ISSN: 2456-9348
Vol (02) _Issue (05) Impact Factor: 4.520

iJETRM
International Journal of Engineering Technology Research & Management

A SURVEY ON DATA MINING AND KNOWLEDGE DISCOVERY APPROACH

Dr. A. K. Upadhyay[*1]
[*1]CSE, ASET, AUMP Gwalior
akupadhyay@gwa.amity.edu

ABSTRACT
The rapid growth and wide availability of digital data has led to an increased research activities in the field of data sciences. Traditional approaches to data management have achieved limited success, because they are unable to handle a huge amount of complex data. This article provides a comprehensive overview of existing problems, methods and future directions of computer science in the field of data mining using structured analysis of historical and modern methods. Data mining remains an important area of research in database systems. We present an overview of processing alternatives, storage mechanisms, algorithms, data structures and optimizations that allow data mining on large data sets. We focus on the calculation of well-known multidimensional statistical models and machine learning models.

Keywords:
Data mining, clustering, data computation, classification, association rule.

INTRODUCTION

The development of information technology has led to the creation of a large number of databases and huge data in various fields. “Research in databases and information technologies has led to the approach to storing and processing these valuable data for further decision making. Data mining is the process of extracting useful information and templates from immense data”[1,2]. It is also called the knowledge discovery process, the development of knowledge from data, the extraction of knowledge or the analysis of data patterns. Data output is a logical process that is used to search through a large number of data to find useful data. The purpose of this method is to find previously unknown templates. Once these templates are found, they can be used to make certain decisions for the development of their business. “Recently, a huge amount of data collected and stored in databases or dataset format, recently increased due to the interests of researchers in the field of data mining, machine learning and pattern recognition, etc”[3,4]. The researchers discussed that the discovery of knowledge in databases or KDD is an interdisciplinary field that focuses on methodologies for retrieving useful data from data. The term "knowledge discovery in databases or KDD for brevity refers to a broad process of finding knowledge in data and emphasizes the "high-level" application of specific methods of data mining. The unifying goal of the KDD process is to extract knowledge from the data in the context of large databases. It has already been proven that the use of data mining provides benefits in many areas of medicine, including diagnosis, prognosis and treatment. Data mining is an interdisciplinary field combining ideas from statistics, machine learning, informatics, visualization and other disciplines”. This symbolizes that this is a very useful approach for the integration of information and theory for the discovery of knowledge from any informatics such as Bioinformatics, Chemo informatics, Nano informatics and informatics of materials. Data collection consists of a set of methods that can be used to extract relevant and interesting knowledge from data. Data mining has several tasks, such as association rules, classification, forecasts and clustering, etc. Classification methods are controlled learning methods that are used to classify a data item in a predefined class label. This is one of the most useful methods for building mining models by relaying on data sets. The use of classification methods is usually used to construct models that are used to predict future data trends. “Data mining appeared in the mid-1990s and appeared as a powerful tool that is suitable for obtaining a previously unknown template and useful information from a huge data set. In various studies, it was stressed that data mining techniques help the data holder to analyze and detect suspicious relationships between their data, which in turn is useful for decision-making”[5,6]. The researchers put forward the idea that data mining is a method that extracts hidden prognostic information from a large database. It uses sophisticated algorithms for the process of sorting large amounts of data sets and collecting relevant information[7].

iJETRM
International Journal of Engineering Technology Research & Management

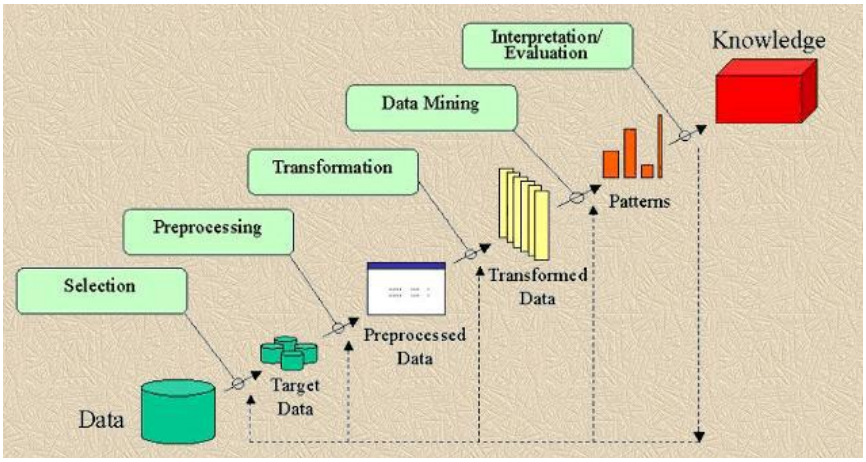

Fig: The process of KDD

As a rule, Data Mining and Discovery in Databases (KDD) are related terms and are used interchangeably, but many researchers assume that both terms differ from each other, because Data Mining is one of the most important stages of the KDD process. "Data mining methods have been used as a hybrid model for the prediction of disease from flares, which means that it is a way of combining two methods of data mining to fill the weakness caused by one technology"[8,9].

The stages in data mining:

Research:

In the first stage, the data intelligence is cleared and converted into another form, and important variables are determined, and then the nature of the data based on the task. Pattern identification: After the data has been studied, refined and defined for specific variables, the second step is the formation of the sample identification. Identify and select the templates that will make the best forecast.

Deployment: templates are deployed to achieve the desired results.

Algorithms and methods of data mining and methods, such as classification, clustering, regression, artificial Intellectual, neural networks, association rules, decision trees, genetic algorithm, Nearest Neighbor method, etc. used to detect knowledge from databases.

Classification:

Classification is the most commonly used data mining technology that uses a set of pre-classified examples to develop a model that can classify a collection of records as a whole. Fraud detection and credit risk applications are particularly suitable for this type of analysis. "In this approach, algorithms for classifying solutions or classification algorithms based on a neural network are often used. The data classification process includes training and classification. In the learning process, learning data is analyzed using a classification algorithm. Classification tests use data to assess the accuracy of classification rules"[9,10]. If accuracy is allowed, the rules can be applied to new data tuples. For the fraud detection application, this will include full records of both fraudulent and actual actions, determined based on the record for each report. The classifier's learning algorithm uses these pre-classified examples to define a set of parameters necessary for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by induction of the decision tree
- Bayesian classification
- Neural networks
- Support for vector machines (SVM)
- Classification by association

Clustering:

Clustering can be called the identification of similar object classes. "Using clustering methods, we can additionally identify dense and sparse areas in the object space and discover the overall structure of the distribution and correlation between the data attributes"[11,12]. The classification approach can also be used for effective means of distinguishing groups or classes of an object, but it becomes expensive, so clustering can be

used as a preprocessing approach to select and classify a subset of attributes. For example, to form a group of customers based on shopping patterns, to categories of genes with similar functionality.

Types of clustering methods

• Methods of separation
• Hierarchical agglomerative methods
• Methods based on density
• Grid-based methods
• Model-based methods

Prediction

The regression method can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables[13]. In independent data of intellectual variables, the attributes are already known, and the response variables are what we want to predict. Unfortunately, many real-world problems are not just a prediction. For example, sales volumes, stock prices and product failure rates are very difficult to predict, because they can depend on the complex interactions of several predictor variables. Therefore, more complex methods (for example, logistic regression, decision trees or neural networks) may be needed to predict future values. The same models can often for regression and classification. For example, CART (classification and regression Trees), the decision tree algorithm can be used to construct both classification trees (to classify the categorical response variables) and regression trees (for predicting continuous response variables). Neural networks can also create models of classification and regression.

Types of regression methods

•Linear regression
•Multidimensional linear regression
•Nonlinear regression
•Multidimensional nonlinear regression

Association Rule:

Association and correlation, as a rule, reveal frequent results of a set of elements among large data sets. "This type of search helps companies make certain decisions, such as catalog design, cross-marketing and consumer behavior analysis. Algorithms of association rules should be able to generate rules with confidence values less than one. However, the number of possible association rules for a given data set is usually very large, and most of the rules are usually insignificant (if any).

Types of association rules

• The rule of a multilevel association
• The multidimensional association rule
• Quantitative association rule

Neural networks

The neural network is a set of connected I / O modules, and each connection has the weight present with it. "At the learning phase, the network learns by adjusting the weights to be able to predict the correct label of the input tuple classes. Neural networks have a remarkable ability to extract meaning from complex or inaccurate data and can be used to extract patterns and identify trends that are too complex for people or other computer methods to notice"[14]. They are well suited for continuous values of inputs and outputs. For example, a manuscript reorganization of a character, for teaching a computer a work of English text and many business tasks in the real world, has already been successfully applied in many industries". Neural networks are best suited to identify patterns or trends in data and are well suited for forecasting or forecasting.

Data Mining Applications:

Data mining is a relatively new technology that has not fully matured. Despite this, there are a number of industries that already use it on a regular basis. "Some of these organizations include retail stores, hospitals, banks and insurance companies. Many of these organizations combine data mining with things such as statistics, pattern recognition and other important tools. Information analysis can be used to find patterns and connections that would otherwise be difficult to find"[15]. This technology is popular with many companies, as it allows them to learn more about their customers and make reasonable marketing decisions. Below is an overview of business tasks and solutions based on data mining technology.

iJETRM
International Journal of Engineering Technology Research & Management

CONCLUSION

Data mining is important for finding patterns, predicting, detecting knowledge and so on, in different business domains. Methods and algorithms for data mining, such as classification, clustering, etc., help to find patterns for determining future business trends. Data mining has a wide field of application in almost every industry where data are generated; therefore, intellectual data analysis is considered one of the most important boundaries in the database and information systems and one of the most promising interdisciplinary developments in the field of information technology.

REFERENCES

[1] Aggarwal Charu C. and Yu Philip S., Privacy Preserving data mining, 2008.
[2] Clifton C., Kantarcioglu M. , Vaidya J., Defining Privacy for Data Mining. Purdue University, West Lafayette.
[3] Elmasri, Navathe, Somayajulu and Gupta Fundamentals of Database Systems, First Impression -2006.
[4] Evfimievski A. Randomization in Privacy-Preserving Data Mining. In SIGKDD Explorations, 4(2): 43-48, December 2002.
[5] Hann J., Kamber M., Data Mining concepts and techniques 2ed.
[6] Jagannathan G., Pillaipakkamnatt K., Wright R.N., A New Privacy-Preserving Distributed k-Clustering Algorithm in proceedings of 2006 SIAM international conference on data mining on SDM-2006.
[7] Lindell Y. and Pinkas B., Privacy Preserving Data Mining, Advances in Cryptology – Crypto '00 Proceedings, LNCS 1880, Springer-Verlag, pp. 20-24, August 2000. A full version appeared in the Journal of Cryptology, Volume 15 - Number 3, 2002.
[8] http://mathworld.wolfram.com/InverseTrigonometricFunctions.html
[9] Oliveira S. R. M. and Zaïane O. R., Privacy Preserving Clustering By Data Transformation. In Proceedings of the 18th Brazilian Symposium on Databases, Manaus, Amazonas, Brazil, October 2003, pp.304-318
[10] Oliveira S. R. M. and Zaïane O. R., Achieving privacy preserving when sharing data for clustering, In Proceedings of the International Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB 2004, Toronto, Canada, August, 2004.
[11] Pinkas B., Cryptographic Techniques for Privacy-Preserving Data Mining SIGKDD Explorations, the newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, January 2003,
[12] Sweeny L., Achieving k-anonymity privacy protection using generalization and suppression. 2002 CMU.
[13] Upadhyay A.K., Gupta R., Kumar R., Analytical model for revised K-clustering algorithm for privacy preservation in data mining. RACE 2007 at BEC Bikaner, IEEE sponsored international conference.
[14] Vaidya J. and Clifton C. Privacy-Preserving K-Means Clustering over Vertically Partitioned Data. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 2003, pp.206-215.
[15] Wikipedia – The free encyclopedia, www.wikipedia.org