Purva Mimaansa | Refereed | Peer Reviewed
ISSN: 0976-0237 | Volume 12, Issue March 2021 | Impact Factor 4.115

A DATABASE PERSPECTIVE ON KNOWLEDGE DISCOVERY

MINAKSHI GUPTA
Assistant Professor in Dept. of Computer Sc. & App, Sanatan Dharma College, Ambala Cantt

ABSTRACT

Knowledge Discovery in Databases is the process of finding knowledge in massive amount of data where data mining is the core of this process. Data mining can be used to mine understandable meaningful patterns from large databases and these patterns may then be converted into knowledge. Data mining is the process of extracting the information and patterns derived by the KDD process which helps in crucial decision-making. Data mining works with data warehouse and the whole process is divided into action plan to be performed on data: Selection, transformation, mining and results interpretation. In this paper, we have reviewed Knowledge Discovery perspective in Data Mining and consolidated different areas of data mining, its techniques and methods in it

KEYWORDS: Decision, Knowledge, Mining, Selection, Warehouse

INTRODUCTION

Knowledge Discovery in Databases (KDD) is the process of finding useful knowledge from large dataset. Data preparation, pattern search, knowledge evaluation and refinement are steps of KDD (Jianhua: 1998). The process of Knowledge Discovery consists of Data Cleaning, Data Integration, Data Selection, Transformation, Data Mining and Pattern Evaluation Phases (Jiawei: 2000). Data mining (DM) is the process where data is analyzed and summarized into useful information. In short, data mining is process of deriving patterns from large databases DM analyses large dataset to extract hidden patterns such as similar groups of data records using clustering technique. This data is used for machine learning and predictive analysis. DM works to analyze data stored in data warehouses and results in effective decision making. "DM is the search for valuable information in large volumes of data". According to Technology Forecast it is the process of extracting previously unknown, useful information which include knowledge, association rules, pattern finding, statistical and mathematical techniques. Query languages or graphical user interface are required to express the DM requests and discovered information , so that results obtained from the DM Engine become understandable and usable for end users.

Data Mining introduced in the year 1990's and it is the combination of many disciplines like database management systems (DBMS), Statistics, Artificial Intelligence (AI), and Machine Learning (ML).Data Mining produce useful patterns by applying algorithmic methods on observational data.

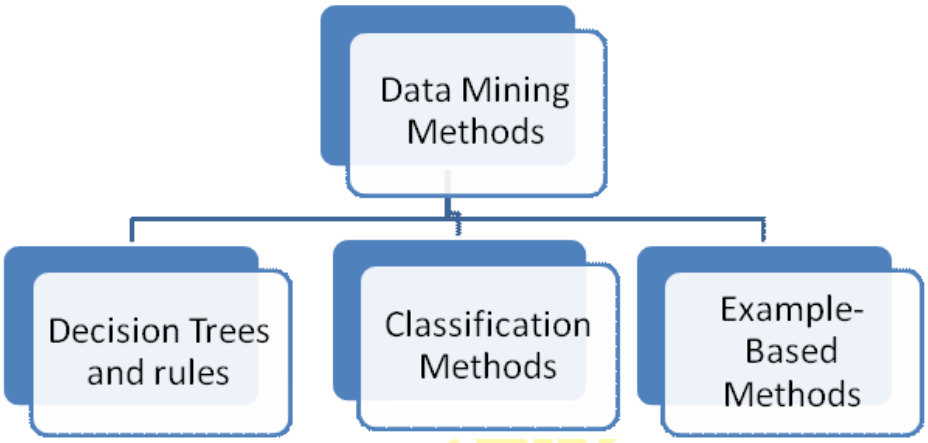
KNOWLEDGE DISCOVERY IN DATABASES

Data mining and knowledge discovery in databases are related to each other and to other related fields such as machine learning, statistics, and databases. Data Mining is one of the steps in the overall process of KDD that consists of collection and pre-processing of data, data mining, interpretation, evaluation of discovered knowledge and finally post processing. The KDD field's basic objective is to make data meaningful by developing methods and techniques of mining but problem being faced by

the KDD process is to map huge and heterogeneous data into understandable, more abstract and useful form. The phrase knowledge discovery in databases emphasizes that knowledge is the end product of a data-driven discovery. The data-mining step of KDD relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data. Data warehousing is one of the fields of databases which helps in business analytics and decision support. Data warehousing helps set the stage for KDD in two ways:

(1) data cleaning and

(2) data access. Approach followed for analysis of data warehouses is called online analytical processing (OLAP)



Primary goals of data mining in practice are prediction and description. In prediction some variables and fields in the database are used to predict unknown values of other variables of interest, and description helps in finding human-understandable patterns describing the data, that, "Classification is learning a function that maps (classifies) a data item into one of several predefined classes". Those classification methods of Data mining are used as part of knowledge discovery applications which includes classifying trends in financial markets, education and identifying objects of interest from large dataset of images. Regression is a predictive technique that maps data item to a prediction variable. Clustering is a descriptive task where we identify a finite set of categories or clusters to describe the data. For example, identifying those students who are short in attendance and has shown poor performance in the sessionals. Cheese man and Stutz in suggested that examples of clustering applications in a knowledge discovery context include discovering similar groups. Summarization involves methods like calculating mean and standard deviations. There are some methods which involve deriving of abstract rules, visualization techniques, and the discovery of functional relationships between variables. Summarization techniques are often applied to interactive exploratory data analysis and automated report generation.

Decision trees and rules

Decision Trees are useful for multiple variable analyses. They split a data set into branch-like

segments.

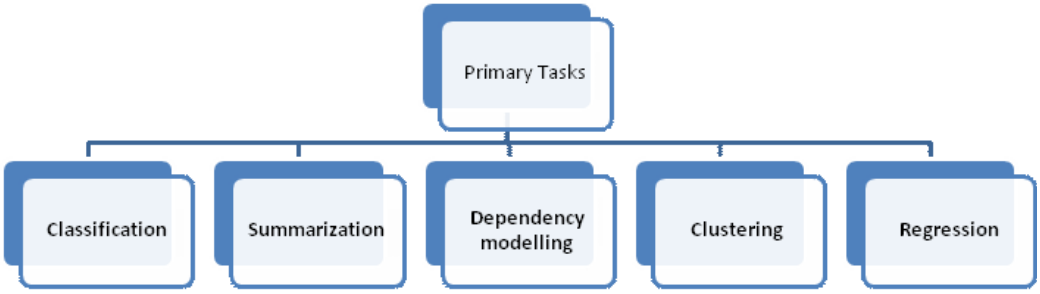
Classification methods

These methods consist of techniques for prediction. Examples includes feed forward neural networks, adaptive sp-line methods, projection pursuit regression, Multi-Layer Perceptions, Generalized Linear Models Bayesian Networks, Decision Trees, and Support Vector Machines.

Example-based methods

Predictive analysis on new examples will be derived from those examples in the model for which predictions are known. Techniques include nearest neighbour classification and re-egression algorithms and case-based reasoning systems.

Primary Tasks of Data Mining



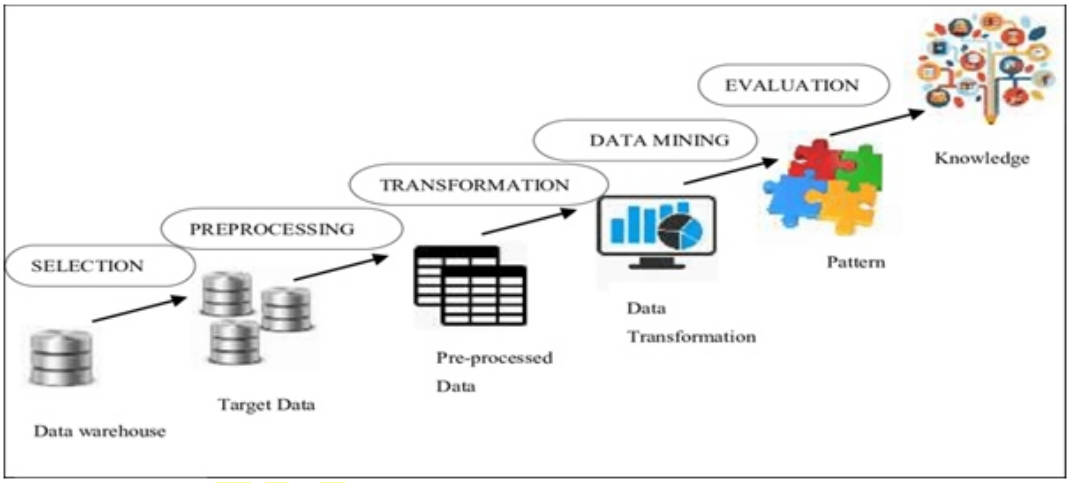
1. Classification - discovering the portrayal of predefined classes and characterizes an information thing into one of them
2. Regression-maps an information thing to a genuine esteemed forecast variable
3. Clustering-recognizing a limited arrangement of classes or bunches to depict the information.
4. Dependency demonstrating finding a model which portrays critical conditions between factors.
5. Summarization-tracking down a minimal portrayal for a subset of information.
6. Deviation and change identification discovering the main changes in the information.

KNOWLEDGE WAREHOUSE

Knowledge warehouse (KW) can be thought of as an "information repository". The knowledge warehouse consists of knowledge components (KCs) that are defined as the smallest level in which knowledge can be decomposed. Knowledge components (objects) are catalogued and stored in the knowledge warehouse for reuse by reporting, documentation, execution the knowledge or query and reassembling which are accomplished and organized by instructional designers or technical writers. The idea of knowledge warehouse is similar to that of data warehouse. As in the data warehouse, the knowledge warehouse also provides answers for ad-hoc queries, and knowledge in the knowledge warehouse can reside in several physical places (Jackson: 2002). A knowledge warehouse (KW) is the component of an enterprise's knowledge management system. The

knowledge warehouse is the technology to organize and store knowledge. The knowledge warehouse also has logical structures like Computer programs and databases to store knowledge that are analogous to the system of tables that implement data storage in the data warehouse (Weiss & Indurkhya: 1998). The primary goal of a KW is to provide the knowledge worker with an intelligent analysis platform that enhances all phases of the knowledge management process (Sang Jun: 2001). Like the DW, the KW may be viewed as subject oriented, integrated, time-variant, and supportive of management's decision making processes. But unlike the DW, it is a combination of volatile and non-volatile objects and components, and, of course, it stores not only data, but also information and knowledge. The KW can also evolve over time by enhancing the knowledge it contains (Sang Jun: 2001). Knowledge warehouse provides the infrastructure needed to capture, cleanse, store, organize, leverage, and disseminate not only data and information but also knowledge (Padhraic: 2000).

KNOWLEDGE DISCOVERY PROCESS



Knowledge discovery in databases (KDD) is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. The term KDD is used to denote the overall process of turning low-level data into high-level knowledge. A simple definition of KDD is as follows: Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Knowledge Discovery has also been defined as the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data'. It is a process of which data mining plays an important role to extract knowledge from huge database (data warehouse). Data mining is the core part of the knowledge discovery in database (KDD) process (Technology Forecast: 1997).

The KDD process may consist of the following steps:

1) data integration,

2) data selection and data pre-processing,

3) Interpretation & assimilation. Data comes on; possibly from many sources therefore it is integrated

and placed in some common data store like data warehouse. Part of it is then selected and pre-processed into a standard format. This 'prepared data' is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of 'patterns'. These are then interpreted to give new and potentially useful knowledge. Although the data mining algorithms are central to knowledge discovery, they are not the whole story. The pre-processing of the data and the interpretation of the results are both of great importance.

DATA MINING AND KDD

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase knowledge discovery in databases was coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the AI and machine-learning fields. In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data-mining step (within the process) is a central point of this article. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, is essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns (William J. Frawley & others: 1992).

THE KDD PROCESS

The KDD process is interactive and iterative, involving numerous steps with many decisions made by the user. Brachman and Anand (1996) give a practical view of the KDD process, emphasizing the interactive nature of the process (Piatetsky-Shapiro: 2000). Here, we broadly outline some of its basic steps: First is developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer's viewpoint. Second is creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed. Third is data cleaning and pre-processing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes. Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found (Venkatadri: 2011). Fifth is matching the goals of the KDD process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clustering, and so on, are described later as well as. Sixth is exploratory analysis and model and hypothesis selection: choosing the data mining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for example, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall

criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities). Seventh is data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps. Eighth is interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models. Ninth is acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge. The KDD process can involve significant iteration and can contain loops between any two steps. The basic flow of steps (although not the potential multitude of iterations and loops) is illustrated in figure 1. Most previous work on KDD has focused on step 7, the data mining. However, the other steps are as important (and probably more so) for the successful application of KDD in practice (Joyce Jackson: 2002).

CONCLUSION

Data Mining is an upcoming field related to several well-established areas of research including e-learning, web mining, text mining etc. Data Mining Techniques are used to analyze Educational data and extract useful information from large amount of data. This paper presents review of the KDD and basic data-mining techniques so as to integrate research in this area. The KDD field is related to development of methods and techniques which make the data relevant. In Educational Sector software's and visualization techniques can be developed using Data Mining Techniques which not only predict student's performance in examinations as well as helps us to cluster those students who need special attention in their studies. Knowledge Discovery in Databases results in better decision-making related to latest technologies useful in classroom teaching as well as faculty enhancement programs and in-house trainings etc .Using data mining techniques we can achieve refined data from distributed databases. Data Mining is an efficient tool for improving institutional effectiveness and student learning. Knowledge ac-quires by Educational Data Mining not only help teachers to manage their classes, improve their teaching skills, students learning processes but also provide feedback to institutions to improve their infrastructures and quality. For making this approach successful and to increase its scope, more data can be collected from Educational Institutions and queries can be performed on it.

REFERENCES

Fan Jianhua, Li Deyi (1998). “An Overview of Data Mining and Knowledge” Discovery, J. of Comput. Sci. & Technol., Vol.13 No.4, Jul. 1998.

Han Jiawei, Micheline Kamber (2000). “Data Mining: Concepts and Technique”. Morgan Kaufmann Publishers.

Sang Jun Lee, KengSiau (2001). “A review of data mining techniques, Industrial Management & Data Systems”, 101/1, pp. 41-46.

Padhraic Smyth (2000). “Data Mining: Data Analysis on a Grand Scale”, July 6,2000.

Weiss S. & Indurkhya N. (1998). “Predictive Data Mining: A Practical guide”, Morgan Kaufmann.

Technology Forecast (1997), Price Waterhouse World Technology Center, Menlo Park, CA.

William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus (1992). "Knowledge Dis-covery in Databases: An Overview", AI Magazine Volume 13 Number 3 (1992) (© AAAI).

Piatetsky-Shapiro, Gregory (2000). "The Data-Mining Industry Coming of Age". IEEE Intelligent Systems.

Venkatadri M., Dr.Lokanatha C. Reddy (2011). "A Review on Data mining from Past to the Future", International Journal of Computer Applications (0975 – 8887), Volume 15– No.7, February 2011.

Joyce Jackson (2002). "Data Mining: A Conceptual Overview, Communications of the Association for Information Systems", Volume 8, 2002, pp. 267-296.

Brijesh Kumar Bhardwaj, Saurabh Pal (2011). "Mining Educational Data to Analyze Students Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6.