International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 08 | August -2020 ISSN: 2582-3930

A Brief Overview on Data Mining & its Techniques

Divya Y A[1], Chaya P[2], Soumya Dath G[3]
[1]Department of ISE, GSSS Institute of Engineering and Technology for Women, Mysuru,
[2]Department of ISE, GSSS Institute of Engineering and Technology for Women, Mysuru,
[3]Department of ISE, GSSS Institute of Engineering and Technology for Women, Mysuru,

Abstract: Data mining is used to extract meaningful information and to develop significant relationships among variables stored in large data set/data warehouse. Knowledge /information are conveying the message through direct or indirect. This paper provides a survey of various data mining techniques. Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization.

Keywords: *Association, Clustering,* Data mining, data mining application, knowledge discovery database.

I. INTRODUCTION

Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data [2].

The essential difference between the data mining and the traditional data analysis (such as query, reporting and on-line application of analysis) is that the data mining is to mine information and discover knowledge on the premise of no clear assumption [1]. Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1.1) shows data mining as a step in an iterative knowledge discovery process.

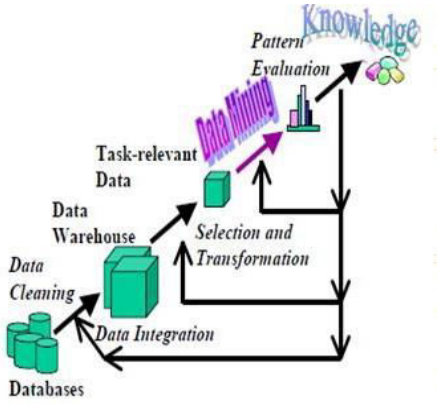

Fig.1: Data mining is the core of Knowledge Discovery Process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

Data mining commonly involves four classes of tasks:

Clustering - is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering.

Classification - is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.

Regression - Attempts to find a function which models the data with the least error.

Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data.

Association rule learning - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

II. The Data Mining Process

Generally, data mining process is composed by data preparation, data mining, and information expression and analysis decision-making phases, the specific process as shown in fig.1[5] .

1) Data preparation

Data preparation generally consists of two processes: data collection and data collation. Data collection is the first step of data mining, and the data can come from the existing transaction processing systems, also can be obtained from the data warehouse; data collation is to eliminate noise or inconsistent data, it is the necessary link of data mining. The data obtained from the phase of the data collection may have a certain degree of "pollution", which refers to that in the data may be its own inconsistency, or some missing data, so the collation of the data is essential.

2) Data mining

Data mining is the core stage of the entire process, it mainly uses the collected mining tools and techniques to deal with the data, thus the rules, patterns and trends will be found.

3) Information expression

Information expression is to use visualization and knowledge information expression technology to provide the mined knowledge information for users, is an important means to show the data mining results. Clear and effective mining result information expression will greatly facilitate the accuracy and efficiency of the decision-making.

4) Analysis and decision-making

The ultimate goal of data mining is to assist the decision making. Decision-makers can analyze the results of data mining and adjust the decision-making strategies combining with the actual situation.

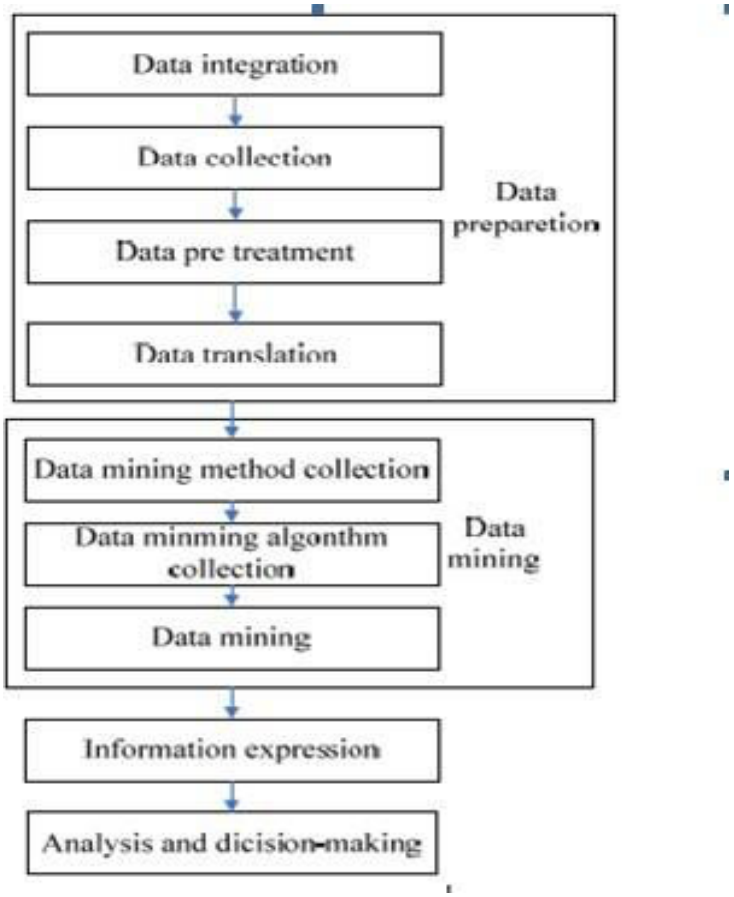

Fig.3: General process of Data Mining

III. DATA MINING ARCHITECTURE:

There are three tiers in the tight-coupling data mining architecture:

1. 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.
2. 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.
3. 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.

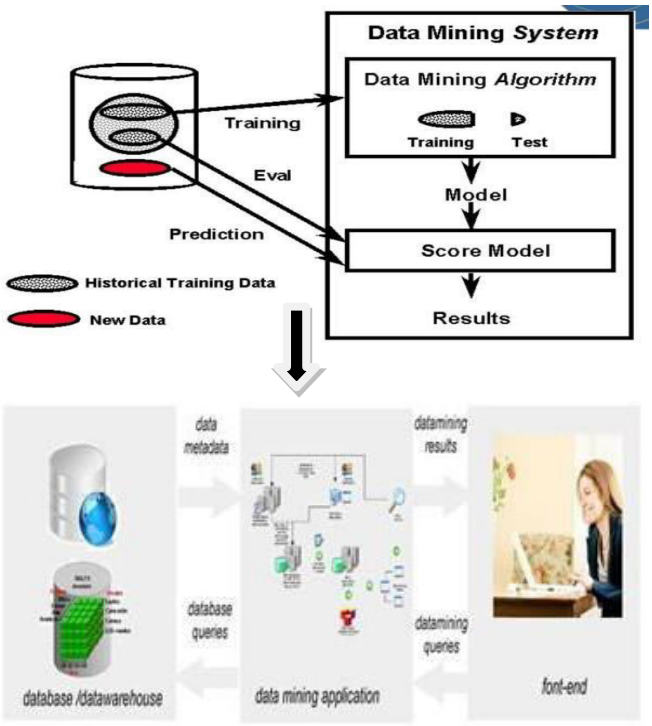

Fig.4: Architecture of Data mining

IV. DATA MINING TECHNIQUES

Data mining based on decision tree

Decision tree learning, used in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are **classification trees** or **regression trees**. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making

A **decision support system** (**DSS**) is a computer-based information system that supports business or organizational decision-making activities. DSSs serve the management, operations, and planning levels of an organization and help to make decisions, which may be rapidly changing and not easily specified in advance.

DSSs include knowledge-based systems. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from a combination of raw data, documents, personal knowledge, or business models to identify and solve problems and make decisions.

Data mining based on neural network:

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases.

Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition. It

imitates the neurons structure of animals, bases on the M-P model and Hebbien learning rule, so in essence it is a distributed matrix structure. Through training data mining, the neural network method gradually calculates (including repeated iteration or cumulative calculation) the weights the neural network connected.

Data mining: K means clustering

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

The k-means Algorithm:

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation‟s proximity to the mean of the cluster. The cluster‟s mean is then recomputed and the process begins again. Here‟s how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers (“means”).
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

V. FUTURE ENHANCEMENT

Over recent years data mining has been establishing itself as one of the major disciplines in computer science with growing industrial impact. Undoubtedly, research in data mining will continue and even increase over coming decades involve Mining complex objects of arbitrary type, fast, transparent and structured data preprocessing, Increasing usability. All aim at understanding consumer behavior, forecasting product demand, managing and building the brand, tracking performance of customers or products in the market and driving incremental revenue from transforming data into information and information into knowledge.

Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

VI. CONCLUSION

Data mining is a hot topic of the computer science research in recent years, and it has a extensive applications in various fields. Data mining technology is an application oriented technology. It not only is a simple search, query and transfer on the particular database, but also analyzes, integrates and reasons these data to guide the solution of practical problems and find the relation between events, and even to predict future activities through using the existing data.

Data mining brings a lot of benefits to businesses, society, governments as well as individual. However privacy, security and misuse of information are the big problem if it is not address correctly.

REFERENCES

[1] Ming-Syan Chen, Jiawei Han, Philip S yu. Data Mining: An Overview from a Database Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, l996, 8(6):866-883.

[2] R Agrawal ,T 1 mielinski, A Swami. Database Mining: A Performance Perspective[J]· IEEE Transactions on Knowledge and Data Engineering, 1993,12:914-925.

[3] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996- Fayyad.pdf Retrieved 2008-12-17..

[4] Y. Peng, G. Kou, Y. Shi, Z. Chen (2008). "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery" *International Journal of Information Technology and Decision Making, Volume 7, Issue 4* 7: 639 – 682. doi:10.1142/S0219622008003204.

[5] Data mining:Ford, C.W.; Chia-Chu Chiang; Hao Wu; Chilka, R.R.; Talburt,J.R.; Information Technology: Coding and Computing, 2005. ITCC 2005 InternationalConference Volume: Digital Object Identifier: 10.1109/ITCC.2005.270 Publication Year: 2005 , Page(s): 122 - 127 Vol. 1

[6] Han, J. & M. Kamber, Data mining: concepts and techniques, San Francisco: Morgan Kaufman (2001).

[7] "Data mining tools", by Ralf Mikut, Markus Reischl, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011

[8] "The applied research on data mining in the financial analysis of university with the analysis of college students „arrears as an example" Chen Hongfei; Wang Xiaoyan; Business Management and Electronic Information (BMEI), 2011 International Conference on
Volume:2 Digital Object Identifier: 10.1109/ICBMEI.2011.5917992 Publication Year: 2011 , Page(s): 633 - 636