International Journal of Research

Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 02Issue 08
August 2015

DATA MINING: TECHNIQUES, TOOLS AND APPLICATIONS

Poonam
Dept. of Computer Science
IB (PG) College,
Panipat, (HR.)
ahlawatp12@gmail.com

ABSTRACT: Data mining is a method which finds useful patterns from large amount of data. Present study discusses few of the data mining techniques, its tools and applications. Data mining is the process of extracting the useful data, patterns and trends from a large amount of data by using techniques like clustering, classification, association and regression. There are a wide variety of applications in real life. Various tools are available which supports different algorithms. In this paper a summary about data mining tools available and the supporting algorithms are included. Comparison between various tools has also been done to enable the users use various tools according to their requirements and applications. Some of the applications are also given in the study.

KEYWORDS: Data mining Techniques; Data mining algorithms; Data mining tools and applications.

INTRODUCTION: 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.

Three steps involved in data mining are

- Exploration- In the first step data is cleaned and transformed into another form, and

important variables and then nature of data based on the problem are determined.

- Pattern Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.
- Deployment: Patterns are deployed for desired outcome

In the factual world, huge amount of data are available in education, medical, industry and many other areas. Such data may make available knowledge and information for decision making. For example, you can find

out drop out student in any university, sales data in shopping database. Data can be analyzed, summarized, understand and meet to challenges. Data mining is a influential concept for data analysis and process of discovery interesting pattern from the huge amount of data, data stored in various databases such as data warehouse, world wide web, external sources .Interesting pattern that is easy to understand, unknown, valid, potential and useful. Data mining is a type of sorting technique which is actually used to extract hidden patterns from large databases. The goals of data mining are fast retrieval of data or information, knowledge Discovery from the databases, to identify hidden patterns and those patterns which are previously not explored, to reduce the level of complexity, time saving, etc. Data mining refers extracting knowledge and mining from large amount of data. Sometimes data mining treated as knowledge discovery in database (KDD).

The enlargement 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 move toward 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.

Data mining is the process of extracting useful information. On the whole it is the process of discovering hidden patterns and information from the existing data. In data mining, one needs to primarily muse on cleansing the data so as to make it practicable for further dispensation. The process of cleansing the data is also called as noise elimination or noise reduction or feature elimination. This can be done by using various tools available supporting various techniques. The important consideration in data mining is whether the data to be handled static or dynamic. In general, static data is easy to handle as it is known earlier and stored. Dynamic data refers to high voluminous and continuously changing information which is not stored earlier for analyzing and processing like static data. It is difficult to maintain dynamic data as it changes with time. Many

algorithms are used to analyze the data of curiosity. Data can be sequential, audio signal, video signal, spatio -temporal, temporal, time series etc.

DATA MINING TECHNIQUES

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns etc., are used for knowledge discovery from databases.

1. **Association**

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

Applications: market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc.

Types of association rules: Different types of association rules based on

- Types of values handled
- Boolean association rules
- Quantitative association rules
- Levels of abstraction involved
- Single-level association rules
- Multilevel association rules
- Dimensions of data involved
- Single-dimensional association rules
- Multidimensional association rules

2. Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. For Example, Teachers classify students' grades as A, B, C, D, or F. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the

company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group.

Classification Techniques

- Regression
- Distance
- Decision Trees
- Rules
- Neural Networks

3. Clustering

Clustering is "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. We can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without irritate. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

4. Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. In data mining independent variables are attributes already known and response variables are what we want to predict unfortunately, many real-world problems are not simply prediction For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., decision trees) may be necessary to forecast future values. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

5. Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

TOOLS FOR DATA MINING TECHNIQUES

There are various open source tools available for data mining. Some of tools work for clustering, some for classification, regression, association and some for all. There are various algorithms for each technique. There are some different tools which can be used to implement some algorithm.

- **Tool 1-Orange**

Orange is the Open source data visualization and analysis tool. Data mining is done through visual programming or Python scripting. Regression method is also being used in Orange where ensembles are basically wrappers around learners.

- **Tool 2- WEKA**

WEKA stands for Waikato Environment for Knowledge Analysis. It is developed in Java programming language. It contains tools for data preprocessing, classification, clustering, association rules and visualization. It is not capable for multi relational data mining. Data file can be used in any format like ARFF (attribute relation file format), CSV (comma separated values), C4.5 and binary and can be read form a URL or from SQL database as well by using JDBC. One additional feature is that data sources, classifiers etc are called as beans and these can be connected graphically.

- **Tool 3-SCaVis**

Scientific **C**omputation and **Vis**ualization Environment. It provides environment for scientific computation, data analysis and data visualization designed for scientists, engineers and students. The program incorporates many open source software packages into a coherent interface using the concept of dynamic scripting. It provides freedom to choose a programming language, freedom to choose an operating system and freedom to share code. There is provision of multiple clipboards, multi-document support and multiple Eclipse-like bookmarks Extensive LaTeX support: a structure

viewer, a build-in Bibtex manager, LaTeX equation editor and LatexTools

- **Tool 4- Apache Mahout**

Its goal is to build machine learning library scalable to large data set. For Classification following algorithms are included: Logistic Regression, Naive Bayes/ Complementary Naive Bayes, Random Forest, Hidden Markov Models, Multilayer Perceptron. For Clustering following algorithms are included: Canopy Clustering, k-Means Clustering, Fuzzy k-Means, Streaming k-Means, Spectral Clustering by Sean Owen and Sebastian Schelter.

- **Tool 5- R Software Environment**

R provides free software environment for statistical computing and graphics mostly for UNIX platforms, Windows and MacOS. It is an integrated suite of software facilities like data manipulation, calculation and graphical display. It provides a wide variety of graphical techniques as well as statistical like linear and nonlinear modeling, classical statistical tests, classification, clustering.

- **Tool 6- ML Flex**

ML uses machine learning algorithms to derive models from independent variables with the purpose of predicting the values of a dependent (class) variable.

- **Tool 7- Data bionic ESOM (Emergent Self Organizing Maps) tool**

On can do Preprocessing, Training, Visualization, Data analysis, Clustering, Projection, and Classification using this tool. Training data is set of points from a high dimensional space called data space. The two most common training algorithms are online and batch training. Both of these training algorithms will search the closest prototype for each data point that is best match. Online training, there is immediately update of best match but in batch training all the best matches are being collected and then update if performed collectively.

- **Tool 8-NLTK (Natural Language Tool Kit)**

NLTK is a leading platform for building Python programs to work with human language data.

It provides easy-to-use interfaces to over 50 corpora. It also provides lexical resources such as Word Net, along with a suite of text processing libraries for classification, tokenization, stemming, and tagging,

parsing, and semantic reasoning. NLTK is available for Windows, Mac-OS X, and Linux. NLTK is a free, open source, community-driven project. It defines various classifier classes: Conditional Exponential Classifier, Decision Tree Classifier, Maxent Classifier, Naive Bayes Classifier, Weka Classifier.

- **Tool 9-ELKI (Environment for Developing KDD- Applications Supported by Index- Structures)**

ELKI is open source data mining software written in Java. The focus of ELKI is research in algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection. ELKI offers many data index structures such as the R*-tree that can provide major performance gain and in order to achieve high performance and scalability. The approach used is the independence of file parsers or database connections, data types, distances, distance functions, and data mining algorithms.

- **Tool 10-UIMA (Unstructured Information Management Architecture) diagram**

Large amount of unstructured information can be analyzed to get relevant information. It enables application to be decomposed into components. Working of framework is to manage these components and flow between them. Basic availability is frameworks, components and infrastructure.

- **Tool 11-GraphLab**

Graph Lab has several algorithms already implemented in its toolkit. One can also implement one's own algorithm on top of our graph programming API .

- **Tool 12-mlpy machine learning Python**

It has algorithms of regression and classification. Cluster analysis can also be done for dimensionally reduction and wavelet transform. Various different algorithms like feature ranking, re sampling algorithm, peak finding algorithm, error evaluation are also available.

- **Tool 13-KEEL (Knowledge Extraction Evolutionary Learning)**

KEEL is open source. It uses java software which have license of GPLv3 (General Public License version 3). It allows users to have the access of behavior of evolutionary learning and basic soft computing based

techniques for various kinds of data mining problems to be handled.

- **Tool 14-Scikit-learn**

Scikit-learn is also a free package. It is in Python which extends the functionality of NumPy and SciPy packages. It also uses the matplotlib package for plotting charts. The package supports most of the core DM algorithms except including classification rules and association rules.

Data Mining Application

Various field adapted data mining technologies because of fast access of data and valuable information from a large amount of data. Data mining application area includes marketing, telecommunication, fraud detection, finance, and education sector, medical and so on. Some of the main applications listed below:

Data Mining in Education Sector: We are applying data mining in education sector then new emerging field called "Education Data Mining". Using these term enhances the performance of student, drop out student, student behavior, which subject selected in the course. Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Use student's data to analyze their learning behavior to predict the results.

Data Mining in Banking and Finance: Data mining has been used extensively in the banking and financial markets. In the banking field, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability. In the financial markets, data mining technique such as neural networks used in stock forecasting, price prediction and so on.

Data Mining in Market Basket Analysis: These methodologies based on shopping database. The ultimate goal of market basket analysis is finding the products that customers frequently purchase together. The stores can use this information by putting these products in close proximity of each other and making them more visible and accessible for customers at the time of shopping.

Data Mining in Earthquake Prediction: Predict the earthquake from the satellite maps. Earthquake is the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along a geologic fault in the interior. There are two

basic categories of earthquake predictions: forecasts (months to years in advance) and short-term predictions (hours or days in advance) .

Data Mining in Bioinformatics: Bioinformatics generated a large amount of biological data. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data.

Data Mining in Telecommunication: The telecommunications field implement data mining technology because of telecommunication industry have the large amounts of data and have a very large customer, and rapidly changing and highly competitive environment. Telecommunication companies uses data mining technique to improve their marketing efforts, detection of fraud, and better management of telecommunication networks.

Data Mining in Agriculture: Data mining than emerging in agriculture field for crop yield analysis a with respect to four parameters namely year, rainfall, production and area of sowing. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The yield prediction problem can be solved by employing Data Mining techniques such as K Means, K nearest neighbor (KNN), Artificial Neural Network and support vector machine (SVM) .

Data Mining in Cloud Computing: Data Mining techniques are used in cloud computing. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage .Cloud computing uses the Internet services that rely on clouds of servers to handle tasks. The data mining technique in Cloud Computing to perform efficient, reliable and secure services for their users.

CONCLUSION-Data mining is a “decision support” process in which we search for patterns of information in data. In other words, Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc in different business domains. Data mining techniques such as classification, clustering, prediction, association and sequential patterns etc it

helps in finding the patterns to decide upon the future trends in businesses to grow. It has wide application field 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 also.

REFERENCES:

- Phridvi Raj MSB., Guru Rao CV (2013) Data mining – Past, Present and Future – A Typical Survey on data streams. INTER-ENG Procedia Technology 12:255 – 263
- Mansi Gera Shivani Goel (2015) Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 18, March 2015 22
- Bharati M. Ramageri Data Mining Techniques and Applications, Indian Journal of Computer Science and Engineering Vol. 1 No. 4 pp 301-305 ISSN: 0976-5166
- Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.
- Dr. Gary Parker, (2004), Data Mining: Modules in emerging fields, Vol. 7CD-ROM.
- Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP-0800.pdf.
- Smita, Priti Sharma (2014) Use of Data Mining in Various Field: A Survey Paper IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 18-21 www.iosrjournals.org
- Yongjian Fu "data mining: task, techniques and application"
- Er. Rimmy Chuchra (2012) "Use of Data Mining Techniques for the Evaluation of Student Performance:A Case Study" International Journal of Computer Science and Management Research Vol 1 Issue 3 October 2012

- J. Han and M. Kamber. "Data Mining, Concepts and Techniques", Morgan Kaufmann, 2000.
- Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar(2012) Data Mining Techniques & Distinct Applications: A Literature Review" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November- 2012 Industry Application of data mining , http://www.pearsonhighered.com/samplechapter/0130862711.pdf
- David L Olson, Dursun Delen (2008)" Advance data minig techniques" springer
- G. V. Otari, Dr. R. V. Kulkarni, "A Review of Application of Data Mining in Earthquake Prediction" G. V. Otari et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012,3570-3574
- D Ramesh, B Vishnu Vardhan, (2013) "Data Mining Techniques and Applications to Agricultural Yield Data" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013
- Ruxandra-Ştefania Petre(2012) "Data mining in Cloud Computing" Database Systems Journal vol. III, no. 3/2012