Chapter 1

Introduction

Huge amount of data have been collected through the advances of database
technologies and data collection techniques. Some of the domains, where large
volume of data are stored are Financial Investment, Health Care, Manufactur­
ing and Production, Telecommunication Network, Scientific Domain, etc. These
databases are full of hidden patterns and knowledge, which can be used for differ­
ent purposes. The subject which deals with hidden patterns in a large database
and finds knowledge from a large database is known as Knowledge Discovery
in Database (KDD). Data Mining can be defined as extracting knowledge from
huge amount of data. The Following subsections will clarify the concept of data
mining more clearly.

1.1 What is Data Mining

In the simplest form, Data Mining can be defined as extraction of knowledge
from huge amount of data. More formally, data mining can be defined as the
non-trivial extraction of implicit, previously unknown and potentially useful in­
formation from database. So, it can be compared with gold mining, diamond
mining, etc. Some other terms with similar meaning are knowledge mining,
knowledge extraction, pattern analysis, data dredging, etc. Some people consider
data mining as a part of the whole KDD (Knowledge Discovery in Database)
(Figure 1.1 on page 3 [HK01]) process, which can be defined as the non-trivial
process of identifying valid, novel, potentially useful and ultimately understandable patterns. Again, many people treat data mining as synonym for KDD.

As shown in the Figure 1.1 on the next page, KDD consists of the following steps.

1. Cleaning and Integration: Noise and inconsistent data are removed. Multiple data sources are combined.

2. Selection and Transformation: Required data are selected and transformed into forms appropriate for mining using different data mining techniques.

3. Data Mining: Different techniques, algorithms are used to extract knowledge from the data.

4. Evaluation and Presentation: Interesting patterns are found depending on some criteria. Patterns are represented using different types of GUI.

There are different architectures of a data mining system. However, the three-tier architecture (Figure 1.2 on page 4 [HK01]) is more popular. In this architecture, the major components are Database and Data Warehouse, Database or Data warehouse server, Data mining engine, Knowledge base, Pattern evaluation and Graphical user interface. A brief description of these components is given below.

- Database, data warehouse: This refers to set of databases, data warehouses, spreadsheets and other sources of data. Data cleaning and integration may be required.

- Database or data warehouse server: This is required to store and fetch relevant data.

- Knowledge base: This refers to knowledge repository, which is required to find interesting patterns. It may include concept hierarchy, meta data, user beliefs, some threshold, etc.
Figure 1.1: KDD Process
Figure 1.2: Three-tier Architecture of Data Mining
• Data mining engine: This is the main module which performs tasks such as association, classification, clustering, evaluation, etc.

• Pattern evaluation: It determines whether a pattern is interesting or not. To find the interestingness of a pattern, it interacts with data mining engine, knowledge base, etc.

• Graphical user interface: This module is responsible to interact with users. The major task of this module is to take the user's query and other parameters. Then it presents the results of the queries in some understandable formats using the available GUI tools.

1.1.1 Definitions

The main purpose of data mining is to find hidden patterns from large databases. However, data mining has been defined in many ways by different authors. Some of the definitions are given below [Puj01].

1. Data mining or knowledge-discovery in databases, as it is also known, is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes and detecting anomalies. By non-trivial, it means that information should not be easily retrievable. As for example, calculating age from date of birth, which is stored in a database, is not non-trivial, but finding average age of employees, who suffer from a particular disease and work in a particular department, may be non-trivial. Another term used in the definition is implicit. It means that information retrieved should not be stored in database explicitly. However, it could be derived from the existing data. Again, information or pattern should be previously unknown and unexpected. As for example “80% people buy bread and butter together” is not unknown. However, “2% people buy bread and spoon together” may be unexpected. Information should be useful to users. In other words, information should be presented in the user understandable format so that they can be used in decision support systems, etc.
2. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among vast amount of data, such as the relationship between patient data and their medical diagnostics. This relationship represents valuable knowledge about database, and the objects in the database, if the database is faithful mirror of the real world registered by the database. This definition gives importance on the relationships among the objects in a database. Suppose, there is a database which stores customers' data and sells data in a supermarket. Finding relationships between customers' age and items bought by them may be interesting. As for example, "customers in the age group of 10 to 20 years prefer food items like maggi, cake, etc." may be useful to the supermarket owner.

3. Data mining refers to using a variety of techniques to identify nuggets of information or decision-making knowledge in the database and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but it has low value and no direct use can be made of it. It is the hidden information in the data that is useful. Huge volume of data is not useful by itself. Data mining techniques find value from this huge volume of data, which can be used by decision makers.

4. Discovering relations that connect variables in a database is the subject of data mining. The data mining system self-learns from the previous history of the investigated system, formulating and testing hypothesis about rules which systems obey. When concise and valuable knowledge about the system of interest is discovered, it can and should be interpreted into some decision support system, which helps the manager to make wise and informed business decision. Here, a data mining system has been considered as a learning system, which learns from the existing data. So, it can be compared with machine learning systems.

5. Data mining is a process of discovering meaningful, new correlation patterns and trends by shifting through large amount of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical techniques. This definition says that data mining is meant to handle large amount of data, which makes it different from other data ana-
lyzing tools. While dealing with large amount of data, it also uses existing statistical and mathematical tools.

1.1.2 True Data Mining

A true data mining system should be able to handle large volume of data and uses advanced techniques to understand the data. So, it can be considered as the advanced stage of OLAP (OnLine Analytical Processing). It is often confused with OLAP. OLAP is generally involved with aggregate-style analytical processing. However, data mining uses advanced techniques to find the patterns in the data in different forms. There are some commercial systems which are used for information retrieval, answering queries, finding aggregate values, statistical analysis, etc. These systems are not true data mining systems and should not be confused with data mining systems.

1.2 Data Mining as a Multi-disciplinary Subject

Data mining is not a single subject. It is the result of confluence of many interdisciplinary subjects (Figure 1.3 on the next page [Puj01]). It uses techniques from various subjects such as machine learning, statistics, neural networks, information retrieval, spatial data analysis, database technology, etc. These subjects are established by themselves and have contributed a lot in developing different data mining algorithms and enhancing their performance.

Let us consider the subject of statistics. Statistics is one important subject from data mining point of view and a theory-rich method for data analysis. It provides theoretical foundations and generates results which is difficult to interpret. However, statistics is the foundation which data mining is based on. There exist statistical tools to find patterns from data, which can be understood by people with strong statistical background. Moreover, these tools deal with small amount of data.
CHAPTER 1. INTRODUCTION

Machine learning also has contributed in the development of data mining. Machine learning is the automation of learning process which includes learning from examples, reinforcement learning, learning with a teacher, etc. Again, machine learning can be of two types - supervised learning and unsupervised learning. In case of supervised learning, the system uses some training set to find description of each class. This description is used to place an unknown object in the appropriate class. On the other hand, unsupervised learning system does not use any training set and prior knowledge. It generates class descriptions from observations and discovery.

Data visualization also is an important subject in the context of data mining.
CHAPTER 1. INTRODUCTION

Data visualization helps analysts get deeper understanding of data. It gives analysts visual representations such as map, charts, etc. for a large volume of data. It also helps analysts concentrate on certain patterns and trends, which are represented by different colors.

Database technology also helps data mining systems in different ways. However, database systems and data mining systems are not same. In most of the cases, data mining systems use database systems as a simple repository of data which data mining algorithms are based on. Database systems use some powerful and established techniques such as SQL, query optimization to retrieve data efficiently. These techniques are used to develop similar techniques for data mining systems. Some database systems integrate some data mining tools within themselves. In that case, data mining system is highly coupled with database systems and both the systems use same memory and disk space. So, it can be seen that database technologies have contributed a lot in development of data mining.

Other disciplines such as neural network, genetic algorithms, fuzzy sets, information science, etc. also have been used to develop efficient data mining algorithms. So, it can be concluded that data mining is not an isolated subject. It is the confluence of multi-disciplinary subjects.

1.3 Data Warehouse and Data Mining

Data warehouse is considered to be a pre-processing step for data mining. According to W H Inmon, "A data warehouse is subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process" [Inm96]. According to this definition, main characteristics of a data warehouse are

- It generally deals with broad subjects like customer, sales, etc. It does not deal with day-to-day activities.

- Data warehouse integrates many heterogeneous sources of data such as relational databases, spreadsheets, flat files, etc.
- Data in the data warehouse are attached with some time element because it stores the historical data.

- Data in data warehouse is permanent. It does not require recovery, concurrency control, etc. Data are just accessed for decision making.

Data warehouse provides a platform on which data mining techniques are based on. It also provides various OLAP tools which can be integrated with data mining techniques. So, a clear understanding of data warehouse is must to understand data mining techniques.

### 1.3.1 Data Cube, Cuboid and View

Multidimensional data model is the basic data structure on which OLAP and data warehouse tools are based on. This model views data in the form of data cube. A data cube allows data to be modeled and view in multiple dimensions and it consists of cuboids/views. In-SQL terminology, cuboids/views are nothing but group-bys. As for an example, suppose, an organization keeps sales data with respect to \( t \), \( l \) and \( b \). Here, the data cube consists of eight possible group-bys: \( tlb \), \( tl \), \( tb \), \( bl \), \( t,l,b \) and none. Each individual group by is called sub-cube or cuboid or view.

DSS queries find the answers from the data cube. It may take long time due to huge size of data warehouse and the complexity of the query itself, which is not acceptable in DSS environment. The requirement of the query execution time is in the order of few seconds. Different techniques like query optimization and query evaluation techniques [CS94, GHQ95, YL95] are being used to deal with this problem. Different indexing techniques like bit-map index, join index are also used to reduce the query response time to a great extent. In data warehouse, the query response time largely depends on the efficient computation of data cube. However, creating data cube on the fly is very much time and space consuming.

One very useful technique used in data warehouse systems is partial materialization (pre-compute), which refers to materialization of some cuboids of a data cube, so that OLAP queries can be answered from these cuboids. However, the
big question is "Which cuboids/views should be materialized?" Partial materialization should i) select the cuboids to be materialized ii) use materialized views to answer the queries and iii) efficiently update materialized cuboids when data warehouse is updated. Selection of cuboids is not easy. Many factors are to be considered to select the cuboids. Among them, access frequencies of the queries, accessing cost of the cuboids, storage requirements, physical database design, etc. are important.

1.4 Challenges of Data Mining

There are many challenges which can be found to be bottlenecks in the development of data mining techniques. Among them, some of the main challenges are given below.

- **Large data set and high dimension**: Data mining algorithms have to deal with huge amount of data - both in terms of size and dimension. That is why, faster and efficient algorithms are required to handle these huge data. Some possible solutions are sampling, partitioning, parallel processing, etc.

- **User interaction and prior knowledge**: Data mining is an interactive and iterative process. Here, user's interactions at various stages are required. Domain knowledge may be used either in the form of high level specification of the model or at the more detailed level.

- **Over-fitting and assessing the statistical significance**: Data sets used for data mining are collected from various sources resulting in the spurious data sets. Therefore, some kinds of regularization methods and sampling techniques are used to design the models for data mining.

- **Understanding the patterns**: Discoveries should be made understandable to the human. The frequently used techniques are rule structuring, natural language processing, visualization of data, etc.

- **Non-standard and incomplete data**: The data can be missing and/or noisy.

- **Mixed media data**: Learning from data is represented by a combination of various media - numeric, symbolic, images and text, etc.
CHAPTER 1. INTRODUCTION

1. Management of changing data and knowledge: Data are often added, modified and deleted from database. The algorithms should take care of the changing patterns of the database.

2. Integration: Data mining being a part of the entire decision making process, it is required to integrate with database and final decision making process.

Researchers have tried and are trying to develop techniques to overcome these challenges. However, there are still ample scopes for exploring best possible solutions.

1.5 Applications of Data Mining

Data mining has wide area of applications. Some of them are highlighted below.

1. Loan Prepayment Prediction: The financial return of loans that a financial institution recovers, depends mainly on life-span of the loan. Data mining techniques help financial institutions predict number of loan repayments in a year as a function of interest rates, borrowers' characteristics, account data, etc. These information can be used to fix the parameters such as interest rate, fees, etc. to maximize profits.

2. Crime Detection: Data mining techniques can be used to solve cases which do not have obvious leads. Suppose crime data are recorded in a database. Then clustering techniques can be used to cluster the similar crimes based on modus operandi and other parameters. If some suspects can be connected to some cases of a cluster, all other crimes of the cluster might have been done by the same suspects. This way, it will be possible to clear up old cases and determine patterns of behavior.

3. Risk Analysis: Insurance companies can use data mining techniques to form clusters of customers depending on various risk factors so that when a new customer comes, he/she can be placed in one of the risk groups.

4. Target Marketing: It will be useless to send the information for a new product to all the customers. Companies can use data mining techniques
CHAPTER 1. INTRODUCTION

to find potential customers, who will respond to the new product mailing campaign. Companies can use data mining techniques to find buying patterns of the customers to promote their products.

- **Banking**: Banks can use data mining techniques in various ways to increase their profits and manage their business properly. Banks can use data mining techniques to detect withdrawal patterns of the customers, patterns of credit card use, identifying loyal customers, determine credit card spending patterns by the customer groups, etc.

### 1.6 Different Types of Data Mining Techniques

Databases contain different types data. So, different types of patterns exist in these databases. Data mining systems try to find these patterns depending on the database types and the users' needs. Basically there are two types of techniques - descriptive and predictive. Descriptive techniques find patterns which describe or characterize the data and the predictive techniques find patterns which are used to make prediction. Some important data mining techniques are described below.

- **Classification**: It refers to the classification of a data item into one of several predefined categorical classes.

- **Regression**: It refers to the mapping of a data item to a real-valued prediction variable.

- **Clustering**: It refers to the mapping of a data item into one of several clusters, where clusters are the natural groupings of data items based on similarity metrics or probability density function.

- **Association rule mining**: It describes association relationship among the attributes.

- **Summarization**: It provides a compact description of a subset of data.

- **Dependency modeling**: It describes the dependencies among variables.
CHAPTER 1. INTRODUCTION

- **Sequence analysis**: It models sequential patterns like time series analysis. The goal is to model the states of the process generating the sequence or extract and report deviation of trends over time.

All of the above techniques are useful in deriving various kinds of interesting patterns from databases and they have got different applications in different domains. However, only association rule mining techniques will be studied in the thesis because of wide area of applications of association rules in various domains. Association rule mining technique and some of its applications have been discussed in this chapter.

### 1.6.1 All Patterns Are Not Interesting

Data mining systems generate huge number of patterns. Obviously, all patterns are not interesting. Interestingness of a pattern depends on many factors such as understandability, validity or usefulness of the pattern. Many a time users give some thresholds to measure the interestingness of patterns. As for example, users may supply different support and confidence values to find interesting association rules. A pattern may also be interesting, if it validates some hypothesis given by user. Similarly, a pattern may be interesting depending on user's belief - an unexpected pattern may be interesting to certain kind of users. As for example, the pattern "80% people buy bread and butter together" is not interesting because it is expected, but "1% people buy bread and spoon together" may be interesting. Two important terms are generally used in relation to interestingness - completeness and optimization. Completeness refers to developing of data mining algorithms which can find all patterns and Optimization refers to developing of data mining algorithms which can find only interesting patterns. The following section discusses association rule mining and its applications.

### 1.7 Association Rule Mining

Association rules were first discussed by Agarwal et al. in 1993. It is often referred to as market-basket problem. Association rules find influence of one
set of items/attributes over another set of items/attributes in a database of transactions. One example may be "When people buy diapers, they also buy beer 60% of the time". Here, meaning of items and transaction depend on applications. Formally, it can be defined as a rule in the form $A_1, A_2, \ldots A_m \rightarrow B_1, B_2, \ldots, B_n$, where $A_i$'s and $B_j$'s are predicates or items. The rule can also be interpreted as conditional probability of occurring of $B_j$'s in a transaction is very high, given that $A_i$'s have already occurred in the transaction. Following are some examples of association rules.

1. Product($X$, bread) $\rightarrow$ Product($X$, butter). Here, items are the things bought by the customers and a transaction is the items bought together. The meaning of the rule is that customers generally buy bread and butter together.

2. age($X$, 20-30), income($X$, 10000 - 15000) $\rightarrow$ product($X$, mobile). Here, items are values of the dimensions from a data warehouse with three dimensions - age, income and product. The rule says that customers with age between 20 and 30 years and income between Rs. 10,000 and 15,000 tend to buy a mobile phone.

Association rules are evaluated by the measures such as support count, confidence, interestingness, etc. So, it can be viewed as multi-objective problem. However, it is viewed as single-objective problem in most of the applications. As far as applications are concerned, association rules have got numerous applications in various domains. The important applications are highlighted in the following subsection.

1.7.1 Some Applications

Association rules originated from market basket data. However, it is also widely being used in other databases and different problem domains. Some of the problem domains, where association rules have been used successfully, are business, engineering, medicine, telecommunication, etc. Association rules are also used for other data mining tasks such as prediction, modeling, decision support, etc. Some of the important applications of association rules are given below.
• **Market Basket Analysis**: Nowadays, it is very essential for the retailers to know the buying preferences and buying patterns of the customers. If a retailer knows the buying patterns of the customers of a region, he can formulate strategies to attract the customers by giving appropriate gifts with different products. Buying patterns also help a retailer organize the products in the shelves so that customers find the related products together. Thus, association rules can help retailers get the buying patterns, preferences of the customers, which in turn will increase sales.

• **Financial Services**: Association rule mining plays a big role in financial sector. Financial experts use association rules to develop investment models, risk models in stock markets, etc. Associating rule mining systems have been used successfully in stock selection, claims processing by insurance companies, currency trading, etc.

• **Fraud Detection**: With the increase of the use of electronic money like debit cards, credit cards, etc., fraud detection has been one of the prime tasks of the crime branches. Association rule mining can find using patterns of cards by card holders, and crime branches can use these patterns to detect the frauds.

• **Partial Classification**: Conventional classifiers may not be effective in a database, where most of the values of the attributes are missing. Association rules can be used to solve these kind of problems. Association rules can be used to see if one set of attributes are related to another set of attributes. In that case, one set of attributes can be replaced by another set of attributes with most of values are present. As for example, if result of one complex and costly medical test can be predicted from a set of simple and cheap medical tests, doctors can prescribe the simple medical tests instead of complex medical tests.

### 1.8 Motivation

Huge amount of data are collected by regional sale system, telecommunication system, World Wide Web and other data collecting tools. These databases con-
tain many useful patterns and knowledge, which can be used by decision makers and analysts to take appropriate decisions. There exists some data analyzing tools. These tools generally use statistical approaches and cannot deal with large data. Data mining techniques overcome these disadvantages because data mining techniques are meant to deal with large amount of data. There are different data mining techniques meant to find different patterns from large databases. All the techniques are useful in their respective domains. However, association rule mining is more interesting and challenging because of its wide areas of applications.

Generally, association rule mining is two step process [AMS+96]. First step finds frequent (or large) itemsets and second step finds association rules among the frequent itemsets. Between them, the first step is more challenging and interesting. That is why, most of the research works concentrate on the first step i.e. finding frequent itemsets from a large databases. For these reasons, mostly frequent itemsets finding techniques have been studied and analyzed in this thesis.

Finding frequent itemsets is even more complex in dynamic databases. An itemset which is frequent in the existing database, may not be frequent when some more records are added to the database, some records are deleted from the database or some records are updated. So, special algorithms are required to deal with such situations. The situation becomes more complex in distributed environment, where data is distributed in different geographic locations/sites. Here also special algorithms are required to deal with distributed environment. The thesis has reported a distributed algorithm to find frequent itemsets in distributed dynamic database.

Most of the databases contain lots of unnecessary or redundant features. These redundant features degrade the performance of machine learning algorithms. So, removing redundant features from a database is considered to be a major preprocessing step in many machine learning algorithms. The concept of frequent itemsets can help to remove redundant features to a great extent. This idea has been explored to remove redundant features from a database.

There exists some algorithms to deal with the above problems. However, the algorithms are not efficient and scalable. This has motivated us to enhance some
existing algorithms and to develop some new algorithms to address the above
issues. Main motivations are listed below.

- **Unnecessary candidates:** It has been found that algorithms generates too
  many unnecessary candidate sets, which is the main reason for exponential
  execution time. It has been tried to reduce the unnecessary candidates.

  **Contribution:** Apriori and BiAssocRule algorithms have been modified
  and probability has been used to reduce candidate sets and execution time.

- **Existing horizontal partition based algorithms are not effective:** Partition
  algorithm partitions a large database horizontally and then find the fre­
  quent itemsets. It has been observed that this technique is not effective
  in databases with larger dimensions and execution time increases with the
  increase in the number of partitions.

  **Contribution:** An algorithm has been developed by using vertical partition.
  It has been found that this technique is more effective for databases with
  larger dimensions and execution time decreases with the increase of number
  of partitions. Vertical partitioning technique also has been used with FP­
  growth [HPY00] algorithm to increase the performance of FP-growth for
  databases with large dimensions.

- **Feature selection is an important step in KDD and machine learning tech­
  niques:** KDD process and machine learning techniques take too much time
  when applied on databases with irrelevant features. So, relevant feature
  selection is one of the important tasks in machine learning techniques.
  There exists some algorithms to select relevant features and they use dif­
  ferent criteria to decide whether a feature is relevant. However, the existing
  algorithms are not good enough in terms of both feature selection and ex­
  ecution time. In addition to that, it has been found that there was very
  little work done to select relevant features using frequency count (support
  count) of the features.

  **Contributions:** One algorithm has been developed to select relevant features
  by using frequency counts (support count) of the features.

- **Lack of efficient algorithm to find frequent itemsets in dynamic database:**
  Finding frequent itemsets in dynamic database is a challenging task. There
exist some algorithms to find frequent itemsets in dynamic databases. However, they suffer from the drawback of having to scan the whole database repeatedly, which in turn, increases the execution time of the algorithms. So an economic solution is required to solve this problem.

Contribution: One algorithm has been proposed, which uses two levels of border sets. The main advantage of the algorithm is that it does not have to scan the whole database repeatedly.

- Lack of algorithm to find frequent itemsets in distributed dynamic databases: Nowadays, most of the databases are distributed. However, there has been a little work to find frequent itemsets in distributed dynamic databases. Again, existing algorithms to find frequent itemsets in dynamic databases cannot be used directly in distributed environment. So, some algorithms are required, which can find frequent itemsets from distributed dynamic databases.

Contribution: A distributed algorithm has been developed to find frequent itemsets for distributed dynamic database.

- Selection of useful cuboids/views to be materialized in data warehouse systems in minimum possible time is very important. However, there are some algorithms to select views to be materialized. These algorithms have used very complex techniques to select the views. So, these algorithms take too much time to select the views.

Contribution: A very simple algorithm has been proposed to select the views. The algorithm selects views based on the concept of density. The algorithm also uses the frequency(support) of the sub-views to calculate the benefits of the views.

1.9 Scope of The Thesis

The thesis embodies an exhaustive experimental study on some of the popular existing frequent itemsets finding algorithms and feature selection algorithms in the light of real and artificial datasets. It also includes some of the enhanced versions of the algorithms such as Apriori, BitAssocRule, Borders, etc. A detailed
CHAPTER 1. INTRODUCTION

comparative study of those enhanced versions with their respective counterparts are also included to establish the superiority of the algorithms. Concentration has been given on the following areas.

- Partitioning is one of the important techniques used in data mining. A vertical partition based frequent itemset generation algorithm has been developed and a detailed comparative study of both these approaches are given.

- Feature selection is one of the important aspects in machine learning techniques. It has been shown how frequency count (support) of the attributes can be used to select relevant features in a database.

- Nowadays, most of the databases are distributed and dynamic. So, distributed algorithms also have been analyzed and studied in detail. One distributed algorithm has been proposed for distributed dynamic databases.

1.10 Organization of The Thesis

The thesis is organized as follows. Chapter 2 reports the related works. Chapter 3 reports some of the popular existing association mining algorithms. It also reports a modified and faster version of an existing robust association mining algorithm. Chapter 4 discusses the partitioning approach in the frequent itemset generation. It also introduces a vertical partitioning approach for frequent itemset generation. A comparative study between both these approaches is also reported in this chapter. Chapter 5 covers a crucial issue of association rule mining i.e. rule mining in dynamic databases. It reports experimental analysis of a popular dynamic association mining algorithm and also it reports an enhanced and distributed version of the algorithm. Chapter 6 is an attempt to report some useful and popular algorithms for feature selection. It also reports a novel feature selection algorithm. A detailed comparative study is also reported in the chapter. Chapter 7 is dedicated to a potential application of association mining techniques. In Chapter 8, concluding remarks and future works are given.