DATA MINING STAGES AND ITS CRUCIAL CONCEPTS

Swati N. Sonune*

Thamraj N. Ghorsad**

Rupali Y. Bisne***

ABSTRACT

The term data mining refers to the analysis of large observational data sets with the goal of finding unsuspected relationships. A data set can be "large" either in the sense that it contains a large number of records or that a large number of variables is measured on each record. Data Mining as an analytic process designed to explore data in search for consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has most direct business applications. In an uncertain and highly competitive business environment, the value of strategic information systems such as these are easily recognized however in today's business environment, efficiency or speed is not the only key for competitiveness. These types of huge amount of data's are available in the form of tera- to peta-bytes which has drastically changed in the areas of science and engineering. We need techniques called the data mining which will transform in many fields. In This paper we discuss the concept of data mining Stages, and some crucial concepts in it like stacked generalization, voting, Feature Selection, Bagging etc.

Keywords: Data mining Stages, Data mining crucial concepts, stacked generalization, Feature Selection, Bagging, Boosting, etc.

* M.E., Department of Computer Science & Engg., RGPV University, Bhopal.

** M.Tech. Department of Computer Science & Engg., RGPV University, Bhopal.

*** MCA., Sarhad College of Art Commerce & Science, Pune.

1. INTRODUCTION

In information era, Data mining is one of the most important steps of the knowledge discovery in databases process and is considered as significant subfield in knowledge management. Research in data mining continues growing in business and in learning organization over coming decades [1]. The human beings are used in the different technologies to adequate in the society. Each and every day the human beings are using the vast data and these data are in the different fields .It may be in the form of documents, may be graphical formats may be the video, may be records (varying array). As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data. As and when the customer will required the data should be retrieved from the database and make the better decision .This technique is actually we called as a data mining [2]. This paper presents the introduction of data mining and the process of data mining which consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and it is concluded with (3) deployment (i.e., the application of the model to new data in order to generate predictions). And also describes some crucial concepts in data mining such as stacked generalization, voting, Feature Selection, Bagging, Boosting, Meta-Learning Drill-Down Analysis, Deployment, etc.

The rest of this paper is organized as follows: Section 2 presents Overview of Data Mining. Section 3 shows stages for data mining process. Sections 4 describe The Crucial Concepts in Data Mining. Section 5 shows application. Section 6 shows conclusions; finally references and acknowledgement are presented in the last section.

2. OVERVIEW OF DATA MINING

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

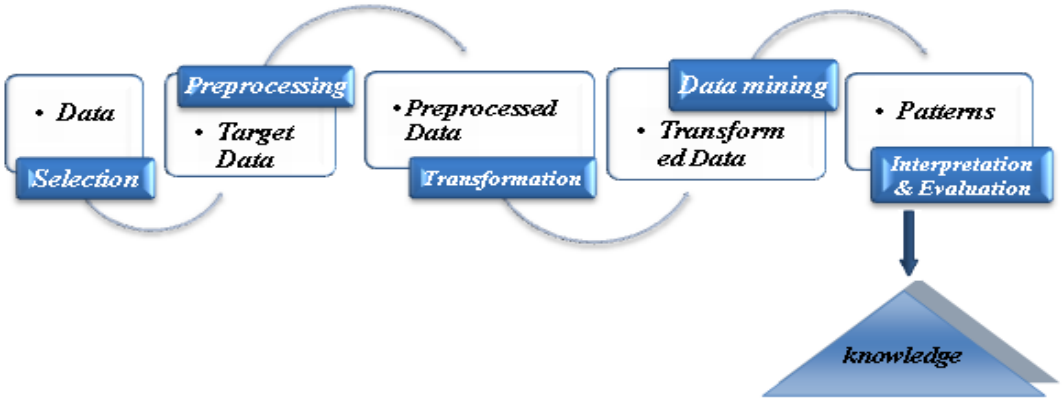

Figure.1. Knowledge discovery Process

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 [1].

Based on figure 1, KDD process consists of iterative sequence methods as follows [1, 7]:

1. Selection: Selecting data relevant to the analysis task from the database.

2. Preprocessing: Removing noise and inconsistent data; combining multiple data sources.

3. Transformation: Transforming data into appropriate forms to perform data mining.

4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; extracting data patterns.

5. Interpretation/Evaluation: Interpreting the patterns into knowledge by removing redundant or irrelevant patterns; translating the useful patterns into terms that human understandable.

3. OVERVIEW OF DATA MINING STAGES

The following three stages are involved in data mining:

Stage 1: Exploration. This stage usually starts with data preprocessing, in which a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined in an acceptable timeframe. Pre-processing is essential to analyze the multivariate datasets before data mining. The target set is then cleaned. Data cleaning removes the

observations containing noise and those with missing data. After that data transformation, selecting subsets of records are performed. In order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

Stage 2: Model building and validation. This stage involves considering various models and choosing the best one based on their predictive performance. Competitive evaluation of models means applying different models to the same data set and then comparing their performance to choose the best. A model is typically rated according to two aspects i.e. Accuracy and Understandability. These aspects often conflict with one another. Validation of the model requires that you train the model on one set of data and evaluate on another independent set of data.

Stage 3: Deployment. A model is built once, but can be used over and over again. In this final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome. The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

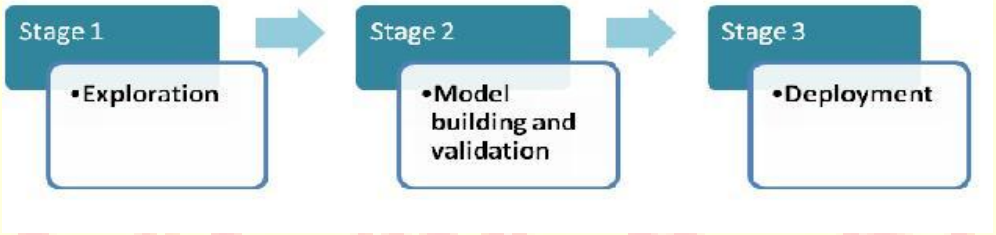

Figure.2. Stages of Data Mining

4. CRUCIAL CONCEPTS IN DATA MINING

4.1. Feature Selection

Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the choice of

attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis.

4.2. Bagging (Voting, Averaging)

Bagging is a variance reduction method for model building. That is, through building multiple models from samples of the training data, the aim is to reduce the variance. Bagging is a technique generating multiple training sets by sampling with replacement from the available training data. Bagging is also known as bootstrap aggregating. The concept of bagging applies to the area of predictive data mining, to combine the predicted classifications from multiple models, or from the same type of model for different learning data.

4.3. Boosting

The concept of boosting applies to the area of predictive data mining, to generate multiple models or classifiers, and to derive weights to combine the predictions from those models into a single prediction or predicted classification. Boosting algorithms build multiple models from a dataset, using some other model builders, such as a decision tree builder. The basic idea of boosting is to associate a weight with each observation in the dataset. A series of models are built and the weights are increased (boosted) if a model incorrectly classifies the observation. The weights of such entities generally oscillate up and down from one model to the next. The final model is then an additive model constructed from the sequence of models, each model's output weighted by some score. Boosting is also susceptible to noise.

4.4. Stacking (Stacked Generalization)

The concept of stacking applies to the area of predictive data mining, to combine the predictions from multiple models. It is particularly useful when the types of models included in the project are very different. Stacking is the abbreviation that refers to Stacked Generalization. The main idea of Stacking is to combine classifier from different learners such as decision trees, instance-based learners, etc. Since each one uses a different knowledge representation and different learning biases, the hypothesis space will be explored differently, and different classifier will be obtained. Thus, it is expected that their errors will not be perfectly correlated, and that the combination of classifier will perform better than the base classifiers.

4.5. Meta-Learning

The concept of meta-learning applies to the area of predictive data mining, to combine the predictions from multiple models. Meta-learning is the study of principled methods that exploit meta-knowledge to obtain efficient models and solutions by adapting machine learning and data mining processes. Meta-learning provides one such methodology that allows systems to become more effective through experience. It shows how this knowledge can be reused to select, combine, compose and adapt both algorithms and models to yield faster, more effective solutions to data mining problems. It can thus help developers improve their algorithms and also develop learning systems that can improve themselves. This meta-learning will be of interest to researchers in business and graduate students in the areas of machine learning, data mining and artificial intelligence.

4.6. Drill-Down Analysis

The concept of drill-down analysis applies to the area of data mining, to denote the interactive exploration of data, in particular of large databases. The process of drill-down analyses begins by considering some simple break-downs of the data by a few variables of interest (e.g., Gender, geographic region, etc.). Various statistics, tables, histograms, and other graphical summaries can be computed for each group.

4.7. Predictive Data Mining

The term Predictive Data Mining is usually applied to identify data mining projects with the goal to identify a statistical or neural network model or set of models that can be used to predict some response of interest. For example, a credit card company may want to engage in predictive data mining, to derive a model or set of models that can quickly identify transactions which have a high probability of being fraudulent. Other types of data mining projects may be more exploratory in nature, in which case drill-down descriptive and exploratory methods would be applied.

4.8. Data Preparation

Data preparation and cleaning is an often neglected but extremely important step in the data mining process. The old saying "garbage-in-garbage-out" is particularly applicable to the typical data mining projects where large data sets collected via some automatic methods (e.g., via the Web) serve as the input into the analyses.

4.9. Data Reduction

The term Data Reduction in the context of data mining is usually applied to projects where the goal is to aggregate or amalgamate the information contained in large datasets into manageable information nuggets. Data reduction methods can include simple tabulation, aggregation or more sophisticated techniques like clustering, principal components analysis, etc.

4.10. Text Mining

Text mining generally consists of the analysis of (multiple) text documents by extracting key phrases, concepts, etc. and the preparation of the text processed in that manner for further analyses with numeric data mining techniques e.g., to determine co-occurrences of concepts, key phrases, names, addresses, product names, etc. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.

4.11. Machine Learning

A learning system uses sample data to generate an updated basis for improved performance on subsequent data from the same source and expresses the new basis in intelligible symbolic form. Machine learning, computational learning theory, and similar terms are often used in the context of Data Mining, to denote the application of generic model-fitting or classification algorithms for predictive data mining.

5. APPLICATION

Data mining is a great way to create new business opportunities. Figure.3 shows Data Mining with Business Application. Data mining technique is used in MBA(Market Basket Analysis).When the customer want to buying some products then this technique helps us finding the associations between different items that the customer put in their shopping buckets. Here the discovery of such associations that promotes the business technique .In this way the retailers uses the data mining technique so that they can identify that which customers intension (buying the different pattern).In this way this technique is used for profits of the business and also helps to purchase the related items.

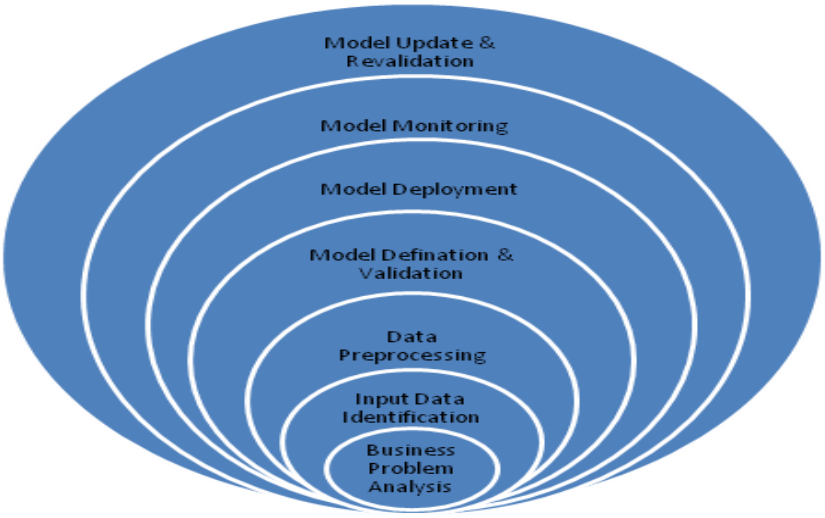

Figure.3. Data Mining With Business Application

6. CONCLUSIONS

Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology. In this paper we describe the Data Mining Stages and its Crucial Concepts**.** We have shown that data mining can be integrated into KM framework and enhanced the KM process with better knowledge. This review would be helpful to researchers to focus on the various issues of data mining.

7. REFERENCES

[1] Tipawan Silwattananusarn and Assoc.Prof. Dr. Kulthidatuamsuk "Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.

[2] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[3] Bharati M. Ramageri "Data Mining Techniques And Applications"/ Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305.

[4] Marie Gaudard, Ph. D., Philip Ramsey, Ph. D., Mia Stephens, MS North Haven Group "Interactive Data Mining and Design of Experiments" March 2006.

[5] Data Mining: A Conceptual Overview by J. Jackson, Communications of the Association for Information Systems (Volume 8, 2002) 267-296.

[6] Osmar R. Zaïane, "Introduction to Data Mining", CMPUT690 Principles of Knowledge Discovery in Databases University of Alberta, Department of Computing Science, 1999.

[7] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases.AI Magazine, 17(3), 37-54.

[8] Han, J. &Kamber, M. (2012). Data Mining: Concepts and Techniques. 3rd.ed. Boston: Morgan Kaufmann Publishers.

[9] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[10] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[11] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

8. ACKNOWLEDGMENTS

The authors wish to thank Ashwini U. Dakhode, for providing pictures, comments, and other helpful information. We thank anonymous reviewers for their helpful comment