Volume 4, Issue 8, August 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review Paper on Knowledge Discovery and Data Mining Techniques

Rajwant Kaur* CSE/IT, PTU Punjab, India
Kiran Jyoti CSE/IT, PTU Punjab, India
Rohit Kumar CSE/IT, PTU Punjab, India

Abstract— In this paper, a review of knowledge discovery in data mining (KDD) or database, KDD process and data mining techniques. Data Mining is a new methodology for improving the quality and effectiveness of business and scientific decision-making process. The term Knowledge Discovery in database or KDD, refers to the broad process of finding knowledge in data and emphasis the "high level" application of particular data mining methods.

Keywords— Knowledge Discovery Process, Data Mining Techniques,

I. INTRODUCTION

Data Mining is a process of discovering hidden and unknown patterns. Data mining finds these patterns and relationships using data analysis tool and techniques to build models. It is a new kind of business information analysis technique. Its aim is to find hidden information by extracting , transforming, analyzing and modeling from large amount of data in business database. Data mining can be defined as "a decision support process in which we search for patterns of information in data". The goal of data mining is to create models for decision-making that predict future behavior based on analyses of past activity. Data Mining tools can analyze massive databases to deliver answers to questions, when implemented on high performance client/server or parallel processing computers.

Data Mining is used in business environment as well as other fields such as weather forecast, medicine, transportation, healthcare, insurance, government and etc. Data Mining produce lots of advantages, when it is used in a specific industry. Data Mining involves multiple steps as shown in figure 1. The process starts with selection of dataor understanding of data that consists of historical data. After that in data preparation data is then cleaned and preprocessed. Cleaning process removes the discrepancies and preprocessing is responsible for relevant information. The next steps is data modelling, in this step the data is modelled to identify the patterns and after that the data is evaluated as per the requirement. In the last step data is finally deployed with new data sets. The process continues until meaningful knowledge is extracted.

The distinction between data mining and the traditional methods with database is that, in traditional methods the database becomes passive and it is used only for storing large amounts of data. In data mining, the database is no longer passive. It offers useful information for business plans by using data analysis process.

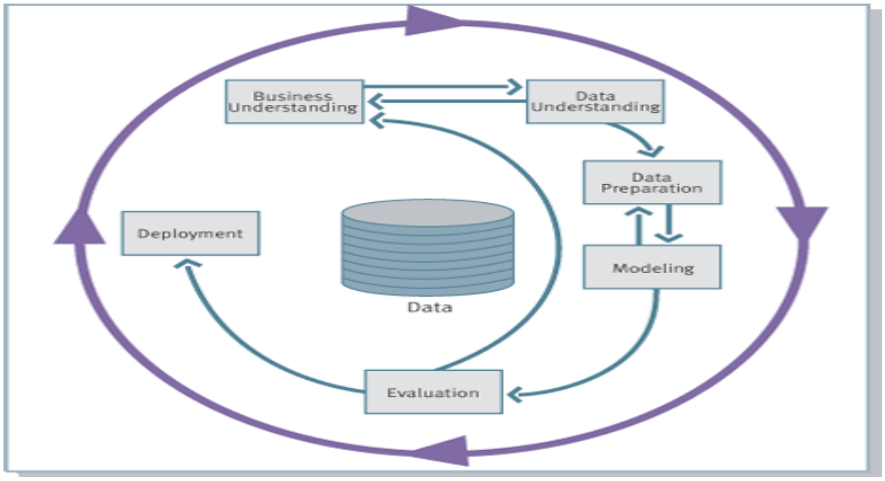

Fig. 1 Steps of Data Mining

Data Mining has six types of models, which are used to solve business problems: classification, regression, time series, clustering, association and sequence discovery. The first two classifications and regression are used to make predictions, while association and sequence discovery are used to describe behavior. Clustering can be used for either forecasting or description. Companies in various industries can gain a competitive edge by mining their expanding databases for valuable, detailed transaction information.

II. KNOWLEDGE DISCOVERY IN DATABASE

The term Knowledge Discovery in database or KDD, refers to the broad process of finding knowledge in data and emphasis the "high level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, sub sampling , and transformations of that database.

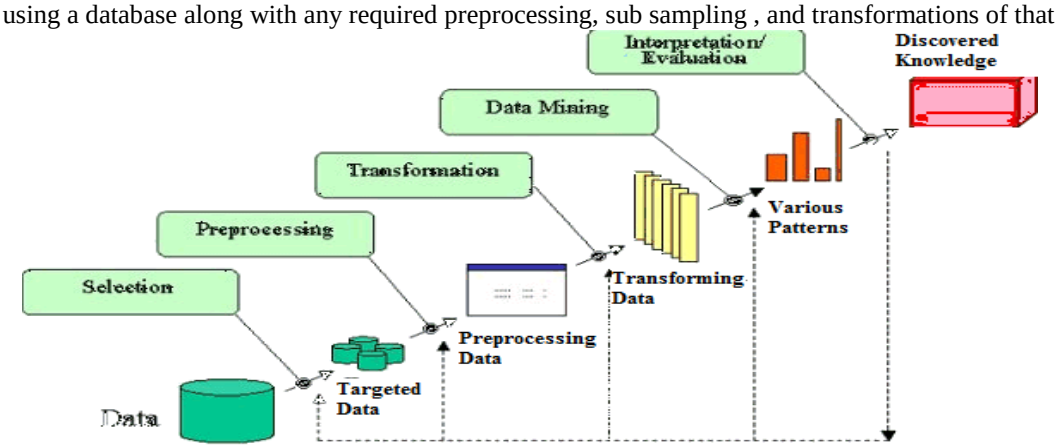

Fig. 2 Steps of KDD Process

- **Developing an understanding of the application domain:** This is the initial preparatory step. It prepares the scene for understanding what should be done with many decisions(about transformation, algorithms, representation, etc.).
- **Selecting and creating a target data set:** Having defined the goals, the data that will be used for the knowledge discovery should be determined. This includes selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed.
- **Data cleaning and preprocessing:** It includes data clearing , such as handling missing values and removal of noise and outliers. If appropriate, collecting the necessary information to model, deciding on strategies for handling missing data fields.
- **Data transformation:** In this stage, the generation of better data for the data mining is prepared and developed. It includes finding useful features to represent the data, depending on the goal of the task.
- **Choosing the appropriate data mining task:** This includes deciding the purpose of the model derived by the data mining algorithm(e.g., summarization, classification, regression, clustering). This mostly depends on the KDD goals, and also on the previous steps.
- **Choosing the data mining algorithm:** This stage includes selecting the specific method to be used for searching patterns, such as deciding which models and parameters may be appropriate, and matching a particular data mining method with the overall criteria of the KDD process.
- **Data Mining:** In this step the implementation of the Data Mining algorithm is done.
- **Interpretation or evaluation:** In this stage, the mined patterns are evaluate and interpret with respect to the goals defined in the first step. Here the pre-processing steps with respect to their effect on the Data Mining algorithm results are considered.
- **Using the discovered knowledge:** It includes incorporating the knowledge into another system for future action. The knowledge becomes active in the sense that we make changes to the system and measure the effects.

III. TECHNIQUES IN DATA MINING

There are several major data mining techniques have been developing and using in data mining projects including association, classification, clustering, prediction, regression. These are briefly examine in the following sections.

1. Association: Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's why association technique is also known as *relation technique*. The association technique is used in *market basket analysis* to identify a set of products that customers frequently purchase together. Association rules mining has many applications other than market basket analysis, including applications in marketing, customer segmentation, medicine, electronic commerce, bioinformatics and finance. The patterns discovered with this data mining technique can be represented in the form of association rules.

- **Apriori algorithm:** Apriori is a classic algorithm used in data mining for learning association rules. It is a algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine

association rules which highlight general trends in the database. Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to count the support of item sets and uses a candidate generation function which exploits the downward closure property of support.

- **Eclat algorithm:** Eclat is a depth-first search algorithm using set intersection. The Eclat algorithm is used to perform item set mining. Item set mining let us find frequent patterns in data like if a consumer buys milk, he also buys bread. This type of pattern is called association rules and is used in many application domains. The basic idea for the eclat algorithm is use tidset intersections to compute the support of a candidate item set avoiding the generation of subsets that does not exist in the prefix tree.
- **FP- growth algorithm:** The FP- growth algorithm allows frequent itemset discovery without candidate itemset generation. The Algorithm, proposed by Hanin, is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree). In his study, Han proved that his method outperforms other popular methods for mining frequent patterns, e.g. the Apriori Algorithm.

2. Classification: Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

- **Decision Trees:** Decision trees are trees that classify instances by sorting them based on feature values. The Microsoft Decision Trees algorithm is a classification and regression algorithm provided by Microsoft SQL Server Analysis Services for use in predictive modeling of both discrete and continuous attributes. Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. Decision trees play well with other modeling approaches, such as regression, and can be used to select inputs or to create dummy variables representing interaction effects for regression equations. For example, Neville (1998) explains how to use decision trees to create stratified regression models by selecting different slices of the data population for in-depth regression modeling.
- **Naïve Bayesian Classification:** It is based on the Bayesian theorem It is particularly suited when the dimensionality of the inputs is high. The Microsoft Naive Bayes algorithm is a classification algorithm provided by Microsoft SQL Server Analysis Services for use in predictive modeling. The algorithm calculates the conditional probability between input and predictable columns, and assumes that the columns are independent. This algorithm is less computationally intense than other Microsoft algorithms, and therefore is useful for quickly generating mining models to discover relationships between input columns and predictable columns. The algorithm considers each pair of input attribute values and output attribute values.
- **Support vector machine:** Support vector machines(SVM) have been promising methods for data classification and regression. SVM performs well on data sets that have many attributes, even if there are very few cases on which to train the model. There is no upper limit on the number of attributes; the only constraints are those imposed by hardware. Traditional neural nets do not perform well under these circumstances. In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

3. Clustering: Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. The algorithm uses iterative techniques to group cases in a dataset into clusters that contain similar characteristics. These groupings are useful for exploring data, identifying anomalies in the data, and creating predictions. For example, you can logically discern that people who commute to their jobs by bicycle do not typically live a long distance from where they work. The algorithm, however, can find other characteristics about bicycle commuters that are not as obvious. In the following diagram, cluster A represents data about people who tend to drive to work, while cluster B represents data about people who tend to ride bicycles to work.

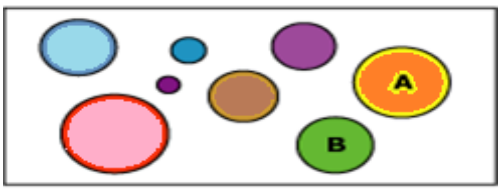

Fig. 3 Clusters

- **Hierarchical —** Groups data objects into a hierarchy of clusters. The hierarchy can be formed top-down or bottom-up. Hierarchical methods rely on a distance function to measure the similarity between clusters.
- **Partitioning** — Partitions data objects into a given number of clusters. The clusters are formed in order to optimize an objective criterion such as distance
- **Locality-based —** Groups neighboring data objects into clusters based on local conditions.
- **Grid-based —** Divides the input space into hyper-rectangular cells, discards the low density cells, and then combines adjacent high-density cells to form clusters.

4. Prediction: The prediction, as it name implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

IV. ARCHITECTURE OF DATA MINING

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses etc. The aim of this technology is usually to find hidden but significant relationships that can lead to a bigger profit. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed.

Data is collected explosively every minute through business transactions and stored in relational database systems. In order to provide insight about the business processes, data warehouse systems have been built to provide analytical reports that help business users to make decisions. The four possible architectures of a data mining system as follows:

- No-coupling: In this architecture, data mining system does not utilize any functionality of a database or data warehouse system. A no-coupling data mining system retrieves data from a particular data sources such as file system, processes data using major data mining algorithms and stores results into file system. The no-coupling data mining architecture does not take any advantages of database or data warehouse that is already very efficient in organizing, storing, accessing and retrieving data. The no-coupling architecture is considered a poor architecture for data mining system however it is used for simple data mining processes.
- Loose Coupling: In this architecture, data mining system uses database or data warehouse for data retrieval. In loose coupling data mining architecture, data mining system retrieves data from database or data warehouse, processes data using data mining algorithms and stores the result in those systems. This architecture is mainly for memory-based data mining system that does not require high scalability and high performance.
- Semi-tight Coupling: In semi-tight coupling data mining architecture, beside linking to database or data warehouse system, data mining system uses several features of database or data warehouse systems to perform some data mining tasks including sorting, indexing, aggregation…etc. In this architecture, some intermediate result can be stored in database or data warehouse system for better performance.
- Tight Coupling: In tight coupling data mining architecture, database or data warehouse is treated as an information retrieval component of data mining system using integration. All the features of database or data warehouse are used to perform data mining tasks. This architecture provides system scalability, high performance and integrated information.

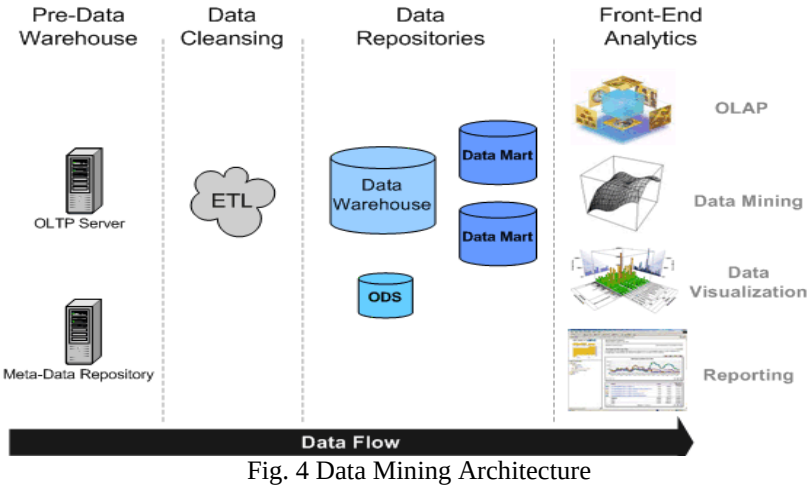

Fig. 4 Data Mining Architecture

Tight Coupling data mining architecture have three tiers, which are as follows:

i. Data layer: as mentioned above, data layer can be database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end-user in form of reports or other kind of visualization.

ii. Data mining application layer is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms.

iii. Front-end layer provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer.

V. BENEFITS OF DATA MINING

Data mining is applied effectively not only in business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government…etc. Data mining has a lot of advantages when using in a specific industry. Besides those advantages, data mining also has its own disadvantages e.g., privacy, security and misuse of information. Let us examine the benefits of data mining in different industries in greater detail.

i Marketing / Retail: Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers. Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

ii Finance / Banking: Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

iii Manufacturing: By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

iv Governments: Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

v Science: Data mining has helped in the various researches in the field of science. Other than researches, data mining has also contributed to help the medical field and the engineering field in improving their processes.

VI. CONCLUSIONS

In this article provided an overview of knowledge discovery in data mining (KDD) or database, KDD process, data mining techniques. Also describes its benefits and architecture of data mining. **Data base mining** or Data mining (DM) (formally termed Knowledge Discovery) is a process that aims to use existing data to invent new facts and to uncover new relationships previously unknown even to experts thoroughly familiar with the data. In this paper we have outlined a review of data mining with its techniques and benefits. This framework will be beneficial for a brief knowledge.

REFERENCES

[1] R.Agrawal, T.Imielinski, and A. Swami Database mining: A performanceperspective IEEE Transactions onKnowledge and Data Engineering, 5(6):914{925, December 1993. Special Issue on Learning and Discovery in Knowledge Based Databases.

[2] ÓscarMarbán, Gonzalo Mariscal and Javier Segovia (2009); *A Data Mining & Knowledge Discovery Process Model*. In Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and AdemKarahoca, ISBN 978-3-902613-53-0, pp. 438-453, February 2009, I-Tech, Vienna, Austria.

[3] ADELA TUDOR, ADELA BARA, IULIANA BOTHA The Bucharest Academy of Economic Studies Bucharest ROMANIA (2011) : Solutions for analyzing CRM systems - data mining algorithms. INTERNATIONAL JOURNAL OF COMPUTERS Issue 4, Volume 5, 2011.

[4] Chien-Hua Wang and Chin-Tzong Pang : Applying Fuzzy Data Mining for an Application CRM H. -J. Zimmermann, *Fuzzy sets, Decision Making, and Expert Systems*, Kluwer, Boston, 1991.

[5] Chris Rygielski , Jyun-Cheng Wang , David C. Yen : Data mining techniques for customer relationship management, Technology in Society 24 (2002) 483–502.

[6] Usama Fayyad , Gregory Piatetsky - Shapiro , and Padhraic Smyth : The KDD Process for Extracting Useful Knowledge from Volumes of Data, COMMUNICATIONS OF THE ACM November 1996/Vol. 39, No. 11.

