CHAPTER 1

INTRODUCTION

Data Mining popularly known as knowledge discovery in database (KDD) are the automated or extraction of patterns denoting knowledge in large database, datawarehouse, internet and other information data repositories. Data Mining is a multidisciplinary field since it works on statistics, machine learning, AI, high performance computing etc. It seems that data mining is technique for discovering interesting patterns from data in various applications. It turns large collection of data into knowledge. Noted that Data Mining differs from Querying large database because the mining principle brings out hidden information where querying brings out the data defined in the database. Thus mining yields processed form of data called information.

1.1 MINING AN EVOLUTION OF IT

A large Collection of database is evolved from databases and database management by Industry in the development of data collection, creation, data management data analysis. The development of data collection and creation supported the development of large data storage, querying data and performing transaction processing. The database and IT helped to produce powerful database systems. The research and development in database system using hierarchical database lead to relational database systems, data modeling tools. For user interfaces, efficient methods like OLTP were developed to Query databases.

Next Querying database technology moved to advanced database systems, data warehousing, web based databases, datamining and database analysis. The systems were in corporate such as Object-Oriented, Object-relational, and Knowledge bases etc. The fast-growing tremendous data by these advanced systems stored in large and in more data
repositories. But it was difficult to comprehend by human ability. So, it required to develop powerful tools. As a result, data collected in large data repositories required important decision making systems. The efforts had been made to develop expert systems, knowledge-based technologies and data mining tools that turns data archives into knowledge.

1.2 KNOWLEDGE DISCOVERY PROCESS

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. Data Mining (DM) is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The Knowledge discovery process is starting with large volume of data and passes through selection of data, preprocessing the data, transformation, mine the data and producing the knowledge

![Fig. 1.1 Knowledge Discovery Process](image)

**Selecting and creating a data set:** Developing an understanding of the application domain is the initial step of KDD process. It prepares the scene for understanding what should be done
with the many decisions (about transformation, algorithms, representation, etc.). Having defined the goals, the data that will be used for the knowledge discovery should be determined. This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail. From this aspect, the more attributes are considered. On the other hand, to collect, organize and operate complex data repositories is expensive and there is a tradeoff with the opportunity for best understanding the phenomena. This tradeoff represents an aspect where the interactive and iterative aspect of the KDD is taking place. This starts with the best available data set and later expands and observes the effect in terms of knowledge discovery and modeling.

**Preprocessing and cleansing:** In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers. There are many methods explained in the handbook, from doing nothing to becoming the major part (in terms of time consumed) of a KDD project in certain projects. It may involve complex statistical methods or using a Data Mining algorithm in this context. For example, if one suspects that a certain attribute is of insufficient reliability or has many missing data, then this attribute could become the goal of a data mining supervised algorithm. A prediction model for this attribute will be developed, and then missing data can be predicted. The extension to which one pays attention to this level depends on many factors. In any case, studying the aspects is important and often revealing by itself, regarding enterprise information systems.
Data transformation: In this stage, the generation of better data for the data mining is prepared and developed. Methods here include dimension reduction (such as feature selection and extraction and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). This step can be crucial for the success of the entire KDD project, and it is usually very project-specific. For example, in medical examinations, the quotient of attributes may often be the most important factor, and not each one by itself. In marketing, may need to consider effects beyond our control as well as efforts and temporal issues (such as studying the effect of advertising accumulation). However, even if do not use the right transformation at the beginning, may obtain a surprising effect that hints to us about the transformation needed (in the next iteration). Thus the KDD process reflects upon itself and leads to an understanding of the transformation needed.

Choosing appropriate Data Mining task and algorithm: This step decides on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps. There are two major goals in Data Mining: prediction and description. Prediction is often referred to as supervised Data Mining, while descriptive Data Mining includes the unsupervised and visualization aspects of Data Mining. Most data mining techniques are based on inductive learning, where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future cases. The strategy also takes into account the level of meta-learning for the particular set of available data.

This stage includes selecting the specific method to be used for searching patterns (including multiple inducers). For example, in considering precision versus understandability, the former is better with neural networks, while the latter is better with decision trees. For
each strategy of meta-learning there are several possibilities of how it can be accomplished. Meta-learning focuses on explaining what causes a Data Mining algorithm to be successful or not in a particular problem. Thus, this approach attempts to understand the conditions under which a Data Mining algorithm is most appropriate. Each algorithm has parameters and tactics of learning (such as ten-fold cross-validation or another division for training and testing). Here the algorithm is employed several times until a satisfied result is obtained, for instance by tuning the algorithm’s control parameters, such as the minimum number of instances in a single leaf of a decision tree.

**Evaluation:** This step evaluates and interprets the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step. Consider the preprocessing steps with respect to their effect on the Data Mining algorithm results. This step focuses on the comprehensibility and usefulness of the induced model. In this step the discovered knowledge is also documented for further usage. The last step is the usage and overall feedback on the patterns and discovery results obtained by the Data Mining.

**Using the discovered knowledge:** This step incorporates the knowledge into another system for further action. The knowledge becomes active in the sense that may make changes to the system and measure the effects. Actually the success of this step determines the effectiveness of the entire KDD process. There are many challenges in this step, such as loosing the “laboratory conditions” under which have operated. For instance, the knowledge was discovered from a certain static snapshot (usually sample) of the data, but now the data becomes dynamic. Data structures may change (certain attributes become unavailable), and the data domain may be modified (such as, an attribute may have a value that was not assumed before).
1.3 DATA MINING TASKS

There are number of data mining functionalities. They specify patterns to be found in data mining activities. It includes summarization or characterization, association, classification and cluster analysis.

Summarization

Summarization is the generalization of target class of data. A set of task-relevant data is summarized and abstracted, resulting a smaller set which gives a general overview of the data and usually with aggregation information. For example, the long distance calls of customer can be summarized into total minutes, total spending, total calls, etc. Such high-level summary information instead of detailed calls is presented to the sales managers for customer analysis.

The summarization can go up to different abstraction levels and can be viewed from different angles. For example, the calling in minutes and spending can be totaled along the calling period in weeks, months, quarters, or years. Similarly, the calls can be summarized into in-state calls, state to state calls, Asia calls, Europe calls, etc., which can be further summarized into domestic calls and international calls. Different combinations of abstraction levels and dimensions reveal various kinds of patterns and regularities.

Classification

Classification is the derivation of a function or model which determines the class of data objects based on its attributes. A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is derived or constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. Such a classification function or model can be used
to classify future objects and develop a better understanding of the classes of the objects in the database.

For example, from a set of diagnosed patients, who serve as the training set, a classification model can be built, which concludes a patients disease from his/her diagnostic data. The classification model can be used to diagnose a new patients disease based on the patients diagnostic data such as age, sex, weight, temperature, blood pressure etc.

**Association**

Association is the discovery of togetherness or connection of objects. Such kind of togetherness or connection is termed as association rule. An association rule reveals the associative relationships among objects, i.e., the appearance of a set of objects in a database are strongly related to appearance of another set of objects. For example, in a telecommunication database, an association rule that “call waiting” is associated with “call display”, denoted as “call waiting \(\rightarrow\) call display” says if a customer subscribes to the “call waiting” service, he or she very likely also has “call display”.

The association rules can be useful for marketing, commodity management, advertising, etc. For example a retail store may discover that people chips on sale to promote the sale of soft drinks.

**Clustering**

The data objects are clustered or grouped on the basis of maximizing intra class similarity and minimizing the interclass similarity. The maximizing interclass similarity states that the object with in a class (or) cluster should have high similarity. The minimizing interclass similarity states that object similarity between two clusters should have minimum similar or dissimilar.
1.4 DATA MINING TECHNIQUES

As a multi-disciplinary field, data mining adopted its techniques from many research areas, including statistics, machine learning, pattern recognition, database systems, neural networks, information retrieval, Data warehouse and visualization, high performance computing and algorithms. Some of the principles are briefly discussed in the ensuring part of this section.

Statistical Approach

A Statistical Model describes the characteristics of objects in a class. It is used to model data and its class for mining table, say statistical model can be used in data mining to avoid noise in the data and can prepare missing value in the datasets. Research on statistics has developed tools for prediction and forecasting information using data models. Statistical methods are useful to find patterns and to verify mining results also[81].

Fig. 1.2 Domain of Data Mining
Bayesian network is a directed graph which represents the causal relationships among the variables computed using the Bayesian probability theorem. Regression is the derivation function which maps a set of attributes of objects to an output variable. Correlation analysis studies the correspondence of variables to each other, such as $X^2$. Cluster analysis finds groups from a set of objects based on distance measure.

A simple Bayesian Network (BN) for a medical problem is given below. Nodes in a Bayesian network represent variables or states, while edges represent the dependencies between nodes, directed from the cause to the effect. The diagram below shows that a patient’s age, occupation, and diet affect the disease which in turn causes symptoms.

![Bayesian Network Diagram](image)

**Fig. 1.3 Bayesian Network**

**Machine Learning Approach**

Machine learning determines how computer can learn based on data. That is computer trained through program to find complex patterns and make intelligent decisions based on data for example computer can be programmed or trained to find hand written characters using machine learning approach.

Computer is trained by supervised learning, unsupervised learning, semi supervised learning, active learning and decision tree induction. The most common machine learning
methods used for data mining include decision tree induction. The most common machine learning methods used for datamining include decision tree induction, inductive concept learning and conceptual clustering. A Decision tree is a classification tree which determines an object class by following a path from root to a leaf node, choosing the branches according to the attribute values of the object. Decision trees are induced from the training set and classification rules can be extracted from the decision trees. Inductive concept learning derives a concise, logical description of a concept from a set of examples. Conceptual clustering finds groups or clusters in a set of objects, based on conceptual closeness among objects.

A simple decision tree is shown below. It determines a car’s mileage from its size, transmission type, and weight. The leaf nodes are in boxes which represent the three classes of mileages. From the decision tree, a medium size, automatic car will have medium mileage.

Fig. 1.4 Model of Decision Tree
Database Approach

Database Systems approach focuses on maintenance and use of databases for organization and end-users. Database systems have data models, query mechanism, data storage and accessing methods. Noting that data mining handles large data sets which uses database technologies to achieve good efficiency [76]. Database-oriented methods do not search for a best model compared to the previous two methods. Instead, data model or database specific heuristics are used to exploit the characteristics of the data in hand. The attribute oriented induction, the iterative database scanning for frequent item sets, and the attribute focusing are representatives of the database-oriented methods.

In attribute-oriented induction, primitive, low level data are generalized into high-level concepts using hierarchies. The iterative database scanning method is employed to search for frequent item sets in a transactional database. The association rules are then derived from these frequent item sets. The attribute focusing method looks for patterns with unusual probabilities by adding attributes selectively into patterns.

In the diagram below, the left side shows a simple conceptual hierarchy for students and the right side shows an attributes-oriented induction.

![Conceptual Hierarchy for Students](image)

<table>
<thead>
<tr>
<th>Name</th>
<th>Major</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raja</td>
<td>Elec.Eng</td>
<td>Junior</td>
</tr>
<tr>
<td>Kumar</td>
<td>Comp.Sci</td>
<td>M.Sc.</td>
</tr>
<tr>
<td>Priya</td>
<td>Mech.Eng</td>
<td>M.Tech</td>
</tr>
<tr>
<td>Ravi</td>
<td>Civil.Eng</td>
<td>Senior</td>
</tr>
</tbody>
</table>
Other Approaches

Many other techniques have been adopted for data mining, including neural networks, rough sets and visualization.

A neural network is a set of interlinked nodes, called neurons. A neuron is a simple computing device that computes a function of its inputs which can be outputs of other neurons or attribute values of an object. By adjusting the connection and the functional parameters of the neurons, a neural network can be trained to model the relationship between a set of input attributes and an output attribute. A neural network can be used, for example, in classification when the output attribute is the object class.

A rough set is a set whose membership is fuzzy. A set of objects can be arranged to form a group of rough sets which may be used, in classification and clustering. Visual exploration is another interesting data mining technique. Data are transformed into visual objects such as dots, lines, areas etc, and displayed in a two or three dimensional space. Users can interactively explore the interesting spots by visual examination.

The above method can be integrated or combined to deal with complicated problems or provide alternative solutions. For example, summarization of data is visually presented as charts, graphs, etc., to help understanding of the results and further examination. Indeed, most data mining systems employ multiple methods to deal with different kinds of data, different data mining tasks, and different application areas.

1.5 DATA MINING APPLICATIONS

Data mining techniques have been applied successfully in many areas, from traditional areas such as business and science.
Business Applications

A lot of organizations now employ data mining as a secret weapon to keep or gain competitive edge. Data mining has been used in marketing, retail data analysis, stock selection, credit approval, etc. Database marketing is one of the most successful and popular business applications of data mining. By mining historical customer databases, patterns and trends are extracted and customer profiles are built which can be used for more effective marketing.

Retail databases contain customer shopping transactions. Data mining can find customer shopping patterns which can be used for sales campaign.

Using data mining techniques, investors can build models which can be used to predict the performance of stocks. By searching trends and patterns in stocks data, data mining can also help investors to find stocks with good performance.

Applications for credit or loan are decided based on the applicant’s information. A decision support model for credit or load approval may be constructed from historical data using data mining tools.

Science Applications

Data mining techniques have been used in astronomy, molecular biology, medicine, geology and many more. For example, Jet Propulsion Lab at the California Institute of Technology has developed data mining systems which classify the sky objects such as stars, in the satellite images.
Other Applications

Data mining techniques have also been used in other areas, such as health care management, tax fraud detection, money laundering monitoring and even in sports. For example, the advanced scout system developed by IBM has been used by coaches of more than a dozen teams in the National Basketball Association (NBA) to improve their games.

Thus, data mining is the process of extracting interesting patterns from large databases. Data mining can give solution to data analysis problems faced by Organizations. As a young and emerging field, a lot more work is needed although a great deal of progress has been made in data mining research and development.

1.6 DATA MINING ISSUES

The data mining issues are the outcomes or problems in data mining research. Many issues are addressed in recent data mining research that is considered as new mining requirements. These results of requirement can further improve data mining research. The main issues are related to mining methodologies, user interaction, efficiency, diversity of datatypes and impact to society. The mining methodology is difficult to user to access the discovered patterns and difficulty of applying user background knowledge in mining process. Sometimes the users do not know to visualize and comprehend or understand data mining results. When diversified data types or heterogeneous data are in datasets, one mining principle may not be sufficient. The impact of mining that is reaching user without knowing background processes, the results are obtained by mouse click.
1.7 MOTIVATION

Zhao et al. [28] has proposed combined association rules mining for the extraction of learned rules. In other ways, a novel idea of combined patterns is proposed to discover actionable knowledge from association rules.

Many business problems involve multiple types of information catering user, the demographics, preferences, behavior, business outcomes and impacts. These multiple and heterogeneous datasets require a strong mining principle to create patterns composed of various source of information. Thus various sources of datasets are need to be combined to generate patterns which is termed as combined mining [28]. Hence a traditional single method may not be useful to mine information in complex data / heterogeneous data. The Combined Mining has some advantages

- Flexible framework for complex data, specific algorithms can be taken to handle particular tasks.
- Effective in discovering patterns from multiple heterogeneous sources
- Application of Fuzzy is useful to obtain combined patterns.
- Combination of multiple sources to a single dataset creates direct relationship to objects.

These advantages have motivated to consider combined pattern mining with fuzzy association rules. Consequently the research work is related to Fuzzy Combined Pattern Mining (FCPM). Further FCPM is used with D3M principle.
1.8 CONTRIBUTIONS OF THESIS

There are problems in KDD method of mining patterns in massive database. To simplify the task of mining functional patterns in data sources, a novelty in combined association rule is done with fuzzy principle. It is then used in D3M to analyze the actionable patterns.

As the first step in the proposed research work, existing algorithms related to D3M with decision making patterns are discussed. Two actionable pattern mining algorithms namely mutually dependent pattern mining algorithm and actionable hierarchy based pattern mining algorithm are compared to study the features of actionable patterns.

It is noted that combined mining is a method of analyzing complex data / various data sources to produce combined patterns. An attempt is made in the next step to investigate the patterns formed by combined association rules and fuzzy combined association rules. It is seen that the supported rules are more and driving time for rules is less.

In third step of research contribution, combined association rules are merged with fuzzy rules to see the effect of mining useful patterns. Then the combined association rules using fuzzy is applied to D3M which is the prime contribution of research work. The effect of post analysis in mining is tested with a different dataset.

The research work proposes a model to generate more rules in D3M for better decision making. It uses ubiquitous data such as resources, domain intelligence to deliver functional outcomes that satisfies needs and support decision making. The objective of D3M is involving ubiquitous data in mining.
1.9 ORGANIZATION OF THE THESIS

The first chapter gives overviews of the principle of data mining technologies and its applications. It highlights the issues of Data Mining.

The second chapter introduces the literature survey of Rule mining, combined pattern mining. It also discusses various methods for motivating the research work.

The chapter three reviews in detail the related works of association rule mining and combined association mining which is the back bone of mining principle.

The fourth chapter presents the investigation of combined association rules with fuzzy combined rules. It helps to analyze complex data for identifying complex knowledge.

The key concept of Domain Driven Data Mining is highlighted in fifth chapter. It has synthesized all intelligences namely domain intelligence, human intelligence, data source etc in effective actionable knowledge discovery.

Fuzzy based combined pattern mining is proposed in the sixth chapter. It is a new technique that mines actionable frequent pattern from large datasets invoking fuzzy principle and D3M concepts.

The chapter seven presents the performance of proposed work with existing algorithms taking data set, support count, and rule generation factors.

The final chapter briefs findings and suggestions for future enhancement.