Chapter 1

Introduction

1.1 Data Mining

With the advancement in sophisticated technologies and tools, the capabilities of generating and collecting large amount of data has increased tremendously for the last two decades. The computerization of several government and business activities and increasing use of bar codes for commercial products has also contributed to the explosive growth of data. This in turn demands for more powerful and affordable tools and technologies that can wisely transform the stored data into useful information which can be used for planning and decision making. Consequently data mining has become a key research area, which captured the interest of a large community of researchers [MIN1996].

Data mining is a non-trivial process of extracting interesting information and hidden patterns from large data repositories like data warehouses, XML repositories or relational databases.
It is sometimes called knowledge discovery because it is the core process in Knowledge Discovery in Database (KDD) [ZHO2003]. This is the process of discovering interesting knowledge regularities, sequential patterns or rules from relevant sets of data and then investigating them from different angles so as to create a more rich and reliable source for knowledge generation and verification. The discovered knowledge can be applied for information management, query processing, decision making, process control and many other applications. Being a core part of KDD some people treat data mining as a synonym for KDD processes. Usually there are three processes in KDD. The first one is called preprocessing which targets to perform data cleaning, integration, selection and transformation over the data before applying data mining technologies. This preprocessing is necessary because the data may come from different sources and may have inconsistencies and redundancies. After that the process of data mining is executed. It applies different algorithms and methods to produce hidden knowledge. Then a post-processing step is applied that evaluates the mined results according to the user requirements and domain knowledge [HAN2000]. The KDD processes are shown in Figure 1.1.

![Figure 1.1: Knowledge Discovery in Databases (KDD) process](image-url)
The kind of knowledge that can be discovered depends upon the data mining task applied. Broadly speaking, there are two categories of tasks: descriptive data mining tasks that describe the general characteristics of the existing data and predictive data mining tasks that attempts to do predictions based on inference over available data. Association rule mining is one of the popular descriptive data mining tasks that find extensive use in marketing and retail communities in addition to many other fields [AGR1993].

1.2 Applications of Data Mining

Data mining has been used extensively by many organizations. Many of these organizations are combining data mining with things such as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions. Popular data mining applications are:

- Data mining is used in financial data analysis for making assessment of the credit rating, loan payment prediction, customer credit policy analysis, classification and clustering of customers for targeted marketing and detection of money laundering.
- In Telecommunication Industry, data mining is used for multidimensional analysis of telecommunication data, fraudulent pattern analysis, identification of unusual patterns, and enhancement of mobile telecommunication services.
- In biological data analysis, data mining is applied for semantic integration of heterogeneous, distributed genomic and proteomic databases, discovery of structural patterns and analysis of genetic networks. Identifying co-occurring gene sequences
and linking genes to different stages of disease development is another task that is made easy with the help of data mining technologies.

- Another field where data mining has paved its way is intrusion detection. Data mining in the area of intrusion detection is popular and the reason for this is two-fold. First, the volume of data dealing with both network and host activity is so large that it makes it an ideal candidate for using data mining techniques. Second, intrusion detection is an extremely critical activity.

- Data collection and storage technologies have recently improved, so now-a-days scientific data can be amassed at much higher speeds and lower costs. This has resulted in the accumulation of huge volumes of high-dimensional data, stream data, and heterogeneous data, containing rich spatial and temporal information. Consequently, scientific applications are shifting from the “hypothesize-and-test” paradigm toward a “collect and store data, mine for new hypotheses, confirm with data or experimentation” process.

1.3 Techniques of Data Mining

During last few decades, most of the commercial activities like banking, shopping, trading etc. have made a significant move towards computerization. Also most of the data related to medical domain, crime detection and even personal information are now being computerized. The invention of powerful database systems and the availability of simple tools to access and manage such databases alleviated the trend of extensively large data repositories. This in turn demands equally powerful and easily accessible techniques for extracting information and knowledge. As a result, data mining has become an interesting research area [FAY1996].
Goal of data mining is to mine useful, comprehensive, previously unknown and interesting information from the stored data. By comprehensive, we mean that the mined information should be wide-ranging that is hidden in the data and provides some facts and information that can be used further for management decision making and process control [FRA1991]. Data mining is the area which is continuously spreading its roots in all fields ranging from bio-informatics [CHE2005] to web-based mining [SRI2000] and from image appreciation [XU2004] to wireless remote sensing [ZHA2006]. Almost all organizations, be it a large scale business or a small private firm, need to figure out associative relationship among the collected data for improving their marketing strategies and polices.

Data mining is a well researched area and several data mining techniques and systems have been developed over time. These methods and techniques can be categorized based on the type of databases to be considered, the type of techniques to be utilized and the type of knowledge to be discovered [MIN1996]. A relational data miner mines knowledge from relational databases whereas an object-oriented miner mines knowledge from object-oriented databases. Based on the tools used, data mining systems can be autonomous knowledge miner, data-driven miner, query-driven miner or interactive data miner. Also these miners can mine several different types of knowledge including association rules, classification rules, characteristic rules, and clustering analysis.

1.3.1 Association Rule Mining

Association rule mining is one of the most important and well researched technique of data mining and was first introduced by Rakesh Agrawal et al. [AGR1993]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of
items in the transactional databases or other data repositories. The aim is to drive a set of strong association rules in the form of \(A_1 \land A_2 \land \ldots \land A_m \rightarrow B_1 \land B_2 \land \ldots \land B_n\), where \(A_i\) (for all \(i\) in \(\{1,2,\ldots,m\}\)) and \(B_j\) (for all \(j\) in \(\{1,2,\ldots,n\}\)) are sets of attributes from the relevant data sets in a database. As mining association rules requires scanning the database again and again to find different association patterns, the amount of processing could be huge and performance improvement becomes an essential concern. Efficient algorithms for mining association rules and several methods for further performance enhancements have been developed. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. There are two important measures for association rules, support and confidence. Since the databases are large and the users concern about only those frequently occurring patterns, usually thresholds for support and confidence are predefined to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively. However additional constraints of interesting rules can also be specified by the users.

### 1.3.2 Classification Rule Mining

Classification is the process that automatically builds a model by detecting common characteristics among a set of objects or data in a database so as to predict the classification or missing attributes. The objective of classification is to first analyze the training data and then develop an accurate description for each class. It is a two step process. In the first step which is also called supervised learning, a model is prepared on the basis of training data to portray the features of a set of data classes or concepts. In the second step, the model is used to predict the classes of future objects or data. There are several techniques for classification.
Classification by decision tree has been well researched and several algorithms have been
designed including statistical approaches [MIN1996] and rough set approach for
classification. An interval classifier has been proposed by Rakesh Agrawal et al. [AGR1992]
to reduce the cost of decision tree generation. The neural network approach and rule
extraction in databases has also been studied a lot.

1.3.3 Clustering

Clustering is another technique similar to classification. However unlike classification,
clustering is an unsupervised learning process. The process of grouping physical or abstract
objects into classes of similar objects is called clustering or unsupervised classification.
Clustering is the technique of partitioning of a large set of objects into smaller components to
simplify design and implementations. It follows the “divide and conquer” methodology. Data
clustering identifies clusters or closely populated regions according to some distance
measurement. Given a large set of multidimensional data points, the data space is usually not
uniformly occupied by the data points. Data clustering identifies the sparse and the crowded
places and hence discovers the overall distribution patterns of the data sets. As a branch of
statistics, clustering analysis has been studied extensively for many years. It is also studied in
machine learning [FIS1987], spatial database [BEC1990] and data mining [FAY1996]
[EST1995] areas with different emphases. The distance based approaches assume that all the
data points are given in advance and can be scanned frequently. For each clustering decision
they inspect all the data points or all currently existing clusters equally no matter how close
or far away they are. In machine learning, the class of an object is not pre-specified. Hence
one needs to define a measure of similarity between objects and then apply it to determine
classes. Hence clustering is often called unsupervised learning. The method of clustering
analysis in conceptual clustering is based on probability analysis. Such approaches make the assumption that probability distributions on separate attributes are statistically independent of each other. But this assumption makes it very expensive approach to update and store the clusters [FIS1987].

1.3.4 Data Generalization and Summarization

Data generalization and summarization deals with general characteristics or a summarized high-level view of the data stored in the databases. Data and objects in databases usually contain detailed information with them. It is often desired to summarize the data present at different level of abstraction and present it at a high concept level. This requirement has demanded for an important functionality in data mining which is called data generalization. It is a process which abstracts a large set of relevant data in a database from a low concept level to relatively high levels. There are two major approaches for efficient and flexible generalization viz. data cube approach [GUP1995] and attribute-oriented induction approach. Generalization and summarization can be performed on multiple dimensional data cube by roll-up or drill-down operations

1.3.5 Pattern Based Similarity Search

Pattern-Based similarity search is most popular in temporal or spatial-temporal data mining. Queries in such databases often require information which is related to distance from a particular object or the objects that are in a user-specified range from other related objects. These kinds of queries require sequence matching or pattern similarity matching tasks. Two types of similarity queries have noted so far namely whole matching and subsequence matching. Several approach have been proposed for similarity search which differ in
similarity measure chosen, basis of comparison used that can be time domain or transformed
domain and generality of the approach. Fourier analysis and Hierarchy Scan are some of the
major approaches used for feature extraction and similarity matching.

1.3.6 Path Traversal Patterns Mining

Today most of the commercial activities like banking, shopping and ticketing are
computerized and there is an environment where a collection of linked documents or objects
are provided for interactive access to these services. For example, World Wide Web is a vast
collection of documents where users navigate from one object to another through links or
other facility provided for desired information. Obviously, understanding user access patterns
in such cases will not only help in improving system design but will also help in better
marketing decisions. Capturing user access pattern in such environments is called path
traversal pattern mining. There are still limited approaches for path traversal pattern mining
because such information providing services are in their formative years. However as these
services are becoming increasingly popular, there is a growing demand for capturing user
traversing behavior for improving quality of services [CAT1995].

1.4 Issues and Challenges in Data Mining

Data mining is a collection of diverse technologies and tools. A proper application of these
techniques is not an easy task. In order to perform successful data mining, the acquaintance
with the anticipated issues and challenges is must. While performing data mining, several
problems may arise that may not only hinder the effective implementation of the system but
can also cause serious hazards. Thus it becomes necessary to have a view of issues and
challenges that one may face while applying data mining technologies
1.4.1 Handling of different types of data

There are several types of data and databases available in practice. Some databases contain complex data such as structured data, hypertext, multimedia data, spatial and temporal data, and legacy data etc. Thus it is not realistic for one data mining system to handle all kinds of data. Specific data mining system should be constructed for knowledge mining on specific kind of data so that data mining system could be able to perform efficient and effective data mining.

1.4.2 Efficiency and scalability of data mining algorithms

To utilize the true power of data mining technologies, the algorithms should be designed in such a way that they can be easily and efficiently scaled to large databases. The running time of data mining algorithm must be predictable and acceptable in case of large databases.

1.4.3 Analysis of data mining results

Different kind of knowledge can be extracted from a large amount of data. But the main concern is that the discovered knowledge should accurately represent the contents of the database and is useful to the users. The uncertainty, noise and exceptional data should be handled sophisticatedly. Also the quality of the discovered knowledge should be analyzed systematically by cross verification and using simulative models.

1.4.4 Interactive mining at multiple abstraction levels

Interactive discovery in data mining system should be encouraged. This means that the user should be able to refine the data mining request, dynamically change data and process and can view the results at multiple levels of abstraction. This facilitates the user in exploring further interesting and hidden knowledge in the database.
1.4.5 Protection of privacy and data security

When data can be viewed from different angles and at different level of abstraction, it opens the doors for potential data security and privacy threats. With increasing interest in on-line activities, most of the consumers’ data has been stored and recorded for the purpose of mining. Thus, keeping this information secure and preserving the privacy has become a challenge. It is important to ensure that the knowledge discovery may not lead to an invasion of privacy.

Efficient data mining techniques address these issues in a proficient manner in addition to mining of valuable knowledge. Researchers are engaged in devising methods that can handle these challenges and provide comprehensive and useful information to the users and policy makers. Evolutionary algorithms, neural networks, genetic algorithms and swarm intelligence are the hot areas that reported outstanding performance in data mining. Swarm intelligence is the field that is capturing the interest of large community of researchers for innovation of new techniques for handling several complex problems. The present research work is also focused on the application of swarm intelligence in the field of data mining.

1.5 Swarm Intelligence

Swarm intelligence is a subfield of artificial intelligence and is a new area of research which studies the cooperative behavior of swarms. Swarms inherently use forms of decentralized control and self-organization to achieve their goals [DOR2006]. The researchers in this field study the collective behavior of systems composed of simple agents, usually limited in capacity, interacting locally with each other and with their environment. In such biological swarms, ants colonies, bees hives, termite, flock birds and fish schools are the most popular
ones. The individual agents in such organizations are by no means complete engineers, but they are simple creatures with limited abilities and have restricted means of communications. Yet they collectively solve complex problems like searching shortest path to food source and predator evasion etc. The agents use simple local rules to govern their actions and through the interactions of the entire group, the swarms achieve their objectives. This self-organization emerges from the collection of actions of the group [GRO2006]. The success and efficiency of swarms to solve difficult problems in nature, has attracted the researchers in computer science to use the principle followed by these swarms for developing swarm-based systems. In recent years, the field of swarm intelligence has received far-reaching attention in research.

1.5.1 Branches of Swarm Intelligence

By impersonating the behavior of biological creatures in computing methodologies, techniques that are robust, scalable, easily distributed and are applicable to hard optimization problems, have been developed. Since last few years, an increasing use of nature inspired computing techniques has been noted in engineering applications. Successful application of swarm intelligence include travelling salesman problem [DOR1996], scheduling [BLU2005], robotics [DOR2009], packet routing [CAR1998] and many more. The most popular swarm intelligence meta-heuristics are Ant Colony Optimization, Particle Swarm Optimization and Prey Model. Among these ACO and PSO have received lots of attention. The Figure 1.2 shows the increase in the popularity of these techniques since last two decades. This is an estimation of the papers written on these subjects as indexed by the Web of Science.
1.5.1.1 Ant Colony Optimization

Ant Colony Optimization (ACO) was introduced in the early 1990s as a novel technique to solve hard combinatorial optimization problems. ACO is an optimization algorithm inspired by the collective foraging behavior of ants to find and exploit the food source that is closest to the nest. ACO is based on cooperative search paradigm that is applicable to the solution of combinatorial optimization problem. Communications among individuals or between individuals and the surroundings is based on the use of chemicals formed by the ants called pheromone [PAR2011]. The different types of Ant Colony models are Ant System for Traveling Salesman Problem (TSP), Ant Miner, Ant Quality, Ant-Cycle, Max-Min ACS, Ant...
System (AS), Ranking Model and Ant Density. To apply ACO for solving a problem following aspects are to be specified:

- A proper definition of the problem that needs to be solved by the ants so that strategies can be made to incrementally build solution to the problem based on some transition rule, amount of pheromone on the path and some local information.
- A method to construct valid solutions that can be considered legally permissible in the real world situation.
- A heuristic function that should be designed based on the problem to measure the significance of the items/terms that can be added to the current solution so as to give right direction to the process of finding optimal solution.
- A rule for updating the pheromones on certain paths that leads to good solution to the problem and fine-tuning the pheromone-trail.
- A transition rule that is based on the heuristic function and amount of pheromone on a path to decide about the movement of ants and construction of solution.

**Real Ant System**

The visual sensitivity of many ant species is not as much developed and there are ants that are completely blind. The communication between them is based on the use of chemicals produced by them. These chemicals are called Pheromones. While walking from the food source to the nest or vice-versa, ants deposit pheromones on the ground, forming a pheromone trail. Ants can smell the pheromone and they tend to choose, probabilistically,
paths marked by strong pheromone concentrations. So more the ants follow a trail, the more attractive that trail becomes to be followed by other ants [DOR1996]. The basic idea was designed and run by Deneubourg et. al. in 1990 [DEN1990], who used a double bridge connecting a food source and the nest. One of the bridges was twice as longer as the other. Initially all ants choose the outgoing path randomly. Some ants choose the short path and reach the food source sooner than the other ants choosing the longer one. The ants reaching the destination through the short path returned to the nest before the others on the long path. Thus the amount of pheromone on the smaller path tend to increase faster than the longer path and results in increased probability of choosing that path by the foragers. Since ants prefer to follow trails with larger amounts of pheromones, eventually all the ants converge to the shorter path.

Artificial Ants

An artificial ant simulates a real ant and a set of artificial ants develops mechanisms of cooperation and learning. Ant Colony Optimization (ACO) was developed by Dorigo et. al. [DOR1996] and it was first used to solve combinatorial optimization problems. Artificial ants are characterized as agents that imitate the behavior of real ants. However, it should be noted that an artificial ant system has some differences in comparison with real ants which are as follows:

a) Artificial ants have memory.

b) They are not completely blind.

c) They follow a discrete time system.
The main idea is that when an ant has to select a path among several available paths, the ant chooses the one which is chosen more frequently by other ants in the past. Thus path with larger amount of pheromones is the shorter path and chosen by most of the ants. The Ant System works in two major steps- solution construction and the pheromone updation as described below:

**Solution Construction**

A solution is constructed in a probabilistic way. Initially a fixed amount of pheromone is deposited on each path. At the very first stage each ant starts randomly. The selection of next path to be followed or an item to be added to the current solution is determined by the ants based on two major things:

a) A problem dependent heuristic function that may determine the quality or predictive power of the item under consideration.

b) The amount of pheromones deposited by ants in the past.

**Pheromone Updation**

Initially equal amount of pheromone is deposited on each path. As each ant moves, it tries to build a best possible solution to the problem. When the ant completes construction of the solution, the amount of pheromone must be updated on each item in the path. When an ant completes its trip, the quality of the solution constructed is checked. The increase in pheromone can be proportional to the quality of the solution; better the solution, larger the increase in the amount of pheromone. This corresponds to increasing the chances of selecting the items in the current solution by other ants in the future. In the real ant system, the pheromone tend to evaporate with time, thus the artificial ant system also needs to decrease
the amount of pheromone on the paths that are not part of the constructed solution. As per the need, the decreasing rate can be set to a constant or it can be determined by the frequency of the items used during past.

1.5.1.2 Particle Swarm Optimization

Birds and fishes exhibit fascinating example of collective behavior. They move in flocks and search for food. Birds randomly look for food while keeping in view the movement of other birds in the flock and following the one closest to the food. This movement in the form of herd or flock has several advantages. This increases the effectiveness of searching for food in addition to local information; the overall swarm demonstrates a fluid coherent motion which seems perfectly synchronized [KEN1995]. Main principles of the collective behavior of swarm intelligence are given in Figure 1.3. These are:

- Velocity Matching: attempt to match velocity with nearby flock mates.
- Locality: the motion of each agent is only influenced by its nearest flock mate. Vision is considered to be the most important sense for flock organization.
- Homogeneity: every agent in flock has the same behavior model. The agents move without a leader, even though temporary leader seems to appear.
- Collision avoidance: avoid collision with nearby flock mates.
- Flock Centering: attempt to stay close to nearby flock mates.
ACO is designed to solve discrete optimization problems but PSO is designed to solve continuous optimization problems. In the beginning, PSO was used for simulation of choreography of bird flocking but the promising performance of the approach has made it a popular technique for continuous optimization meta-heuristic. In PSO, the birds are represented by a population of particles, called a swarm [KEN1995]. Each particle has a central location and velocity within the search space and it represents a solution. As birds move through the environment in search for food and escaping predators, particles move through the search space in search of high quality solutions. The search process is directed by the best solution found so far by the particle itself and all the neighboring particles.

The basic PSO algorithm is straightforward. Initially, the particles are spread randomly all over the input space. Each particle stores the personal best solution it has ever visited with its corresponding fitness value. The fitness value is an indicator of the quality of the solution. In addition to the personal best solution, the best solution among all the personal best solutions

Figure 1.3 Main principles of collective behavior of swarms
among all the particles in a neighborhood is also preserved with its fitness value. The fitness value of this best solution is higher than all the solutions in the neighborhood. PSO algorithm update the particles’ location and velocity according to the standard equations that are defined precisely for each variant of PSO viz. Linearly Decreasing PSO (LDPSO), Constricted PSO (CPSO) and Comprehensive Learning PSO (CLPSO). In each iteration, vectors of random numbers are generated for each particle. As such, particles move stochastically in the direction of their own best solution and the neighborhoods’ best solution.

In the original PSO algorithm, the velocity of each particle is kept within the range \([-V_{\text{max}}, +V_{\text{max}}]\) to avoid these vectors of random numbers growing to extremely large numbers. Updation of the particles is halted when sufficiently good performance is obtained.

Since the beginning of PSO, several variants have been introduced and each variant defines different velocity update rules. Discrete Particle Swarm Optimization [KEN1997], Linearly Decreasing PSO [SHI1998], Constricted PSO [CLE2002] and Comprehensive Learning PSO [LIA2006] are some of the best known variants of PSO that are proposed over time by researchers.

In DPSO, the particle position is discrete but its velocity is continuous. In LDPSO, a time-decreasing function is added as a multiplicator in the first term of the velocity update rule. The CPSO adds a constriction factor in the velocity update rule to avoid unlimited growth of the particles’ velocity. On the other hand, CLPSO combines all the particles’ historical best information to update the particle velocity. This strategy aims at avoiding premature convergence by enabling the diversity of the swarm to be preserved. The chaotic PSO is another variant introduced by Alatas et al. [ALA2009] which uses a chaotic number
generator each time a random number is needed by the classical PSO algorithm in order to improve global searching capability by escaping the local solutions.

1.5.2 Swarm Intelligence for Data Mining

The popularity and admired performance of swarm intelligence techniques has also prompted the researcher to make use of these approaches for more efficient data mining. The swarm intelligence based data mining is categorized into two approaches viz. effective search and data organizing. The first category consists of techniques in which individual agents wanders through a solution space and try to find solutions for data mining task at hand in some effective manner that combines both exploration and exploitation. In ACO, the search space is discrete and a solution is defined implicitly by the path taken by an ant. For PSO, the search space is continuous and the locations within the search space are updated explicitly. The second category called data organizing consists of approaches in which the swarms move data instances that are placed on a multi-dimensional feature space in order to create a suitable clustering or mapping solution of the data. The major techniques in this category are ant-based sorting and prey models [MAR2011]. In both categories, the approaches start by defining the environment in which the swarm individuals will work, followed by an instantiation of problem parameters. They first initialize the search space parameters such as pheromone levels, number of ants and after that the individual agents commence their task of creating data mining solutions, optimizing some predefined objective functions.

1.5.2.1 Data Mining Applications of ACO

ACO in data mining has been used primarily for supervised classification. Although it has also been used for clustering, the bulk of the research addresses classification. The most
Chapter 1  Introduction

prominent use of ACO for classification rule discovery was made by Parpinelli in AntMiner in 2002 [PAR2002]. Several variants of AntMiner have been proposed for improving the quality of classification rule discovery. AntMiner2 [LIU2002], AntMiner3 [LIU2003] and AntMiner+ [MAR2007] are the different versions of AntMiner. The reported benchmarking experiments revealed that the accuracy of AntMiner+ is superior to that obtained by the other versions and competitive or even better than the results achieved by the other classification techniques such as C4.5 and RIPPER. In addition to classification rule discovery, ACO has also been used for rule extraction. TACO-Miner proposed by Özbakir extracts rules from neural networks or support vector machines. The aim is either to explain the non-linear model and as such open up the black box or improve the performance as noise can be removed in this manner [ÖZB2009].

The ACO meta-heuristic has also been applied to clustering and sorting. The approach is similar to the use of ACO for TSP problem, where paths are constructed by the ants trying to minimize the total distance. In TSP, the distance between cities is minimized but in clustering, the nodes are data instance and the distance between them is defined by their dissimilarity. However ACO has been widely used for classification and clustering, but the use of ACO for association rule discovery is rare. The major problem that remains for ACO based techniques is the need for discretization of all variables and the learning time. Because ACO is a discrete optimization technique, each variable can only have a limited number of values. This is a problem as data mining is performed on huge data stores where each variable can have many values. However the discretization technique can be used for handling the continuous variables but still the efficiency issues matters a lot. The learning time of the ACO based techniques is quite long as compared to other techniques. Although it
is stated that for most applications, the training duration is not an issue and may last up to hours, for some real-time applications it is of key importance.

1.5.2.2 Data Mining Applications of PSO

PSO has found several applications in data mining including classification and clustering. PSO is used for classification rule discovery in the work reported by Sousa et al. [SOU2004]. This approach has made use of a sequential covering algorithm with each rule being a conjunction of terms. Sensitivity and specificity are used for evaluating the quality of the rules. Three variants of PSO viz. DPSO, LDPSO and CPSO have been tested and results show potential increase in efficiency. However there is the drawback of increase in execution time. A second application of PSO to classification is developed with a different model, its representations and with extending benchmarking experiments. LDPSO is used by De Falco et al. [FAL2007] in their algorithm that see the classification problem as finding the optimal coordinates of \( c \) centroids in the \( m \)-dimensional search space where \( c \) and \( m \) are number of classes and variables respectively. PSO based classification is quite competitive, yielding the best results and obtaining an average ranking of four and having an execution time similar to other techniques.

Another very important application of PSO is in the domain of cascading classifiers. Cascading classifiers have been used to solve pattern recognition problems in the last years. The main motivations behind such a strategy are the improvement of classification accuracy and the reduction of the complexity. The issue of class-related reject thresholds for cascading classifier systems is an important problem. It has been demonstrated in the literature that class-related reject thresholds provide an error-reject trade-off better than a single global
threshold. The use of PSO for clustering was made by Van der Merwe [VAN2003]. Using two artificial and four UCI datasets, the author demonstrated the potential in terms of improved quantization error. A drawback is that, as with k-mean clustering, the number of clusters needs to be predefined. But the experiments are limited to low dimensional data only. Clustering high dimensional data is addressed with PSO by Lu et al. [WAN2010].

1.6 Problem Formulation

Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) are two key technologies of swarm intelligence which have proved their efficiency and effectiveness in several data mining tasks. Classification rule discovery and clustering are most prominent areas that have shown notable improvements with the application of ACO and PSO techniques. But there has been a lack of use of these techniques in the field of association rule mining. However limited work has been done by some researchers [VIJ2011][SAD2012][KUO2007], but it is not fully utilized. The association rule mining may be considered as an optimization problem where the objective can be to optimize any of the interestingness measures. But due to inadequate research in this direction, the importance of these swarm intelligence technologies in association mining is underutilized. The present research work is thus focused on developing association rule mining strategies that apply Ant Colony meta-heuristic for generating high quality and comprehensive rules. Efforts have been made to take the advantage of positive feedback mechanism followed by ants and some new approaches have been suggested that can generate rules more efficiently.
1.7 Objectives of the Thesis

To address the problem discussed above the following objectives have been set for this thesis:

i). To analyze the performance of existing methods for mining association rules in data mining.

ii). To find the suitability of Ant Colony Optimization Technique in mining association rules.

iii). To develop new approaches for mining association rules using Ant Colony Optimization.

iv). To compare the performance of the proposed approaches with other existing approaches through experiment in some specific domain.

1.8 Organization of the Thesis

The thesis is organized into five chapters and four appendices:

Chapter 1 gives general idea about the broad area of the present research work viz. Data Mining and Swarm Intelligence. The benefits and challenges of data mining are explored and different techniques of data mining are outlined. The two major branches of swarm intelligence namely Ant Colony Optimization and Particle Swarm Optimization are described. The applications of these techniques in data mining are also summarized.

Chapter 2 presents the survey of different approaches for association rule mining available in literature. The different ACO based approaches used for data mining have also been
surveyed. The aim is to provide an overview of these approaches and to summarize the problems and issues related to each approach.

Chapter 3 describes the association rule mining technology in detail. The different approaches for mining association rules proposed in literature are discussed with their pros and cons. Some methods for increasing the efficiency of association rule mining have been discussed outlining the recent advancements. A novel approach using graph has been proposed that generates frequent itemsets for generating association rules.

Chapter 4 introduces the concept of Ant Colony Optimization. The use of ACO for finding the optimal path from a graph of nodes is discussed and a novel approach is proposed to find frequent itemsets from a database represented as graph using the ACO based approach.

Chapter 5 proposes two more approaches that make use of ACO for generating association rules. The first proposed approach efficiently generates frequent itemsets from large database repositories. The second proposed approach generates association rules directly without generating frequent itemsets. Experimental evaluation of the proposed approaches has been performed on the retail data set taken from UCI repository and the results are compared with that of the other existing similar approaches.

Chapter 6 concludes the thesis and presents the future research directions for extension of the presented work.

Different datasets taken from UCI repository used for evaluation of the proposed approaches are given in appendices from Appendix-A to Appendix-E. At last, the thesis ends with a complete list of references given in lexicographical order.