Chapter 1

Introduction to Data Mining
1.1 Introduction

The world now is overwhelmed with data, the digital revolution has made digitized information easy to capture, process, store, distribute and transmit. The amount of data seems to go on and on increasing and the progress in digital data acquisition and storage technology has resulted in the growth of huge databases. The knowledge Discovery from huge number of databases and massive volume of data is a challenge. Within these masses of data lies hidden information of strategic importance. When there are so many trees, how do we draw meaningful conclusions about the forest? The newest answer is data mining, which is being used both to increase revenues and to reduce costs. The potential returns are enormous. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud.

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The first and simplest analytical step in data mining is to describe the data, summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables (such as values that often occur together). As emphasized in a later section, collecting, exploring and selecting the right data are critically important. But data description alone cannot provide an action plan. We must build a predictive model based on patterns determined from known results, then test that model on results outside the original sample. A good model should never be confused with reality but it can be a useful guide to understanding business. The final step is to empirically verify the model. Data Mining is an attempt to make sense of the information explosion embedded in this huge volume of data [BS 2004]. Many
people treat Data Mining as a synonym for another used term, i.e. KDD, or Knowledge Discovery in Databases. While others see Data Mining as an essential step in the process of Knowledge Discovery in Databases. Knowledge Discovery of Databases as a process is depicted in Figure 1.1 and we see Data Mining as a step of the process.

Figure 1.1 Data mining: Knowledge Discovery Process
1.2 Challenges in Data Mining:

Collecting the data for mining is very hard process by itself as the ongoing operations everyday generate tremendous and huge amount of data. Data Mining helps the end users extract interesting business information or patterns from large databases, and the larger the volume of data that can be processed by data mining techniques, the greater the confidence in the result [BS 2004 & PM]. Data mining process can be of one or more of the following functions such as classification rules, regression, time series analysis, prediction, clustering, summarization, association rules and sequence discovery. The number of generating rules would be very high and only few of the discovered patterns are of the interest to the end user. Many researchers have identified some measures of interestingness of discovered rules. These measures are support, confidence, statistical significance, simplicity [LHML 1996] and these measures are called objective measures and unexpectedness, actionability and novelty which are called subjective measures.

1.3 Type of Data that are collected

Based on the types of data that our mining techniques are applied to, data mining can be classified into different categories:

1.3.1 Relational database: till now most data are stored in relational database and relational database is one of the biggest resources of the data mining objects. As we know relational database is highly structured data repository, data are described by a set of attributes and stored in tables. Data mining on relational database mainly focus on discovering patterns and trends.

1.3.2 Transactional database: transactional database refers to the collection
of transaction records, in most cases they are sales records. With the popularity of computer and e-commerce, massive transactional database are available now. Data mining on transactional database focuses on the mining of association rules, finding the correlation between items in the transaction records.

1.3.3 Spatial database: spatial databases usually contain not only traditional data but also location or geographic information about the corresponding data. Spatial association rules describe the relationship between one set of features and another set of features in a spatial database. Algorithms for mining spatial association rules are similar to association rule mining except consideration of spatial data, the predicates generation and rules generation processes are based on Apriori.

1.3.4 Temporal and time-series database: it differs from traditional transaction data, for each temporal data item the corresponding time related attributes is associated. Temporal association rules can be more useful and informative than basic association rules. For example rather than the basic association rule \{diapers\} \rightarrow \{beer\}, mining from the temporal data we can get a more insight rule that the support of \{diapers\} \rightarrow \{beer\} jumps to 50% during peak hours 7-10 pm everyday, obviously retailers can make more efficient promotion strategy by using temporal association rule.

1.3.5 World Wide Web: Information on the web increases in a phenomena speed and web becomes ubiquitous, most researchers turn to the field of mining web data. Web mining is usually divided into three main
categories, web usage mining, web structure mining and web content mining. Web usage mining concentrates on mining the access patterns of users, so that the structure of the web site can be modified based on the navigation patterns. Different application of mining web logs have been developed to find navigation patterns. Besides improving the web site structure, web usage mining is also valuable for cross marketing strategies, web advertisements and promotion campaigns.

Web structure mining focuses in analysis of structures and links in web documents. The basic idea is that those pages that are linked together have some kinds of relationship. With those links, a typical structure mining is to classify those web documents into authoritative pages and hub pages. Web content includes text, graphic, media etc.. Consequently web content mining includes text mining, multimedia mining and graphic mining.

1.4 Data Mining and Data Warehousing:

Frequently, the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart (Figure 1.2). There is some real benefit if to be mined data is already part of a data warehouse. The problems of cleansing data for a data warehouse and for data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined. The data mining database may be a logical rather than a physical subset of your data warehouse, provided that the data warehouse DBMS can support the additional resource demands of data mining. If it cannot, then we will be better off with a separate data mining database.
1.5 Data Mining, Machine Learning and Statistics:

Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification. Both communities have made great contributions to the understanding and application of neural nets and decision trees. Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions. New techniques include relatively recent
algorithms like neural nets and decision trees, and new approaches to older algorithms such as discriminant analysis. By virtue of bringing to bear the increased computer power on the huge volumes of available data, these techniques can approximate almost any functional form or interaction on their own. Traditional statistical techniques rely on the modeler to specify the functional form and interactions. The key point is that data mining is the application of these and other AI and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. Data mining is a tool for increasing the productivity of people trying to build predictive models.

1.6 Data Mining Techniques:

1.6.1 Classification: classification is the most commonly applied data mining technique. Classification is a method of categorizing or assigning class labels to a pattern set under the supervision of a teacher. Decision boundaries are generated to discriminate between patterns belonging to different classes. The patterns are initially partitioned into training and test sets, and the classifier is trained on the former. The test set is used to evaluate the generalization capability of the classifier. A decision tree classifier is one of the most widely used supervised learning methods used for data exploration. It’s easy to interpret and can be represented as if-then-else rules. It approximates a function by piecewise constant regions and does not require any prior knowledge of the data distribution.

Early examples of classification techniques are [Mitchell 1982], [Quinlan 1986]. [Mitchell 1982] induces a single classification rule from two complementary trees (a specialization tree and generalization tree) that converge.
on a common node containing the rule. [Quinlan 1986] induces a decision tree. An object is classified by descending the tree until a branch leads to a leaf node containing the decision.

Other examples of classification techniques are [[Quinlan 1993], [Shapiro 1991], [Fifield 1992]]. C4.5 in [Quinlan 1993] is an industry-quality descendant of ID3 in [Fifield 1992] that has seen widespread use in the research community. KID3 in [Shapiro 1991] induces exact decision rules, i.e. those that are always correct, and strong decision rules i.e. those that are almost always correct.

1.6.2 Association: The task of association rule mining is to find certain association relationships among a set of objects (called items) in a database. The association relationships are described in association rules. Each rule has two measurements, support and confidence. Confidence is a measure of the rule’s strength, while support corresponds to statistical significance.

The task of discovering association rules was first introduced in 1993 [AIS 93]. Originally, association rule mining is focused on market “basket data” which stores items purchased on a per-transaction basis. A typical example of an association rule on market “basket data” is that 70% of customers who purchase bread also purchase butter. Later, association rule mining is also extended to handle quantitative data.

Finding association rules is valuable for crossing-marketing and attached mailing applications. Other applications include catalog design, add-on sales, store layout, and customer segmentation based on buying patterns. Besides application on business area, association rule mining can also be applied to other areas, such as medical diagnosis, remotely sensed imagery.
1.7 Motivation and Problem Definition:

Knowledge Discovery of Databases (KDD) is the process of extracting previously unknown but useful and significant information from large massive volume of databases. Data Mining is a stage in the entire process of KDD which applies an algorithm to extract interesting patterns. Usually, such algorithms generate huge volume of patterns. These patterns have to be evaluated by using interestingness measures to reflect the user requirements. Interestingness is defined in different ways, (i) Objective measures (ii) Subjective measures. Objective measures such as support and confidence extract meaningful patterns based on the structure of the patterns, while subjective measures such as unexpectedness and novelty reflect the user perspective. Objective measures of interestingness may not highlight the most important patterns produced by the data mining system, subjective measures generally operate by comparing the beliefs of a user against the patterns discovered by the data mining algorithm. It should be noted that both objective and subjective measures should be used to select interesting rules. Objective measures can be used as a kind of first filter while subjective measures can be used as a final filter to elect truly interesting rules [ZXS]. Identifying interesting rules from a set of discovered rules is not a simple task because a rule could be interesting to one user but of no interesting to another. The interestingness of a rule is a subjective matter because it depends on the user’s existing concepts and information about the domain and user’s interest. In this work we introduce another measure of rule interestingness that is shocking rules and we propose an algorithm for incremental association rules mining that integrates shocking interestingness criterion during the process of building the model. One of the main features of the proposed approach is to capture the user background knowledge, which is monotonically augmented. The proposed algorithm makes use of
interestingness measure as the basis of extracting interesting patterns. This important feature of the proposed algorithm is attractive and desirable in many real life applications as the volume of data keeps on growing and changing over the time and therefore the user background knowledge is monotonically augmented. This changing environment updates the user understandability and comprehensibility about the domain.