INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

# DATA MINING AND ITS TECHNIQUE: AN OVERVIEW

[1]Bilawal Singh, [2]Sarvpreet Singh

**Abstract:** Data mining has assumed a global proportion in every sector such as higher education, science, biodiversity information technology, mathematics geology and many more…Data mining provides a practical means for classification and distinction of tremendous amount of data related to any field of information .data mining is more of exploratory analysis. It basically aims at expanding the learning culture and helps in predicting the future need of that data.

## 1. INTRODUCTION

Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge mining".
Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets 'patterns discovery. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is about solving problems by analyzing data already present in databases [1]
In its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.[2] Data mining monitors each similar pattern from data gathered and helps in making prediction about the performance. Data mining answers "who" first then. "Why"… a matter of deductive Reasoning than inductive reasoning. Data mining results will enhance theoretical based research by giving its real name or numbers.

Data mining consists of five major elements:

· Extract, transform, and load transaction data onto the data warehouse system.
· Store and manage the data in a multidimensional database system.
· Provide data access to business analysts and information technology professionals.
· Analyze the data by application software.

· Present the data in a useful format, such as a graph portable. [3]

## 2. KNOWLEDGE DISCOVERY PROCESS

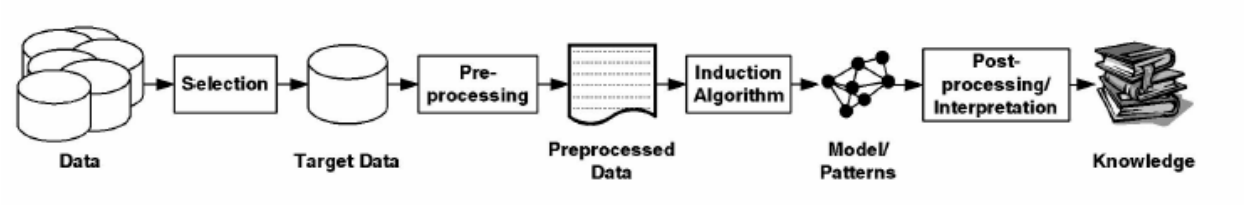



Figure 1. Process of Data Mining [4]

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Three steps involved are

- Exploration
- Pattern identification
- Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

Pattern Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

Deployment: Patterns are deployed for desired outcome.[5]

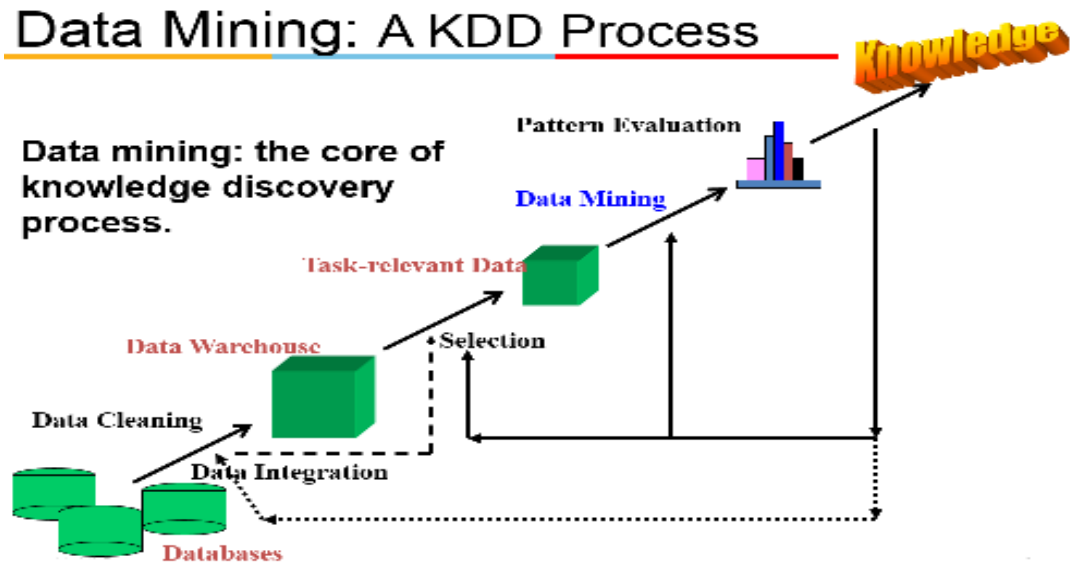



Figure 2.Data mining: KDD process

1. DATA CLEANING: To Remove Noise and Inconsistent Data

2. DATA INTEGRATION: Where Multiple Data Sources May Be Combined

3. DATA SELECTION: Where Relevant To The Analysis Task Are Retrieved From The Database.

4. DATA TRANSFORMATION: Where Data Are Transformed Or Consolidated Into Appropriate For Mining By Performing Summary Or Aggregations Operations

5. DATA MINING: An Essential Process Where Intelligent Methods Are Applied In Order To Extract Data Patterns.

6. PATTERN EVALUATION: To Identify The Truly Interesting Patterns Representing Knowledge Based On Some Interestingness Measure.

7. KNOWLEDGE PRESENTATION: Where Visualization and Knowledge Representation Techniques Are Used To Present the Mined Knowledge To The User

## 3. TECHNIQUES AND ALGORITHMS IN DATA MINING

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

1. MACHINE LEARNING algorithms

Machine learning is a set of tools that, broadly speaking, allow us to "teach" computers how to
Perform tasks by providing examples of how they should be done Machine learning uses statistics (mostly inferential statistics) to develop self-learning algorithms.
Data mining uses statistics (mostly Descriptive statistics) on results obtained from algorithms, it used to solve the problem. Data mining as a field emerged to solve problems in the miscellaneous domain (particularity in business), acquired different techniques and practices that are used in different field of studies.

Machine learning is a diverse and exciting field, and there are multiple ways of defining it:

1. The Artificial Intelligence View. Learning is central to human knowledge and intelligence, and, likewise, it is also essential for building intelligent machines. Years of effort in AI has shown that trying to build intelligent computers by programming all the rules cannot be done; automatic learning is crucial. For example, we humans are not born with the ability to understand language — we learn it — and it makes sense to try to have computers learn language instead of trying to program it all it.

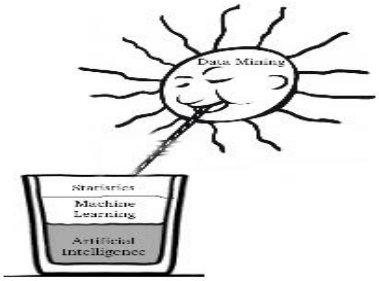
2. The Software Engineering View. Machine learning allows us to program computers by Example, which can be easier than writing code the traditional way.

3. The Stats View. Machine learning is the marriage of computer science and statistics: computational techniques are applied to statistical problems. Machine learning has been applied

to a vast number of problems in many contexts, beyond the typical statistics problems. Machine Learning is often designed with different considerations than statistics (e.g., speed is often more important than accuracy).

Often, machine learning methods are broken into two phases:
1. Training: A model is learned from a collection of training data.
2. Application: The model is used to make decisions about some new test data.

2. ARTIFICIAL INTELLIGENCE Artificial Intelligence is a science to develop a system or software to mimic human to respond and behave in a circumference. As field with extremely broad scope, AI has defined its goal into multiple chunks. Later each chuck has become a separate field of study to solve its problem.
Here is a major list of AI goal (a.k.a. AI problems)
1. Reasoning
2. Knowledge representation
3. Automated planning and scheduling
4. Machine learning
5. Natural language processing
6. Computer vision
7. Robotics
8. General intelligence or strong AI




As mentioned in the list Machine learning is field emerged from one the AI goal to help machine or software to learn on it own to solve problems it's can come across. Natural language processing is another such field emerged from AI goal to help machine to communicate with real human. Computer vision is a field emerged from AI goal to identify and distinguish objects that the machine could see. Robotics is a field emerged from AI goal to give a physical appearance for a machine to do physical actions.

## 3. NEURAL NETWORKS

Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions. The greatest breakthroughs in neural network in recent years are in their application to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications [6]. This powerful predictive modeling technique creates very complex models that are really difficult to understand by even experts. Neural Networks are used in a variety of applications. It is shown in figure. Artificial neural network have become a powerful tool in tasks like pattern recognition, decision problem or predication applications. It is one of the newest signals processing technology. ANN is an adaptive, nonlinear system that learns to perform a function from data and that adaptive phase is normally training phase where system parameter is change during operations [7]

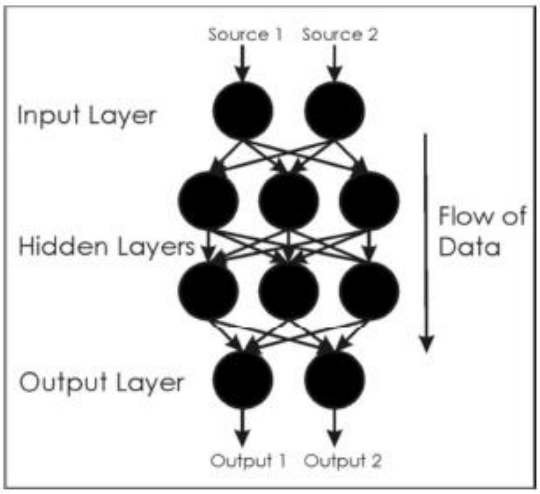



Figure 3. Neural Networks

## 4. TOOLS USED IN DATA MINING FOR EXTRACTION OF USEFUL DATA

Most data mining tools can be classified into one of three categories:
I. Traditional Data Mining Tools
II Dashboards
III Text-Mining Tools

I. Traditional Data Mining Tools:

Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only.

II. Dashboards:

Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

III .Text-Mining Tools:

The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text — from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes. [2]

IV. Weka tool:

The Waikato Environment for Knowledge Analysis (WEKA) came about through the perceived need for a unified workbench that would allow researchers easy access to state-of the-art techniques in machine learning. At the time of the project's inception in 1992, learning algorithms were available in various languages, for use on different platforms, and operated on a variety of data formats. The task of collecting together learning schemes for a comparative study on a collection of data sets was daunting at best. It was envisioned that WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation[8].


Figure 4. Weka GUI Interface

The WEKA project aims to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and practitioners alike. It allows users to quickly try out and compare different machine learning methods on new data sets. Its modular, extensible architecture allows sophisticated data mining processes to be built up from the wide collection of base learning algorithms and tools provided. Extending the toolkit is easy thanks to a simple API, plug-in mechanisms and facilities that automate the integration of new learning algorithms with WEKA's graphical user interfaces. The workbench includes algorithms for regression, classification, clustering, association rule mining and attribute selection. Preliminary exploration of data is well catered for by data visualization facilities and many preprocessing tools. These, when combined with statistical evaluation of learning schemes and visualization of the results of learning, supports process models of data mining such as CRISP-DM [9].

V. RAPIDMINER TOOL:

Rapid Miner [10], formerly YALE (Yet another Learning Environment), is an environment for providing data mining and machine learning procedures including: data loading and transformation (ETL), data preprocessing and visualization, modeling, evaluation, and deployment. The data mining processes can be made up of arbitrarily nest able operators, described in XML files and created in Rapid Miner's graphical user interface (GUI). Rapid Miner is written in the Java programming language. It also integrates learning schemes and attribute evaluators of the Weka machine learning environment and statistical modeling schemes of the R-Project. Rapid Miner can be used for text mining, multimedia mining, feature engineering, data stream mining and tracking drifting concepts, development of ensemble methods, and distributed data mining. Rapid Miner is found in the: electronics industry, energy industry, automobile industry, commerce, aviation, telecommunications, banking and insurance, production, IT industry, market research, pharmaceutical industry and other fields.

VI. DBMINER TOOL:

DBMiner[11], a data mining system for interactive mining of multiple-level knowledge in large relational databases, has been developed based on our years-of-research. The system implements a wide spectrum of data mining functions, including generalization, characterization, discrimination, association, classification, and prediction.

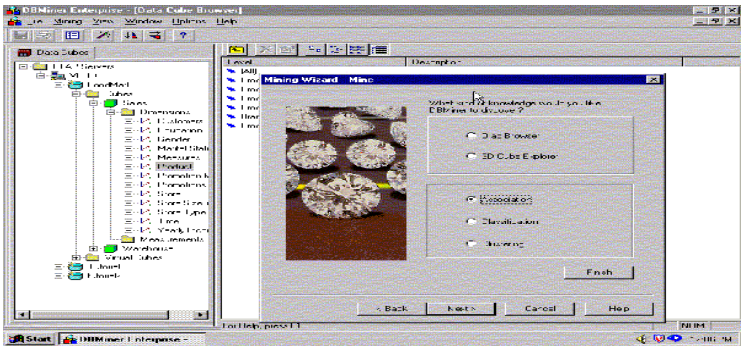


Figure 5. DBMINER GUI Interface

DBMiner performs interactive data mining at multiple concept levels on any user-specified set of data in a database using an SQL-like Data Mining Query Language, DMQL, or a graphical user interface. Users may interactively set and adjust various thresholds, control a data mining process, perform roll-up or drill-down at multiple concept levels, and generate different forms of outputs, including generalized relations, generalized feature tables, multiple forms of generalized rules, visual presentation of rules, charts, curves, etc.

VII. WITNESS MINER TOOL:

WITNESS Miner [12] is a graphical data mining tool comprising a collection of data structures and algorithms written specifically for the tasks required in knowledge discovery. Designed to be easy to use, it provides a visual method of constructing streams, containing data preparation and data mining tasks that form the knowledge discovery process. The key features of this tool are: decision trees, clustering, discretization, rule induction using modern heuristic techniques, the ability to handle missing values, host of standard data processing tools, HTML output and in the case of the decision tree, XML output options, feature subset selection.

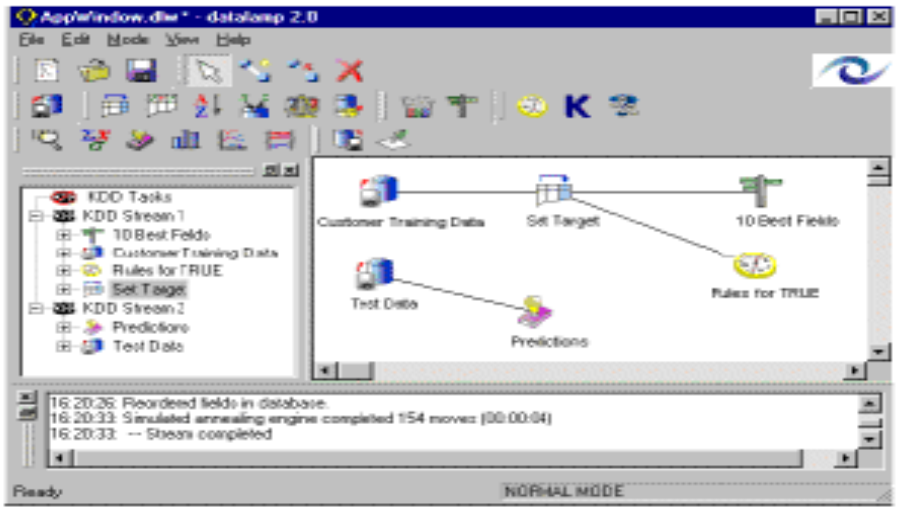



Figure 6 .WITNESS Miner GUI Interface

WITNESS Miner offers a way of making sense of data in Manufacturing, Finance, Health, Retail and Government. It provides knowledge from raw data through, data analysis, easy data modeling, powerful rule evaluation and high quality reporting. Most importantly, it allows an exploration of data to determine fundamental relationships that affect business. The WITNESS Miner module offers easy to understand rules generated directly from the data.

## 5. APPLICATIONS OF DATA MINING

There are approximately 100,000 genes in a human body and each gene is composed of hundreds of individual nucleotides which are arranged in a particular order. Ways of these nucleotides being ordered and sequenced are infinite to form distinct genes. Data mining technology can be used to analyze sequential pattern, to search similarity and to identify particular gene sequences that are related to various diseases. In the future, data mining technology will play a vital role in the development of new pharmaceuticals and advances in cancer therapies.

Financial data collected in the banking and financial industry is often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. Typical cases include classification and clustering of customers for targeted marketing, detection of money laundering and other financial crimes as well as design and construction of data warehouses for multidimensional data analysis.

The retail industry is a major application area for data mining since it collects huge amounts of data on customer shopping history, consumption, and sales and service records. Data mining on retail is able to identify customer buying habits, to discover customer purchasing pattern and to predict customer consuming trends. Data mining technology helps design effective goods transportation, distribution polices and less business cost.
Data mining in telecommunication industry can help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources and improve service quality. Typical cases include multidimensional analysis of telecommunication data, fraudulent pattern analysis and the identification of unusual patterns as well as multidimensional association and sequential pattern analysis.[13]

## 6. REFERENCES

[1] “Data Mining as a Civic Duty – Online Public Prisoners Registration Systems”, by Joseph, Zernik, International Journal On social Media: Monitoring, Measurement, Mining, Vol. - 1, No.-1, Pp. 84-96, September2010.
[2] A Study of Data Mining Tools in Knowledge Discovery Process by Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam International Journal of Soft Computing and Engineering (Ijsce) Issn: 2231-2307, Volume-2, Issue-3, July 2012.
[3] Data Mining Techniques: A Survey Paper by Nikita Jain, Vishal Srivastava, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.
[4] Improving Academic Performance Of Students By Applying Data Mining Technique by N.V.Anand & G.V.Uma Chennai European Journal Of Scientific Research Issn 1450-216x Vol.34 No.4 (2009), Pp.526-534 © Euro journals Publishing, Inc. 2009.
[5] Data Mining Techniques And Applications by Mrs. Bharati M. Ramageri , Indian Journal Of Computer Science And Engineering Vol. 1 No. 4 301-305 ISSN : 0976-5166.
[6] ”A Survey And Critique Of Techniques For Extracting Rules From Trained Artificial Neural Networks”, Knowledge-Based Systems by R. Andrews, J. Diederich, A. B. Tickle,Vol.- 8,No.-6, Pp.-378-389,1995.
[7] A Survey On Data Mining Techniques by M. Suganthi , India International Journal Of Innovation And Scientific Research Issn 2351-8014 Vol. 10 No. 1 Oct. 2014, Pp. 1-5© 2014

[8] The WEKA data mining software: An update, Mark Hall, Eibe Frank, G. Holmes, B. Pfahringer, P. Reutemann, IH Witten, ACM SIGKDD Explorations, Newsletter, Pages 10-18, volume 11 issue 1, June 2009.

[9] C. Shearer. The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4),2000

[10] http://rapid-i.com/

[11] DBMiner: A System for Data Mining in Relational Databases and Data Warehouses, Data Mining Research Group, Intelligent Database Systems Research Laboratory School of Computing Science, Simon Fraser University, British Columbia, Canada, http: db.cs.sfu.ca DBMiner.

[12] www.uea.ac.uk/polopoly_fs/1.3589!introductionkdd.pdf

[13] Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques*, London: Academic Press, 5, 2001.