International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 3 Issue 13, December 2016

ISSN: 2394-4404

Classification Of Data Mining Techniques & Tools: A Survey

Fatima

Research Scholar, Magadh University Bodh Gaya, Gaya, Bihar, India

Dr. Javed Ikbal Khan

Associate Professor, Deptt. of Mathematics, M.G. College, Gaya, Bihar, India

Abstract: ***Data mining is a procedure which finds functional patterns from large amount of data. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It uses machine learning, statistical and visualization technique to discover and present knowledge in a form which is easily comprehensible to humans. This review of literature focuses on data mining techniques, issues, tools, and applications. In this paper we have focused a variety of techniques, approaches and different areas of the research which are helpful and patent as the important field of data mining Technologies.***

Keywords: Data mining, data mining task, data mining techniques, data mining tools.

I. INTRODUCTION

The field of data mining and knowledge discovery is emerging as a new, fundamental research area with important applications to science, engineering, medicine, business, and education. Data mining attempts to formulate analyze and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from unstructured data. [1]Text mining is a method which is used in different fields like machine learning, information retrieval, statistics and computational linguistics. Web mining is a sub discipline of text mining used to mine the semi structured web data in form of Web content mining, Web structure mining and Web usage mining. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process [2][3]. Many organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions [4]. The primary disadvantage to data mining is that users may discover things which are based on chances instead of a direct connection. In order for data mining to be used effectively, the users must be able to tell the difference between chance and a direct correlation. [3] Some disadvantages of DM are privacy and security issues. Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people. Data mining is useless if you don't have any data to analyze. While most organizations already collect data to some extent, this is not enough if you want to use data mining successfully. The information must be specific and refined [6].

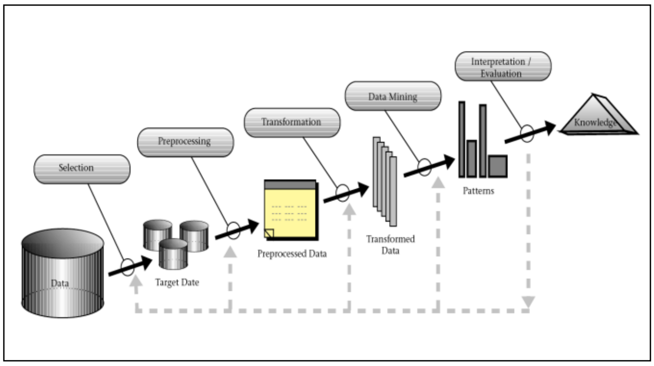


Figure 1: Process of KDD

Sometime, data may be in different formats as it comes from different sources, irrelevant attributes and missing data. Therefore, data needs to be prepared before applying any kind of data mining. Data mining is also known under many other names, including knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing.[7]Many researchers and practitioners use data mining as a synonym for knowledge discovery but data mining is also just one step of the knowledge discovery process. All the techniques follow an automated process of knowledge discovery (KDD) i.e., data cleaning, data integration, data selection, data transformation, data mining and knowledge representation. [8]

II. DATA MINING TASK

Data mining satisfy its main goal by identifying valid, potentially useful, and easily understandable correlations and patterns present in existing data. This goal of data mining can be satisfied by modelling it as either Predictive or Descriptive nature. The Predictive model works by making a prediction about values of data, which uses known results found from different datasets.

A. CLASSIFICATION

Classification involves the discovery of a predictive learning function that classifies a data item into one of several predefines classes. It involves examining the features of a newly presented object and assigning to it a predefined class. Define classification has a two-step process. First a model is built describing a predetermined set of data classes or concepts and secondly, the model is used for classification.

B. REGRESSION

Statistical Regression is another Predictive data-mining model also known as is a supervised learning technique. This technique analyzes of the dependency of some attribute values, which is dependent upon the values of other attributes mainly present in same item. The development of a model can predict these attribute values for new cases. The difference between regression and classification is that regression deals with numerical/continuous target attributes, whereas classification deals with discrete/categorical target attributes.

C. CLUSTERING

Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Clustering is commonly used to search for unique groupings within a data set. The distinguishing factor between clustering and classification is that in clustering there are no predefined classes.

D. SUMMARIZATION

Summarization is called as the abstraction or generalization of data. The summarization technique maps data into subsets with simple descriptions. The summarized data set gives general overview of the data with aggregated information. Summarization can scale up to different levels of abstraction and can be viewed from different angles. It is a key data-mining concept involving techniques for finding a compact description of dataset.

E. ASSOCIATION

Associations or Link Analysis technique are used to discover relationships between attributes and items. In these techniques, the presence of one pattern implies the presence of another pattern i.e. item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of a model.[11] This association rules are also build by programmers use to build programs capable of machine learning.

III. BASIC CLASSIFICATION TECHNIQUES

A. DECISION TREE CLASSIFIER

Decision tree is a flow-chart-like tree structure Leaf nodes represent class labels or class distribution. Decision tree is a classifier in which each non-terminal node represents either a test or decision for the given data item. Which branch to be select next is depends upon the outcome of the test. To classify a given data item, need to from start at the root node and follow the assertions down until we reach a terminal node or leaf node. Decision trees use recursive data partitioning. The important things in decision tree are attribute selection measure. There is important parameter used for attribute selection. The attribute with highest information gain is used to be selected as a root.

B. NAIVE BAYESIAN CLASSIFIERS

The Naive Bayesian classifier, or simple Bayesian classifier are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. The Naive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Naive Bayes Rule is the basis for many machine-learning and data mining methods.

C. NEURAL NETWORK AS A CLASSIFIER

Neural network approach has been widely adopted as classifiers. The neural network provides several advantages, like arbitrary decision its nonparametric nature, boundary capability, easy adaptation to different types of data. Neural nets consist of three layers such as input layer, hidden layer and output layer. There are numerous advantages of ANN some of these include

- ✓ High Accuracy.
- ✓ Independent from prior assumptions about the distribution of the data.
- ✓ Noise tolerance.
- ✓ ANN can be implemented in parallel hardware.

D. DECISION TREE CLASSIFICATION

Decision tree classification approach is most useful in classification problems. [3] It is a flow chart like tree structure. Trees are constructed in a top down recursive divide and conquer manner. In this classification method used in different type algorithm to classify the data sets, the algorithms are: [1]

- ✓ ID3(Iterative Dichotomiser)
- ✓ C4.5(a Successor of ID3)
- ✓ Classification and Regression Trees(CART)

The algorithm follows a top-down approach, which starts with a training set of tuples and their associated class labels.

E. SUPPORT VECTOR MACHINES (SVMS)

Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. Support Vector Machines are based on the concept of decision planes that define decision boundaries.

IV. REVIEW LITERATURE

Dynamic networks have recently being recognized as a powerful abstraction to model and represent the temporal changes and dynamic aspects of the data underlying many complex systems. Significant insights regarding the stable relational patterns among the entities can be gained by analysing temporal evolution of the complex entity relations.[9]Web crawlers are essential to many Web applications, such as Web search engines, Web archives, and Web directories, which maintain Web pages in their local repositories. We propose a set of crawling algorithms for effective and efficient crawl ordering by prioritizing important pages with the well-known Page Rank as the importance metric reflected in the gradually descending curves in the performance of semantic focused crawlers. [10] a nature-inspired theory to model collective behavior from the observed data on blogs using swarm intelligence, where the goal is to accurately model and predict the future behavior of a large population after observing their interactions during a training phase. Specifically, an ant colony optimization model is trained with behavioral trend from the blog data and is tested over real-world blogs. Promising results were obtained in trend prediction using ant colony based pheromone classier and CHI statistical measure. [11] Over the past decade, there has been an explosion of interest in network research across the physical and social sciences. For social scientists, the theory of networks has been a gold mine, yielding explanations for social phenomena in a wide variety of disciplines from psychology to economics. Here, we review the kinds of things that social scientists have tried to explain using social network analysis and provide a nutshell description of the basic assumptions, goals, and explanatory mechanisms prevalent in the field. [12] A collective approach to learning a Bayesian network from distributed heterogeneous data. Bayesian network is learnt at the central site using the data transmitted from the local site. The local and central Bayesian networks are combined to obtain a collective Bayesian network, which models the entire data.

V. DATA MINING TOOLS

A. RAPID MINER (FORMERLY KNOWN AS YALE)

Written in the Java Programming language, template-based frameworks are done by using rapid miner. Users hardly have to write any code. Offered as a service, rather than a piece of local software, this tool holds top position on the list of data mining tools. Rapid Miner provides functionality data pre-processing and visualization, predictive analytics and statistical modelling, in addition of data mining. WEKA and R scripts makes even more powerful learning schemes, models and algorithms. Rapid Miner is distributed under the AGPL open source license and can be downloaded from the number one business analytics software Source Forge.

B. ORANGE

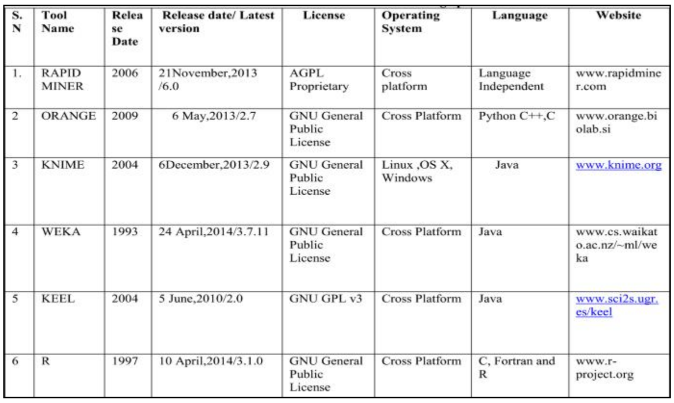
Python is getting in popularity because it’s simple and easy to learn. When it comes to looking for a tool for your work and if you are a Python developer, look there is no further option than Orange, a Python-based, powerful and open source tool for both learner and experts in particular domain. You will get addict to this tool’s while visual programming and Python scripting. It also has components for machine learning, add-ons for bioinformatics and text mining. It’s packed with copy of features for data analytics.

C. NLTK

Comes to language processing tasks, nothing can beat NLTK. NLTK provides a pool of language processing tools. It includes data mining, machine learning, data scraping, sentiment analysis and other various language processing tasks.

D. WEKA

WEKA was developed for analyzing data from the agricultural domain. WEKA is an original non java version. With the Java-based version, the tool is used in many different applications. This includes visualization and algorithms for data analysis and predictive modelling. Its free under the GNU General Public License, It is a big plus compared to Rapid Miner, because users can customize it according to their requirement. WEKA supports several standard data mining tasks those are: data pre-processing, clustering, classification, regression, and feature selection.



S. N	Tool Name	Release Date	Release date/ Latest version	License	Operating System	Language	Website
1.	RAPID MINER	2006	21November,2013 /6.0	AGPL Proprietary	Cross platform	Language Independent	www.rapidminer.com
2	ORANGE	2009	6 May,2013/2.7	GNU General Public License	Cross Platform	Python C++,C	www.orange.biolab.si
3	KNIME	2004	6December,2013/2.9	GNU General Public License	Linux ,OS X, Windows	Java	www.knime.org
4	WEKA	1993	24 April,2014/3.7.11	GNU General Public License	Cross Platform	Java	www.cs.waikato.ac.nz/~ml/weka
5	KEEL	2004	5 June,2010/2.0	GNU GPL v3	Cross Platform	Java	www.sci2s.ugr.es/keel
6	R	1997	10 April,2014/3.1.0	GNU General Public License	Cross Platform	C, Fortran and R	www.r-project.org

Table 1: List of DM Tools

E. R-PROGRAMMING

Project R, a GNU project, is written in R programming. In no particular order it is written in C and FORTRAN. For statistical computing and graphics R programming acts as a free software programming language and software environment. The R language is widely used among data developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity and substantially in recent years. Besides data mining provides statistical and graphical techniques, including sequence and non sequence modeling, attic statistical tasks, time-series analysis, classification, clustering, and others.

IV. CONCLUSION

This learning gives an overall thought about the data mining techniques which can be used on various server log files to find the most frequent patterns. The data mining techniques can be used to find the user behavior over the internet. In our future work are most robust technique used for data mining.

REFERENCES

[1] Safdar, M., & Khan, M. J. I. Opinion Mining For Customer Feedback: A Survey.

[2] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-7854, 1st Edition, 2006.

[3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996.

[4] Article: Exforsys Inc "Data Mining applications" Published on: 26th Jul 2006 Source: http://www.exforsys.com/tutorials/data-mining/datamining-applications.html

[5] Article: Exforsys Inc" What is Data Mining" Published on: 27th Jul 2006 Source: http://www.exforsys.com/tutorials/data-mining/datamining-overview.html

[6] Article: Exforsys Inc "Advantages of Data Mining" Published on: 26th Jul 2006 Source: http://www.exforsys.com/tutorials/datamining/data-mining-advantages.html

[7] Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.),. Advances in Knowledge Discovery and Data Mining, AAAI Press, Cambridge, 1996.

[8] Kittipol Wisaeng. "An Empirical Comparison of Data Mining Techniques in Medical Databases", International Journal of Computer Applications (0975 – 8887), Volume 77– No.7, September 2013.

[9] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.

[10] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[11] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[12] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012