value is in one record look for records with similar predictor values in the historical database and use the prediction value from the record that is "nearest" to the unclassified record.

Example of the nearest neighbor algorithm is that if you look at the people in your neighborhood, you may notice that, in general, you all have somewhat similar incomes. Thus if your neighbor has an income greater than 100,000 chances are good that you too have a high income. Now the chances that you have a high income are greater when all of your neighbors have incomes over 100,000 than if all of your neighbors have incomes of 20,000. Within your neighborhood there may still be a wide variety of incomes possible among even your "closest" neighbors but if you had to predict someone's income based on only knowing their neighbors your chance of being right would be to predict the incomes of the neighbors who live closest to the unknown person.

The nearest neighbor prediction algorithm works in very much the same way except that "nearness" in a database may consist of a variety of factors not just where the person lives. It may, for instance, be far more important to know which school someone attended and what degree they attained when predicting income. The better definition of "near" might in fact be other people that you graduated from college with rather than the people that you live next to. Nearest Neighbor techniques are easy to use and understand because they work in a way similar to the way that people think - by detecting closely matching examples. They also perform quite well in terms of automation, as many of the algorithms are robust with respect to dirty data and missing data. Lastly they are particularly adept at performing complex ROI calculations because the predictions are made at a local level where business simulations could be performed in order to optimize ROI. As they enjoy similar levels of accuracy compared to other techniques the measures of accuracy such as lift are as good as from any other.

## Nearest Neighbor for Prediction

One of the essential elements underlying the concept of clustering is that one particular object (whether they be cars, food or customers) can be closer to another object than can some third object. It is interesting that most people have an innate sense of ordering placed on a variety of different objects. Most people would agree that an apple is closer to an orange than it is to a tomato and that a Toyota Corolla is closer to a Honda Civic than to a Porsche. This sense of ordering on many different objects helps us place them in time and space and to make sense of the world. It is what allows us to build clusters - both in databases on computers as well as in our daily lives. This definition of nearness that seems to be ubiquitous also allows us to make predictions.

The nearest neighbor prediction algorithm simply stated is:

Objects that are "near" to each other will have similar prediction values as well. Thus if you know the prediction value of one of the objects you can predict it for it's nearest neighbors.

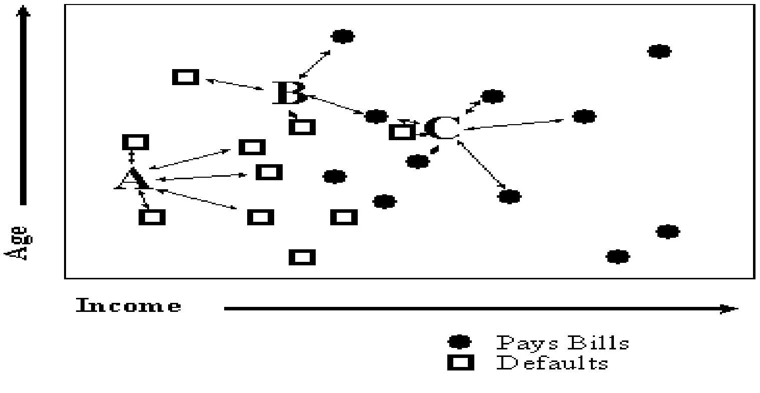

**Figure 3.4: The nearest neighbors for three unclassified records**

### 3.8.1.3 Clustering

In an unsupervised learning environment the system has to discover its own classes and one way in which it does this is to cluster the data in the database as shown in the following diagram. The first step is to discover subsets of related objects and then find descriptions e.g. Dl, D2, D3 etc. which describe each of these subsets.

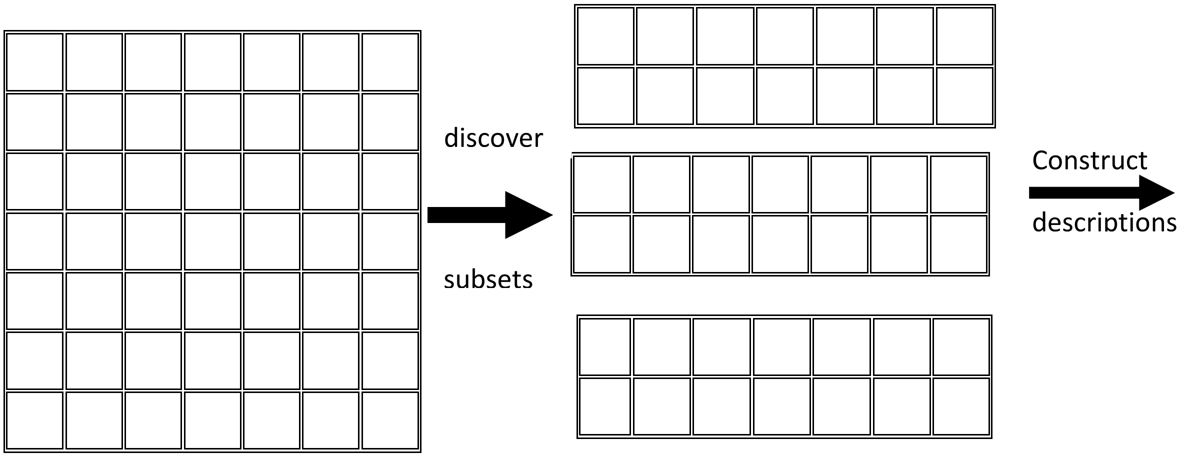

**Figure 3.5: Discovering clusters and descriptions in a database**

Clustering is basically a partition of the database so that each partition or group is similar according to some criteria or metric. Clustering according to similarity is a concept, which appears in many disciplines. If a measure of similarity is available there are a number of techniques for forming clusters. Membership of groups can be based on the level of similarity between members and from this the rules of membership can be defined. Another approach is to build set functions that measure some property of partitions i.e. groups or subsets as functions of some parameter of the partition. This latter approach achieves what is known as optimal partitioning.

Many data mining applications make use of clustering according to similarity for example to segment a client/ customer base. Clustering according to optimization of set functions is used in data analysis e.g. when setting insurance tariffs the customers can be segmented according to a number of parameters and the optimal tariff segmentation achieved.

### 3.8.1.3.1 Hierarchical Clustering

There are two main types of clustering techniques, those that create a hierarchy of clusters and those that do not. The hierarchical clustering techniques create a hierarchy of clusters from small to big. The main reason is that clustering is an unsupervised learning technique, and as such, there is no absolutely correct answer. Now depending upon the particular application of the clustering, fewer or greater numbers of clusters may be desired. With a hierarchy of clusters defined it is possible to choose the number of clusters that are desired. Also it is possible to have as many clusters as there are records in the database. In this case the records within the cluster are optimally similar to each other and certainly different from the other clusters. Such a clustering technique misses the point in the sense that clustering is to find useful patters in the database that summarize it and makes it easier to understand. Thus one of the main points about clustering is that there should be many fewer clusters than there are original records.

The hierarchy of clusters is usually viewed as a tree where the smallest clusters merge together to create the next highest level of clusters and those at that level merge together to create the next highest level of clusters. Figure 3.6 shows how several clusters might form a hierarchy. From such hierarchy the user can determine what the right number of clusters is that adequately summarizes the data while still providing useful information

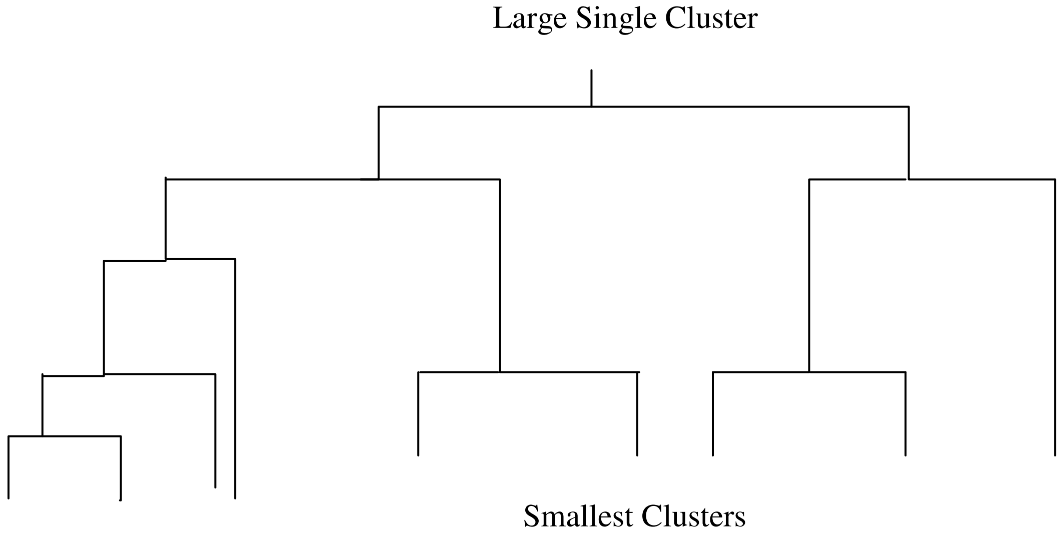

**Figure 3.6: Hierarchical Clustering**

This hierarchy of clusters is created through the algorithm that builds the clusters. There are two main types of hierarchical clustering algorithms:

- Agglomerative: Agglomerative clustering techniques start with as many clusters as there are records where each cluster contains just one record. The clusters that are nearest to each other are merged together to form the next largest cluster. This merging is continued until a hierarchy of clusters is built with just a single cluster containing all the records at the top of the hierarchy.
- Divisive: Divisive clustering techniques take the opposite approach from agglomerative techniques. These techniques start with all the records in one cluster and then try to split that cluster into smaller pieces and then in turn to try to split those smaller pieces into more smaller ones.

### 3.8.1.3.2 Non-Hierarchical Clustering

There are two main non-hierarchical clustering techniques. Both of them are very fast to compute on the database but have some drawbacks. The first are the single pass methods. They derive their name from the fact that the database must only be passed through once in order to create the clusters (i.e. each record is only read from the database once). The other class of techniques is called reallocation methods. They get their name from the movement or "reallocation" of records from one cluster to another in order to create better clusters. The reallocation techniques do use multiple passes through the database but are relatively fast in comparison to the hierarchical techniques.

Some techniques allow the user to request the number of clusters that they would like to be pulled out of the data. Predefining the number of clusters rather than having them driven by the data might seem to be a bad idea as there might be some very distinct and observable clustering of the data into a certain number of clusters which the user might not be aware of.

For instance the user may wish to see their data broken up into 10 clusters but the data itself partitions very cleanly into 13 clusters. These non-hierarchical techniques will try to shoe horn these extra three clusters into the existing 10 rather than creating 13 which best fit the data. One of the advantages of these techniques is that most of the times the user does have some predefined level of summarization that they are interested in (e.g. “25 clusters is too confusing, but 10 will help to give me an insight into my data”). The fact that greater or fewer numbers of clusters would better match the data is actually of secondary importance.

### 3.8.2 Next Generation Techniques: Trees, Networks and Rules

#### 3.8.2.1 Decision Trees

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. For instance if we were going to classify customers who churn (don’t renew their phone contracts) in the Cellular Telephone Industry a decision tree might look something like that found in Figure 3.7.

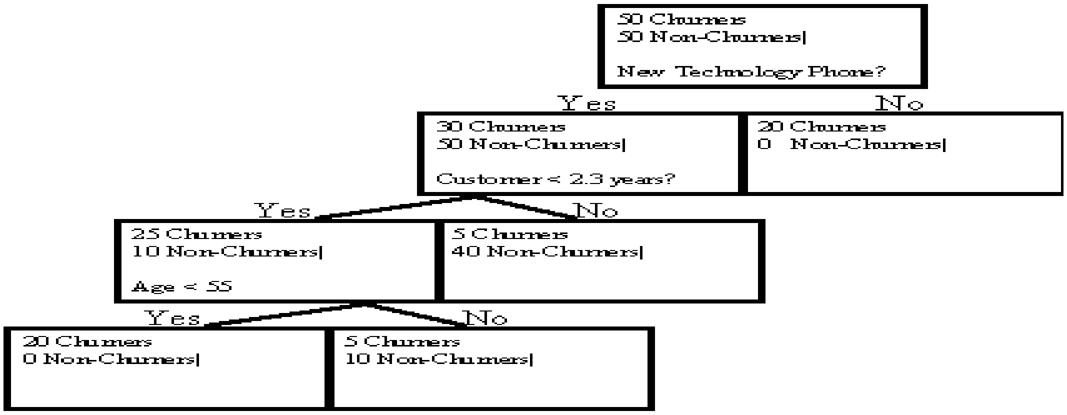

**Figure 3.7: Decision Tree for Cellular Telephone Industry**

There are some interesting things about the tree:

- It divides up the data on each branch point without losing any of the data (the number of total records in a given parent node is equal to the sum of the records contained in its two children).

- The number of churners and non-churners is conserved as you move up or down the tree
- It is pretty easy to understand how the model is being built (in contrast to the models from neural networks or from standard statistics).
- It would also be pretty easy to use this model if you actually had to target those customers that are likely to churn with a targeted marketing offer.

You may also build some intuitions about your customer base. e.g. “customers who have been with you for a couple of years and have up to date cellular phones are pretty loyal”.

### Prediction using Decision Tree

Although some forms of decision trees were initially developed as exploratory tools to refine and preprocess data for statistical techniques like logistic regression. They have also been used for prediction. This is interesting because many statisticians still use decision trees for effectively building a predictive model as a byproduct but then ignore the predictive model in favor of techniques that they are most comfortable with. Sometimes veteran analysts will do this even excluding the predictive model when it is superior to that produced by other techniques. With a host of new products and skilled users, tendency to use decision trees only for exploration seems to be changing.

## 3.8.2.2 Neural Networks

Neural networks is an approach to computing that involves developing mathematical structures with the ability to learn. The methods are the result of academic investigations to model nervous system learning. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data. This can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if' questions.

Neural networks have already been successfully applied in many industries. Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting needs. The structure of a neural network is shown in figure 3.8.

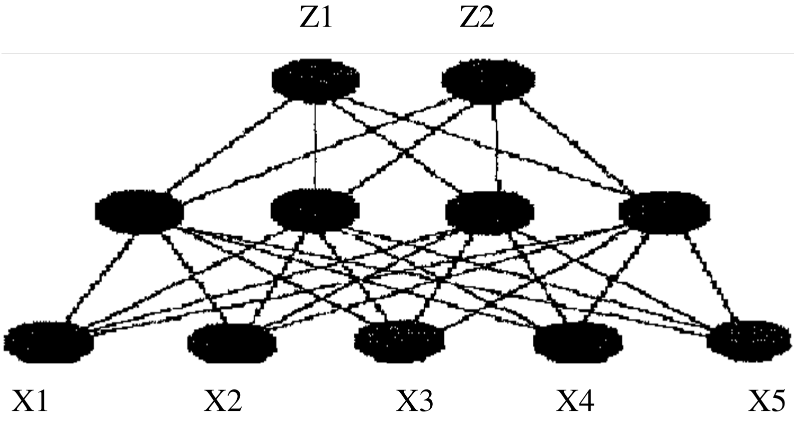

**Figure 3.8: Structure of a neural network**

Here, the bottom layer represents the input layer, in this case with 5 inputs labels Xl through X5. In the middle, there is the hidden layer, with a variable number of nodes. The hidden layer performs much of the work of the network. The output layer in this case has two nodes, Z1 and Z2 representing output values determined from the inputs. For example, predict sales (output) based on past sales, price and season (input).

In Figure 3.9, there is a drawing of a simple neural network. The round circles represent the nodes and the connecting lines represent the links. The neural network functions by accepting predictor values at the left and performing calculations on those values to produce new values in the node at the far right. The value at this node represents the prediction from the neural network model.

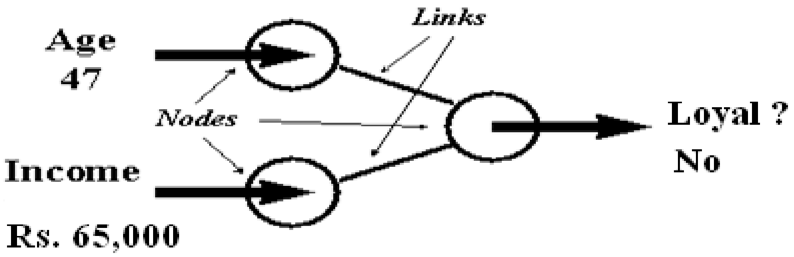

**Figure 3.9: Simplified View of Neural Network.**

In this case the network takes in values for predictors for age and income and predicts whether the person will default on a bank loan.

### Prediction using Neural Network

In order to make a prediction the neural network accepts the values for the predictors on the input nodes. These become the values for those nodes whose values are then multiplied by values that are stored in the links (sometimes called links and in someways similar to the weights that were applied to predictors in the nearest neighbor method). These values are then added together at the node at the far right (the output node) a special threshold function is applied and the resulting number is the prediction. In this case if the resulting number is 0 the record is considered to be a good credit risk (no default) if the number is 1 the record is considered to be a bad credit risk (likely default).

A simplified version of the calculations made in figure 3.9 might look like what is shown in figure 3.10. Here the value age of 47 is normalized to fall between 0.0 and 1.0 and has the value 0.47 and the income is normalized to the value 0.65. This simplified neural network makes the prediction of no default for a 47 year old making 65,000. The links are weighted at 0.7 and 0.1 and the resulting value after multiplying the node values by the link weights is 0.39. The network has been trained to learn that an output value of 1.0 indicates default and that 0.0 indicate non-default. The output value calculated here (0.39) is closer to 0.0 than to 1.0 so the record is assigned a non-default prediction.

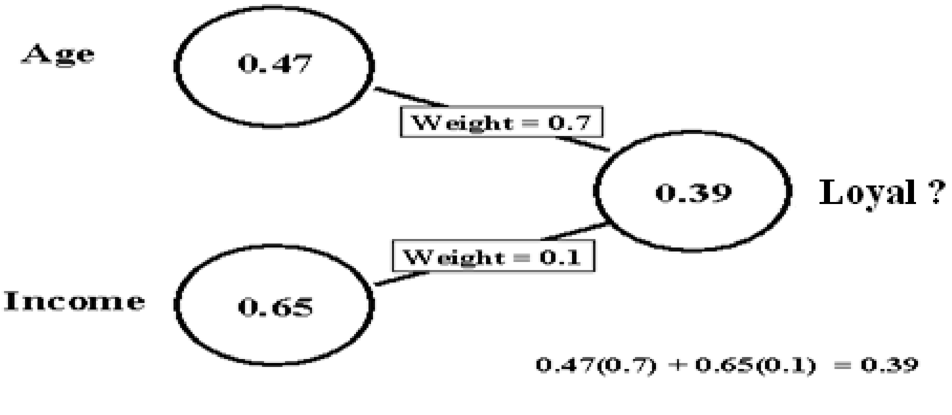

**Figure 3.10: Neural Network for Prediction of Loyalty.**

#### 3.8.2.3 Rule Induction

Rule induction is one of the major forms of data mining and is the most common form of knowledge discovery in unsupervised learning systems. Rule induction on a data base can be a massive undertaking where all possible patterns are systematically pulled out of the data and then an accuracy and significance are added to them that tell the user how strong the pattern is and how likely it is to occur again. In general these rules are relatively simple such as for a market basket database of items scanned in a consumer market basket you might find interesting correlations in your database such as:

- If bagels are purchased then cream cheese is purchased 90% of the time and this pattern occurs in 3% of all shopping baskets.
- If live plants are purchased from a hardware store then plant fertilizer is purchased 60% of the time and these two items are bought together in 6% of the shopping baskets.

The rules that are pulled from the database are extracted and ordered to be presented to the user, based on the percentage of times that they are correct and how often they apply. The bane of rule induction systems is also its strength, that it retrieves all possible interesting patterns in the database. This is a strength in the sense that it leaves no stone unturned but it can also be viewed as a weakness because the user can easily become overwhelmed with such a large number of rules that it is difficult to look through all of them. One almost need a second pass of data mining to go through the list of interesting rules that have been generated by the rule induction system in the first place in order to find the most valuable gold nugget amongst them all. This overabundance of patterns can also be problematic for the simple task of prediction because all possible patterns are culled from the database there may be conflicting predictions made by equally interesting rules. Automating the process of culling the most interesting rules and of combing the recommendations of a variety of rules is well handled by many of the commercially available rule induction systems on the market today and is also an area of active research.

### Prediction using Rule Induction

Once the rules are created and their interestingness is measured, then prediction is performed with the rules. Each rule by itself can perform prediction - the consequent is the target and the accuracy of the rule is the accuracy of the prediction. Rule induction systems produce many rules for a given antecedent or consequent, so there can be conflicting predictions with different accuracies. This is an opportunity for improving the overall performance of the systems by combining the rules. This can be done in a variety of ways by summing the accuracies as if they were weights or just by taking the prediction of the rule with the maximum accuracy.

## 3.9 Selection of an Appropriate Technique

Clearly one of the hardest things to do when deciding to implement a data mining system is to determine which technique to use when. When is data mining appropriate at all as opposed to just working with relational databases and reporting? When would just using OLAP and a multidimensional database be appropriate?

Some of the criteria that are important in determining the technique to be used are determined by trial and error. There are definite differences in the types of problems that are most conducive to each technique but the reality of real world data and the dynamic way in which markets, customers and hence the data that represents them is formed means that the data is constantly changing. These dynamics mean that it no longer makes sense to build the "perfect" model on the historical data since whatever was known in the past cannot adequately predict the future because the future is so unlike what has gone before.

In some ways this situation is analogous to the business person who is waiting for all information to come in before they make their decision. They are trying out different scenarios, different formulae and researching new sources of information. But this is a task that will never be accomplished - at least in part because the business the economy and even the world are changing in unpredictable and even chaotic ways that could never be adequately predicted. Better to take a robust model that perhaps is an underperformer compared to what some of the best data mining tools could provide with a great deal of analysis and execute it today rather than to wait until tomorrow when it may be too late.

## 3.10 Conclusion

This Chapter discusses the various data mining concepts, functionalities, tools and techniques. The disciplines of statistics and data mining have also been discussed to prove that these areas are highly interrelated and share a symbiotic relationship. This chapter also helps to gain a major understanding of the various data mining algorithms and the way these can be utilized in various business applications and the way these algorithms can be used in the descriptive and predictive data mining modeling. The chapter ends with the discussion on the selection of appropriate data mining techniques in a particular scenario.