WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Review Study on Data Mining

Kamal Dhar Dwivedi[1], Akhil Pandey[2], Dileep Tiwari[3]

[1]Scholar, Department of MCA, IET, Dr R M L Avadh University Ayodhya
[2]Scholar Department of MCA, IET, Dr R M L Avadh University Ayodhya
[3]Assistant Professor, Department of MCA, IET, Dr R M L Avadh University, Ayodhya

ABSTRACT

With progression in technology specifically in last three decades or so, an enormous magnitude of information has been transitioned into a digital form, which resulted in formation of enormous data repositories. With accrual of information in these repositories a challenge persisted as how to extract meaningful knowledge from it. Data mining as a tool was used to tackle the situation. Data mining considered as stepping stone to procedure of knowledge discovery in databases, this is a procedure of extracting hidden information from enormous sets of databases to excavate eloquent patterns and rules. Data mining has now become an indispensable component in almost every field of human life. The present article provides an analysis of the available literature on data mining. The concept of data mining as well as its various methodologies is summarized. Some applications, tasks and issues related to it have also been illustrated.

Keywords Data mining, Knowledge discovery in database, Knowledge base.

Introduction

he readiness of ample magnitude of data in almost every field and the desire to excerpt beneficial information and knowledge from it substantiated as main motivation that pulled the eyes of researchers in recent past towards data mining. The information and knowledge extracted can be momentously useful for the applications ranging from small business management to complex engineering design to science exploration. Data mining is the analysis and scrutiny of mammoth data sets, with an aim to uncover significant pattern and rules that were previously unidentified. The core aim is exploiting the data processing power of computer with human's capability to perceive patterns [1]. The epoch of data mining applications was conceived in the year 1980 predominantly by research driven tools engrossed on solo chore [2]. In recent times data mining is being dominant among Statisticians, MIS communities' data analysts. It was during the first workshop on KDD in 1989 Piatetsky-Shapiro coined the phrase "knowledge discovery in database". The recognition of data mining and KDD shouldn't be astonishing, considering the scale of data been collected from various obtainable sources, the collected data is magnanimous to be examined manually and many a times automatic data analysis supported by classic statistics and machine learning could face concerns once the procedure is hefty and collected knowledge comprises of problematical entities. The bellicose, massive volume of data collected from numerous sources and kept in vast and various repositories. The data collection thus exceeds the human aptitude for analysis without a powerful analysis tool, as a consequence these repositories become 'data vaults', that are not often visited. As decision makers lack tools to extract the treasurable knowledge mounted within enormous volume of data, hence vital decisions lack the utilization of information rich data. Data mining tools perform analysis of data and determine the vital patterns that were earlier anonymous [3]. As every arena of human life has become data intensive which stemmed in making data mining as an indispensable constituent? Though Data mining and KDD have been used conversely yet KDD can be seen as an inclusive procedure of extracting beneficial knowledge from data, while as Data mining can be seen as core of KDD, which includes Algorithms that explore data, build models and discovery unknown patterns.

Data Mining Tasks

Data mining tasks are grouped into two main categories:

- Predictive
- Descriptive

These two are considered primary objectives of data mining. Fayyad et.al 1996 defines six main functions of data mining:

- Classification
- Regression

- Clustering
- Dependency modeling
- Deviation detection.
- Summarization.

Classification, regression and anomaly detection categorized under predictive category while as clustering, Dependency modeling categorized under descriptive category. Predictive model forecasts using some variable in dataset so as to predict unknown values of other relevant variable while as descriptive model classifies patterns or relationship and encompasses human understandable pattern and trends in [4].

Classification: classification is among the classical data mining technique that is established on machine learning. It finds mutual properties amongst a set of objects in a database and categorizes them into diverse classes in accordance with the classification model. Its main objective is to scrutinize the training data and develop an accurate description or model for each class using feature available in data. This method uses mathematical techniques like decision trees, Neural networks and statistics [5].

Regression: It is one among data mining techniques that defines the association between dependent and independent variables. Prediction is accomplished with regressions support. Statistically regression is the mathematical model that constitutes connection amongst the values of dependent variable and values of other predictor or independent variable. In regression the predicted variable may be continuous variable. In regression real valued prediction variables are mapped from items of a learning function. Statistical regression, Neural Network, Support Vector Machine regression is some of the commonly used regression strategies. More complex techniques such as Logistic regression, Decision Trees or Neural Networks could also be utilized for forecasting future values, these techniques could also be combined for attainment of better result.

Clustering: It is a data mining technique which groups physical or abstract objects into classes of similar objects. Clustering is a method of dividing set of data (records/tuples/objects/samples) into several groups (clusters) based on foreordains similarities. The principal aim of clustering is finding groups (clusters) of objects based on affinity so that within individual cluster there is great resemblance to each other while clusters are diverse enough from one another. In machine learning terminology, clustering is a form of unsupervised learning.

Dependency Modeling (Association Rule Mining): it's amongst the finest acknowledged data mining techniques and is categorized under unsupervised data mining technique, which aims at finding connections or relations between items or records belonging to a large dataset and labels significant dependencies among variables [6]. Association rule mining is implication of the form X

→ Y, where x and y are distinct items or item sets manufacturing if-then statements regarding attribute values. In market basket analysis this rule has been commonly used, it tries to analyze customers purchasing certain items and provides insight into the combinations customer frequently purchases together.

Anomaly detection: synonymous to is its name it deals with the unearthing of most substantial changes or aberrations from the standard behavior. Summarization: Though not amongst the techniques of data mining, but is a resultant of these techniques and deals with determining a compact depiction for a subset of data synonymously referred to as generalization or description. Sequential Patterns: Sequence discovery is a data mining technique that is used to determine sequential patterns or associations or regular events/trends between variable data fields over a business period.

Data Mining Life Cycle

The life cycle of a data mining project consists of six phases [8]. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data Understanding: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information [5].

Data Preparation: It covers all activities to construct the final dataset from the initial raw data.

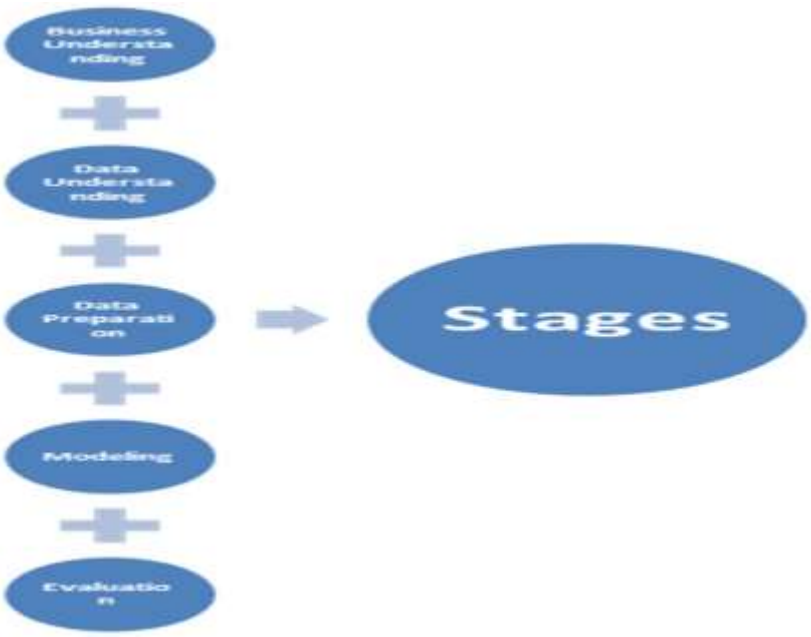

Fig 1: Stages for data mining approach

Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment: 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. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Conclusion

Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation.

References

1. Domingos, Pedro, and Geoff Hulten. "Mining highspeed data streams." Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.
2. Hand, David J., Heikki Mannila, and Padhraic Smyth. Principles of data mining. MIT press, 2001.
3. Rygielski, Chris, Jyun-Cheng Wang, and David C. Yen. "Data mining techniques for customer relationship management." Technology in society 24.4 (2002): 483- 502.
4. Keogh, Eamonn, Stefano Lonardi, and Chotirat Ann Ratanamahatana. "Towards parameter- free data mining." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
5. Alhammady, Hamad. "A novel approach for mining emerging patterns in data streams." Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on. IEEE, 2007.
6. Džeroski, Sašo. Relational data mining. Springer US, 2009. [13]. Venkatadri, M., and Lokanatha C. Reddy. "A review on data mining from past to the future." International Journal of Computer Applications 15.7 (2011): 19-22.

7. Silwattananusarn, Tipawan, and Kulthida Tuamsuk. "Data mining and its applications for knowledge management: a literature review from 2007 to 2012." arXiv preprint arXiv:1210.2872 (2012).

8. Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266