



## Data Mining And Its Significance In Industrial Applications

Prasanna Kumara\*

Research Scholar, Department of Computer Science,  
Rayalaseema University Kurnool, India  
[prasanna.s\\_kumara@yahoo.com](mailto:prasanna.s_kumara@yahoo.com)

Alok Ranjan

Research Scholar, Department of Computer Science,  
Rayalaseema University Kurnool, India  
[mibalok07@gmail.com](mailto:mibalok07@gmail.com)

**Abstract:** Data Mining is the process of analyzing the data from different perspective and summarizing it into meaningful information. In fact, Data Mining is a technique involving non-trivial extraction of novel, implicit, and actionable knowledge from a huge sets of data, and is an emerging technology in the recent era, with the increased use of databases to store and retrieve information efficiently. This is also termed as 'Knowledge Discovery in Database', which enables the data exploration, data analysis, and data visualization of huge database at a high level of abstraction. These techniques are used to predict future trends and behaviors, thus allowing the business to make proactive and knowledge driven decisions. This paper focus on the Data Mining techniques, its architecture, its process, methodologies employed, kinds of data mined, functionalities, general and technical issues, its significance and application areas in relevance to today's business environment, and the scope for research in the related areas.

Keywords: Database, Data Mining, Data Warehouse, KDD, Recent techniques

### I. INTRODUCTION

The development of Information Technology lead to a large amount of data generation, and are stored in the repositories, thus most of the organizations become 'data rich and information poor' The increased use of computers in a variety of business resulted in the massive collection of data, and this mass collection of data led to the creation of structured databases and Database Management Systems (DBMS) [1]. The conversion of huge volume of data into highly valued information, which are used to aid in decision making in various business processes. The extraction of knowledge from large sets of data help the business organizations to focus on most important information's in their data warehouses [2]. The DBMSs efficiently manage enormous amount of data, efficient and effective retrieval of required information from the large collection whenever required and contributes to massive gathering of all sorts of information.

Data mining is the science of extracting useful information from large data sets or databases. The different kinds of information which could be mined are: i) Business Transactions Data, ii) Scientific Data, iii) Media and Personal Data, iv) Surveillance video and images, v) Spatial data, vi) World Wide Web repositories [3] etc. Data mining is used for building empirical models, which are based not on the underlying theory about the process or mechanism that generated the data, instead is data-driven, and it provides a description of the observed data. Its fundamental objective is to provide insight and understanding about the structure of the data and its important features, and to discover and extract patterns contained in the data set. Data mining brings together a multitude of disciplines, such as database systems, statistics, artificial intelligence, data visualization, and others. The discovered knowledge can be applied to Information Management, Query Processing, Decision-Making, Process Control and many other applications [4].

The Data Mining techniques are the result of long process of research and product development. Data Mining is also known as Knowledge Extraction, Data/Pattern Analysis, Data Archeology, Data Dredging, Information Harvesting, Business Intelligence etc. Data Mining is not specific to one kind of data or media; instead, it is applicable to any kind of information in the repository.

### II. TECHNIQUES USED IN DATA MINING

Data Mining is a convergence of three key technologies i.e. Increasing Computing Power, Statistical and Learning Algorithms, and Improved Data Collection and Management. The idea of Data Mining is drawn from Artificial Intelligence, Machine Learning, Pattern Recognition, Statistics and Database Systems [5].

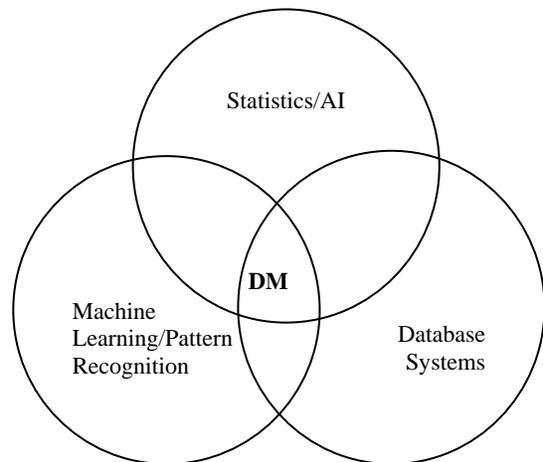


Figure 1: Data Mining Techniques

The Data Mining is originated, as the traditional techniques, and may be unsuitable for due to Enormity, High Dimensionality, Heterogeneous and distributed nature of data

in the business applications. Since the diversity of applications contributing to Data Mining, Data Mining research is expected to generate a wide variety of Data Mining Systems [6]. The techniques used in Data Mining are: i. Decision Trees: A Predictive Mode that makes a prediction based on a series of decision, ii. Artificial Neural Networks: Systems that Detect Patterns, Make Predictions and Learn, iii. Rule Induction: Techniques for exposing all possible predictive patterns in a database for knowledge discovery in unsupervised learning systems. iv. Genetic Algorithms: Optimization Techniques that use process such as genetic combination, mutation and natural selection in a design based on the concepts of evolution, v. Association Analysis: Tool for discovering the relations between variables in large databases, vi. Cluster Analysis: Tool for data analysis [7].

### III. DATA MINING PROCESS AND STANDARDS

The Data Mining process is interactive and iterative which involves various steps: i. Problem Identification and Definition, ii. Creating a target data set, iii. Data Cleaning and Preprocessing, iv. Data Reduction and Projection, v. Data Mining to Extract Patterns/Models vi. Result Interpretation and Evaluation, vii. Using Discovered Knowledge [8].

Various Data Mining standards have come in recent days up and are used by many of the data mining vendors, and the developer of the applications based on Data Mining techniques. The Data Mining standards are mainly concerned with one or more mining issues. The Data Mining Process has been formalized by an industry group called CROSS Industry Standard Process for Data Mining (CRISP-DM). The main steps in this process include: i) Business (or Problem) Understanding, ii) Data Understanding, iii) Data Preparation (includes Data Cleaning and Pre-Processing), iv) Modeling (applying Machine Learning and Data Mining Algorithms), v) Evaluation (check for the performance), vi) Deployment, and vii) Monitoring.

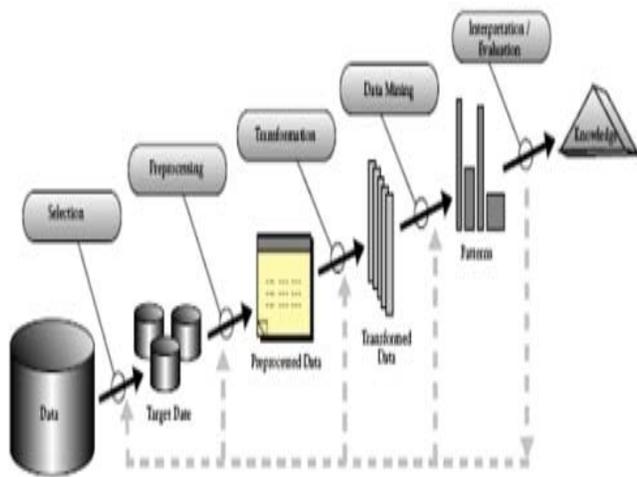


Figure 2: Data Mining Process of Knowledge Discovery

Many researchers in Data Mining and Knowledge Discovery are creating new, more automated methods for the discovery of knowledge to meet the requirements of the

industries [9]. The business needs keep growing and require customer-information analysis and customer preferences prediction. Data Mining standards are categorized into process standards that address different aspects or dimension of Data Mining Application Development [10]. They are:

- a. **Process Standards:** CROSS Industry Standard Process for Data Mining (CRISP-DM) - Captures Data Mining Process: begins with business problem and ends with the deployment of knowledge gained in the process.
- b. **XML Standards:** Predictive Model Markup Language and (PMML) and Common Warehouse Model for Data Mining (CWM-DM) - Model for representing Data Mining and Statistical Data, and Model for Meta Data that specifies metadata for building settings, model representations, and results from model operations Models are defined through the Unified Modeling Language respectively.
- c. **Protocol for Transport of Remote and Distributed Data:** Data Space Transport Protocol (DSTP) - DSTP is used for distribution, enquiry and retrieval of data in a data space.
- d. **Model Scoring Standard:** Predictive Scoring and Update Protocol (PSUP) - PSUP can be used for both on line real time scoring and updates as well as scoring in an off line batch Environment.
- e. **Web Standards:** XML for analysis (XMLA) - Standard web service interface designed specifically for Online Analytical Processing and Data Mining functions (uses Simple Object Access Protocol), Semantic Web - Semantic Web provides a framework to represent information in machine processable form and can be used to extract knowledge from Data Mining Systems, and Data Space - Provides an infrastructure for creating a web of data. It is built around standards like XML, DSTP, PSUP, and helps to handle large data sets, which are present on remote and distributed locations.
- f. **Grid Standards:** Open Grid Service Architecture - Developed by Globus - This standard talk about Service based open architecture for distributed virtual organizations. It will provide data mining engine with secure, reliable and saleable high bandwidth access to the various distributed data sources and formats across various administrative domains.

### IV. DATA MINING APPLICATIONS

Data Mining is used for wide variety of purposes both in private and in public sectors [11]. Its areas of applications are:

- a. **Risk Analysis:** Insurance Companies and Banks use Data Mining for Risk Analysis. An Insurance Company searches in its own insurants and claims databases for relationships between personal characteristics and claim behavior. The companies are mostly interested in the characteristics of insurants with a highly deviating claim behavior, premium policy etc.\
- b. **Direct Marketing:** Data Mining techniques are used to discover the relationship between one's personal characteristics.

- c. **Production Quality Control:** Data Mining can be used to determine the combinations of production factors that influence the quality of the product [12]. This information allows the process engineers to explain why certain products fail the final test and to increase the quality of the production process.
- d. **Surveillance:** Data Mining techniques could be used by various public and private agencies to monitor the behavior, activities and/or changing information, usually of individuals or groups.
- e. **Science:** The Data Mining techniques are used in Scientific and Engineering fields such as astronomy, bio-informatics etc., and in Medical field in the discovery of drugs.
- f. **Security:** Data Mining techniques are mostly used in the Security and Fraud Detection applications. Many Financial organizations, Telecom companies have developed the applications based on Data Mining and to detect the frauds[13].
- g. **Criminal and Forensic Investigations:** The data Mining technique has been used in the forensic department to identify the fingerprint, patterns, images taken from a crime scene, images taken from suspects or images to be used as control samples etc. Data Mining techniques are used in the investigation of the domestic violence, death, threat to person or property, fraud or any other crimes that require complex detection.
- h. **Tax Administration:** Data Mining techniques have been used by the Income Tax department to identify the audit targets to enforce the tax payment.
- i. **Aviation Safety:** Discovery of the accidental trends in aviation, and precaution measure to ensure to prevent such type of event does not happen in future.

## V. GENERAL AND TECHNICAL ISSUES

The key issues raised by the Data Mining technology is not a business or technological one, instead a social one [13]. The major issues identified are:

- a. **Individual Privacy:** Data Mining techniques helps to analyze the routine business transactions and gather significant amount of information about individuals buying habits and preferences.
- b. **Data Integrity and Database Structure:** The key challenges in Data Mining implementations are integrating conflict or redundant data from different sources, and the optimal set up of the Relational Databases over Multidimensional databases.
- c. **Cost:** Data Mining and Data Warehousing tend to self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries, and hence are more expensive.
- d. **Outliers:** Many data entries may not fit into the Data Mining Models, and may lead to issues with large databases. The inclusion of such outliers in a Data

Mining Models may not behave well for data that are not outliers.

- e. **Interpretation and Visualization of Results:** Data Mining experts are required to correctly interpret the results of Data Mining outputs, and proper visualization is a must to easily view and understand the output of Data Mining algorithms.
- f. **Large Datasets:** Most of the modeling applications grow exponentially on the data set size, and hence are too inefficient for larger datasets [14]. The algorithms designed for small datasets may create problems when applied to the massive datasets associated with the Data Mining.
- g. **Missing and Irrelevant Data:** The pre-processing phase of KDD, replaces the missing data with the estimated one. The various approaches used to handle the missing data may lead to invalid results in the Data Mining task. Some attributes in the database might not be of interest to the Data Mining task being developed.
- h. **Changing Data:** Most Data Mining algorithms do assume a static database, which is often not true. The algorithms must be able to run anytime the database changes.
- i. **Integration:** The KDD process is not currently integrated into normal data processing activities [15]. KDD requests may be treated as special, unusual or one-time needs, thus making them inefficient, ineffective and not general enough to be used on an ongoing basis. Integration of Data Mining functions into traditional DBMS systems is certainly a desirable goal.

## VI. CONCLUSION

Data Mining is a mixture of concepts and algorithms from Machine Learning, Statistics, Artificial Intelligence and Data Management. With the emergence of Data Mining Techniques, researchers and practitioners started applying this technology on data from various fields such as banking and finance, retail, marketing, insurance, fraud detection, science and technology etc. to discover the interesting patterns related to their business needs. The usage of Data Mining techniques in the industries opens new possibilities in decision making to the top management, and to advisory and control systems. The applications of Data Mining in industries have opened up new possibilities in decision-making and predict the business trends.

## VII. REFERENCES

- [1] Shen-Ming Gu, Yun Zheng, Lin-Ting Guan, Yue-ting Zhuang: "The Explore of Some Cases with Data Mining Techniques", International Conference on Electronic Computer Technology, 2009, 978-0-7695-3559-3/09, DOI 10.1109/ICECT.2009.47.
- [2] Mrs. Bharati M. Ramageri: "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305, ISSN : 0976-5166.
- [3] B.N Lakshmi, G.H. Raghunandan: "A Conceptual Overview of Data Mining", Proc. National Conference on Innovation in Emerging Technology, 2011, pp. 27-32.

- [4] Jochen Hipp, Ulrich Günter, and Udo Grimmer: “Data Quality Mining, unpublished.
- [5] Alex Berson, Stephen Smith, and Kurt Thearling: “Data Mining Techniques- An Overview”, unpublished.
- [6] Ho Tu Bao: “Introduction to Knowledge Discovery and Data Mining”, Institute of Information Technology, National Center for Natural Science and Technology, unpublished.
- [7] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth :“From Data Mining to Knowledge Discovery in Databases”, American Association for Artificial Intelligence, 738-4602-1996, AI Magazine: pp, 37-54
- [8] Dr. Sohail Asghar and Khalid Iqbal: “Automated Data Mining Techniques: A Critical Literature Review”, International Conference on Information Management and Engineering, 978-0-7695-3595-1/09, 2009.
- [9] Bob Mann, Roy Williams, Malcolm Atkinson. Ken Brodlie, Amos Storkey, Chris Williams: “Scientific Data Mining, Integration, and Visualization”, UK e-Science Technical Report Series, ISSN 1751-5971.
- [10] Arati Kadav, Jaya Kawale, Pabitra Mitra “Data Mining Standards”, unpublished
- [11] Waldemar Wojcik and Konrad Gromaszek: “Data Mining Industrial Applications”, Lublin University of Technology, Poland, pp. 431-442.
- [12] Matthew Sparkes: “Literature Survey: Successful Applications of Data Mining”, Jan 2004.
- [13] C. Lakshmi Devasena, T Sumathi, and Dr. Hemalatha: “An Experiential Survey on Image Mining Tools Techniques and Applications”, International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-339, Vol. 3 No. 3 Mar 201, pp. 1155-1167.
- [14] Edith Namikka and George J. Gibbon: “Identification of Data Mining Techniques for Industrial Process Analysis and Control”, unpublished
- [15] Radu Platon and Mouloud Amazouz: “Application of Data Mining Techniques in Industrial Process Optimization”, A Literature and Software Review, July 2007.