Volume 6, Issue 4, April 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Mining– A Survey

Amala Jayanthi.M*, Swathi.S, Tharakai.R
Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India

Abstract— Data mining is a emerging interdisciplinary sub domain of knowledge management of computer science. Data mining is a analyzing process of discovering hidden information and pattern from large scale of data of any kind. For past few decades data mining influences on various field as organizations such as Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition, Business, Education, Medical, Scientific etc. This paper discusses on the idea, techniques, issues, applications and tools of data mining.

Keywords— Knowledge discovery, Predicitive Mining, Descriptive Mining, Tools, Techniques.

I. INTRODUCTION

Knowledge or Information for decision making in a business is very poor even though data storage grows exponentially. Data mining also known as Knowledge Discovery of Data (KDD) is computer assisted process that analyses the enormous amount of data and extract the exact knowledge from the data. The Knowledge extracted allows to predict the behaviour and future behaviour .This allows the business owners to take positive, knowledge driven decisions. Data mining is applied on various industries like retail, finance, health care, aerospace, education etc. Knowledge is extracted from the historical data by applying pattern recognition, statistical and mathematical techniques that results in the knowledge in the form of facts, trends, associations, patterns, anamolies and exceptions**.**

II. KNOWLEDGE DISCOVERY PROCESS FROM DATA

Knowledge is retrieved from large volume of data undergoing following process,

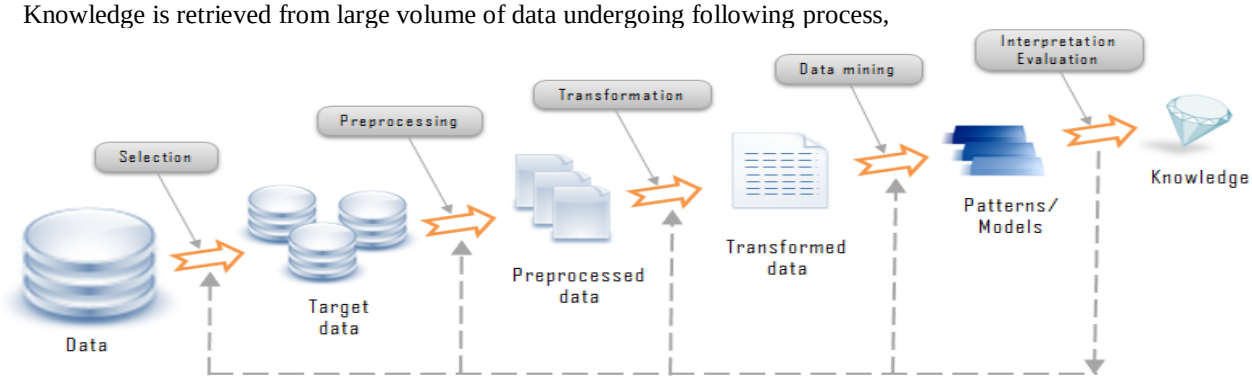

Fig 1. KDD Process

A. Data Pre-processing:

Data Pre-processing make ready the real world data for the mining process.

1) *Data Cleaning*
Data without quality will results in poor quality knowledge. Real world data is dirt with noise, error, missing, inconsistency, irrelevancy. Data cleaning treats the dirt data to have quality data for knowledge discovery.
2) *Data Integration*
Data Integration involves combining of data from multiple, heterogeneous data sources into to a single coherent data source that provides unified view on the multiple data stores(Data warehouse).
3) *Data Selection:*
Data Selection is the process of selecting relevant data that favours the business problem.
4) *Data Transformation :*
Data Transformation transforms the interested data that is selected in previous step to the relevant form for mining.

B. Data Mining:

Data Mining is application of intelligent methods that extract data patterns.

C. Pattern Evaluation:

The patterns that are generated by the data mining are evaluated for the interest according to the target business problem.

D. Knowledge Presentation:

Knowledge Presentation uses visualization techniques that visualize the interesting patterns and helps the user to understand and interpret the resultant patterns.

III. DATA MINING FUNCTIONALITIES

The KDD process is ultimately data mining methods to extract patterns from data. Each methods have different aim, which decides the outcome of the KDD process entirely. The outcome of the KDD process can be any of the following tasks based on the customer desire.

These tasks are categorized as predictive and descriptive mining,

A. Predictive Mining:

Supervised learning task where the unknown value of a class or future values of interest is predicted from the existing data. It can also validate a newly invented hypothesis.

1) *Classification:*
   Classification results in classification model termed as classifiers that classifies the data as classes and concepts .The resultant model is used to predict the class label of the instances for which the class label is unknown. Decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques are the some of the kinds of classification methods used to decide the classifier of the sample.
2) *Regression :*
   Regression is a data mining function. It is used to predict the missing or unavailable numerical data values by mapping the data into a function. Linear regression and Multi- Linear regression are some of the regression techniques. Regression model is usually a graph which is used to analyze the future trends based on the past and present data.
3) *Prediction:*
   Prediction models a predictor that predicts the unknown data and future data from the available data.
4) *Decision Trees* :
   Decision tree is a model that is tree like structure with test nodes and branches . Test nodes are called as predictor and the leaf nodes are the classes or concepts.. The predictors helps to classifies the data as classes or concepts.

B. Descriptive mining:

It is a task of summarizing the data and its features as patterns using data mining and data aggregation methods.

1) *Clustering :*
   Clustering is a task of grouping data of similar characteristics into a cluster while the different data may grouped into different respective clusters.Search for the cluster is a unsupervised learning i.e Class label is unknown. Thus the data are organized into an effective representation that categories the sample data .
2) *Association rule mining:*
   Association rule mining unwraps the patterns that occurs frequently among the data set. It focus in extracting associations, correlations, frequent sequence, frequent itemset and frequent patterns with interestingness among the data set in the data repositories.
3) *Summarization :*
   Summarization is the process of reducing the huge volume of data in a meaningful and intelligent fashion with important and relevant features. Summarization techniques like tabulation of the mean and the standard deviations are often implied to analyse and visualize the data, and to generate the report automatic.

IV. TOOLS OF DATA MINING

Data mining tools are categorized as stand-alone and client/server solutions. Client/server solutions are dominating .They are specially designed forbusiness users. They are available for different platforms, including Windows, MAC OS, Linux, or spe-cial mainframe supercomputers. Many number of JAVA-based systems are being developed that are platform-independent for researchers and applied researchers.

A. WEKA:

The original version of WEKA was non-JAVA and was developed to analyze data from the agricultural domain. the JAVA version of WEKA, l is very sophisticated and used in various applications to visualize , analyze and predict. Its a open ware under the GNU General Public License, Users can customize the tool.

B. Rapid Miner :

Rapid Miner is Java based tool that offers advanced analytics through template-based frameworks. This Tool has been offered as a service, rather than a local software. RapidMiner also provides functionalities like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment..

C. R – Programming:

R – Programming is developed from C and Fortran. It's a freeware that provide software programming language and software environment for statistical computing and graphics. Data miners to develop statistical software and data analysis

with the help of R-Programming. It is very ease to use.It also provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification and clustering apart from data mining.

D. Orange:

Orange, a Python-based, powerful and openware. It has components for machine learning, bioinformatics and text mining. It's wrapped with characteristics for data analytics.

E. KNIME:

KNIME is a Java based. KNIME does all the three process of extraction, transformation and loading of data .It provides a GUI that allows to assemble the nodes for data processing It is an open source that is able to do data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept. It is also able to perform business intelligence and financial data analysis. KNIME is easy to extend and to add plugins.

F. NLTK:

NLTK is phython based can be customized. NLTK provides a pool of language processing tools including data mining, machine learning, data scraping, sentiment analysis and other various language processing tasks.

V. APPLICATIONS OF DATA MINING

Data Mining is used in many domain in constant basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many domains like health care, finance insurance, retail stores combines the data mining with as statistics, pattern recognition, and other important tools to perform data analytics. Data mining is used primarily for decision making.

A. Medicare and health care:

Using data mining methods, it is able to find the correlation between the diseases, to analyse the effectiveness of the treatments given, to identify the new drugs to analyse the market of the drugs and etc.

B. Education:

Educational Data Mining is a blooming field which knowledge from educational Environment data . The goals of EDM are identified as predicting students' learning behaviour, emotions, skills. This study improves the educating methods by understanding the ward and to take accurate decisions respectively

C. Market Basket Analysis:

Market basket analysis is a technique that uses association rule mining to understand the purchasing behaviour of the customer. It also allows the seller to understand his business, customer's needs and to make profitable change accordingly.

D. Financial Banking :

Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

E. Research Analysis:

Data mining is very useful in data pre-processing and integration of databases. Data mining allows the researcher to identify co-occurring sequences and the correlation between any activities .Data visualisation and visual data mining help the researcher with a clear view of the data.

F. Fraud Detection:

Huge amount of dollars is being lost because of fraud detection. Traditional methods are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information.Vaild and useful information is called as knowledge. The results are categorized into fraudulent or non-fraudulent.

In addition to these data mining plays a very vital role in Bio- Informatics, Criminal Investigation,Corporate Surveillance, Insurance,Agriculture,Customer Segementation,Lie Detection Intrusion Detection, Manufacture engineering and etc.

VI. CONCLUSION

In this paper the process of KDD and relevance of data mining in various sectors is discussed.The data mining functionalities –predictive mining (classification, regression, prediction and decision trees) and descriptive minining (clustering, association, summarization) is also summarized. The essential of data mining in commercial, educational, medical, scientific fields are highlighted.

REFERENCES

[1] http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/

[2] Thair Nu Phyu ,*Survey of Classification Techniques in Data Mining* ,Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I,IMECS 2009, March 18 - 20, 2009, Hong Kong,ISBN: 978-988-17012-2-0

[3] Lior Rokach,Oded Maimon, *clustering methods, data mining and knowledge discovery handbook*, Department of Industrial Engineering,Tel-Aviv University.

[4] Sarat M. Kocherlakota, Christopher G. Healey,*Summarization Techniques for Visualization of Large Multidimensional Datasets*,Technical Report TR-2005-35Knowledge Discovery Lab,Department of Computer Science, North Carolina State University,Raleigh, NC 27695-8207

[5] T.Karthikeyan,N. Ravikumar *A Survey on Association Rule Mining*, International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, ISSN (Print) : 2319 5940,ISSN (Online) : 2278-1021, Issue 1,January 2014.

[6] http://bigdata-madesimple.com/14-useful-applications-of-data-mining/

[7] Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)3rd Edition .

