INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

DATA MINING TOOLS: REVIEW

SNEHAL A DESHMUKH

Prof. Ram Meghe College of Engineering & Research, Badnera.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: The rapid development of information technology and adoption of its several applications has created the revolution in business and various fields significantly. Due to the increase in the data, it is important to extract knowledge/information from the large data repositories. Hence, Data mining has become an essential factor in various fields including business, education, health care, finance, scientific etc. Data mining and its applications can be viewed as one of the emerging and promising technological developments that provide efficient means to access various types of data and information available worldwide It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. Various popular data mining tools are available today. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. This paper presents an overview of the data mining tools like Weka, R, Orange etc.

Keywords: Data Mining, Knowledge discovery process, Data Mining Tools, Weka, R, Orange

Corresponding Author: MS. SNEHAL A DESHMUKH

Access Online On:

www.ijpret.com

How to Cite This Article:

Snehal A. Deshmukh, IJPRET, 2015; Volume 3 (9): 1238-1246

PAPER-QR CODE

INTRODUCTION

Today Information Technology plays a vital role in every aspects of the human life. It is very essential to gather data from different sources. This data can be stored and maintained to generate information and knowledge. This information and knowledge has to be disseminated to every stake holders for the effective decision making process. Due to the increase in the data, it is important to extract knowledge/information from the large data repositories. Hence, Data mining has become an essential factor in various fields including business, education, health care, finance, scientific etc. "Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. Data collected from various areas such as marketing, health, communication, etc., are used in data mining. Data Mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organisations focus on the most important information in their data warehouse." Questions those traditionally were too time consuming to resolve can be answered by the data mining tools in an effective manner. This helps to find the hidden patterns, predictive information that facilitates the experts with solution outside their expectations. The goal of data mining is to extract knowledge from dataset in human-understandable structures. In recent years data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and engineering.

2. KNOWLEDGE DISCOVERY PROCESS

The process of discovering useful knowledge from a huge data is called as Knowledge Discovery in Database (KDD) and which is often referred to as Data mining. While data mining and knowledge discovery in databases are normally treated as synonyms, but, in fact data mining is a part of knowledge discovery process.

Data collected from multiple sources often heterogeneous is integrated into a single data storage called as target data. Data relevant to the analysis is decided on and retrieved from the data collection. Then, it is pre-processed and transformed into an appropriate standard format. Data mining is a crucial step in which intelligent algorithm/techniques are applied to extract meaningful pattern or rules. Finally, those patterns and rules are interpreted to new or useful knowledge or information.

The KDD process comprises of few steps as shown in Fig. 1and explained as follows:

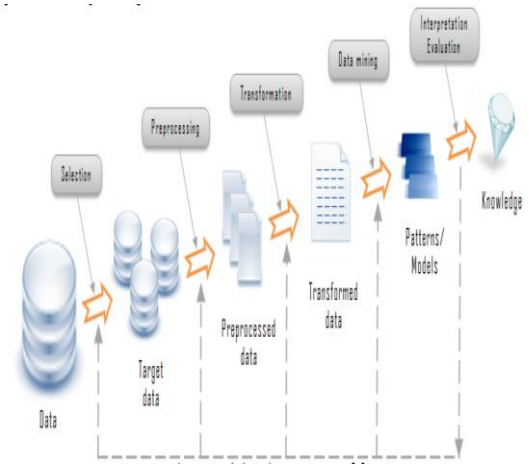

Fig. 1. Knowledge discovery project

2.1. SELECTION

This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail. From this respect, the more attributes are considered, the better.

2.2. PROCESSING

In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers. It may involve complex statistical methods or using a Data Mining algorithm in this context. For example, if one suspects that a certain attribute is of insufficient reliability or has many missing data, then this attribute could become the goal of a data mining supervised algorithm. A prediction model for this attribute will be developed, and then missing data can be predicted. The extension to which one pays attention to this level depends on many factors.

2.3. TRANFORMATION

In this stage, the generation of better data for the data mining is prepared and developed. Methods here include dimension reduction (such as feature selection and extraction and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). This step can be crucial for the success of the entire KDD project,

and it is usually very project-specific. However, even if we do not use the right transformation at the beginning, we may obtain a surprising effect that hints to us about the transformation needed (in the next iteration). Thus the KDD process reflects upon itself and leads to an understanding of the transformation needed.

2.4. DATA MINING

We are now ready to decide on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps. This stage includes selecting the specific method to be used for searching patterns. For example, in considering precision versus understandability. This approach attempts to understand the conditions under which a Data Mining algorithm is most appropriate. Each algorithm has parameters and tactics of learning.

2.5. INTERPRETATION EVALUATION

This stage we evaluate and interpret the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step. Here we consider the preprocessing steps with respect to their effect on the Data Mining algorithm results (for example, adding features in Step 4, and repeating from there). This step focuses on the comprehensibility and usefulness of the induced model. In this step the discovered knowledge is also documented for further usage.

3. APPLICATION

Advantages of using data mining in various applications such as Banking, Manufacturing and production, marketing, health care etc., are as follows:

1) Banking: Data mining supports banking sector in the process of searching a large database to discover previously unknown patterns; automate the process of finding predictive information. Data mining helps to forecast levels of bad loans and fraudulent credit cards use, predicting credit card spending by new customers and predicting the kinds of customer best respond to new loan offered by the backs.

2) Manufacturing and production: Data mining helps to predict the machine failures and finding key factors that control optimization of manufacturing capacity.

3) Marketing: Data mining facilitates marketing sector by classifying customer demographic that can be used to predict which customer will respond to a mailing or buy a particular product and it is very much helpful in growth of business.

4) Health-Care: Data mining supports a lot in health care sector. It supports health care sector by correlating demographics of patients with critical illnesses, developing better insights on symptoms and their causes and learning how to provide proper treatments

5) Insurance: Data mining assist insurance sector in predicting fraudulent claims and medical coverage cost, classifying the important factors that affect medical coverage and predicting the customers' pattern which customer will buy new policies.

5. OPEN SOURCE TOOLS

Data mining has a wide number of applications ranging from marketing and advertising of goods, services or products, artificial intelligence research, biological sciences, crime investigations to high-level government intelligence. Due to its widespread use and complexity involved in building data mining applications, a large number of Data mining tools have been developed over decades. Every tool has its own advantages and disadvantages

The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult. The three open source tools available for data mining are briefed as below

5.1. WEKA TOOL

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data pre-processing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. Weka is probably the most successful open source data mining software which has inspired by the development of other programs with more sophisticated graphical user interface and better visualization methods. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values.

Technical Specification:

· First released in 1997.

· Latest version available is WEKA 3.6.11.

· Has GNU general public license.

· Platform independent software.

· Supported by Java

· Can be downloaded from www.cs.waikato.ac.

Specialization:

· Weka is best suited for mining association rules .

· Stronger in machine learning techniques.

· Suited for machine Learning.

5.2. ORANGE

Orange is an open source data mining and visualisation software with active community and which helps novice and experts for their analysis. It has the ability to work under various platforms like windows, Mac Os C and GNU/Linux operating systems and it's packed with data analytics features. It enables design of data analysis process through user friendly visual programming or python scripting. Hence, this can be used as a scripting language for respective tasks of data mining. It represents most major algorithms for data mining and contains different visualisation, from scatter plots, bar charts, trees to dendrograms, networks and heat maps. It remembers user's choices, suggests most frequently used combinations, and intelligently chooses which communication channels to use. It has specialised add-ons like Bio orange for bio informatics.

Technical Requirements:

· Developed in 2009.

· Latest version available is Orange 2.7

· Licensed by GNU General Public License

· Compatible with Python, C++, C.

· Can be downloaded from www.orange.biolab.si

Specialization

· Open source data visualization and analysis for novice and experts

· It contains components for machine learning and add-ons for bioinformatics and text mining. Along with it's also packed with features for data analytics.

· Specialized for data visualization along with mining

5.3. R

R is an open source programming language and environment for statistical computing and graphics. R provides a wide variety of graphical and statistical techniques such as linear and non-linear modelling, classical statistical tests, time series analysis, classification clustering and is highly extensible. Researchers in various fields of applied statistics have adopted R for statistical software development and data analysis. Extensibility and superb data visualisation are the two main reasons for the success of R.

1) Technical Specification

· First released in 1997

· Latest version Available is 3.1.0

· Licensed by GNU General Public License

· Cross Platform

· C, Fortran and R

· www.r-project.org

Specification:

· It has a large number of users, in particular in the fields of bio-informatics and social science. It is also a free ware replacement for SPSS.

· Suited for Statistical Computing

6. CONCLUSION

The objective of data mining is to design and work efficiently with large data sets. Data mining is the component of wider process called knowledge discovery from database. Data Mining is the process of analysing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" Open-source data mining suites of today have come a long way from where they were only a decade ago. They offer nice graphical interfaces, focus on the usability and interactivity, support extensibility through augmentation of the source code or (better) through the use of interfaces for add-on modules. They provide flexibility either through visual programming within graphical user interfaces or prototyping by way of scripting languages. The study presented the specific details along with description of various open source data mining tools enlisting the area of specialization

7. REFERENCES

1. Hand David, Mannila Heikki, Smyth Padhraic.: “Principles of data mining”, Prentice hall India, pp.1, 2004.

2. Sethi I. K., “Layered Neural Net Design Through Decision Trees, Circuits, and Systems”, IEEE International Symposium,1990.

3. Meheta M., Aggarwall R., Rissamen I. : “SLIQ:A fast Scalable Classifier for Data Mining”, In Proc. International Conference Extending data base Technology(EDBI), Avignon, France, March 1996.

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

5. 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.

6. S. R. Mulik, S. G. Gulawani:“ PERFORMANCE COMPARISON OF DATA MINING TOOLS IN MINING ASSOCIATION RULES”, International Journal of Research in IT, Management and Engineering (IJRIME), Volume1Issue3 ISSN: 2249- 1619

7. Ralf Mikut and Markus Reischl Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011.

8. Witten, I.H., Frank, E.: "Data Mining: Practical machine Learning tools and techniques", 2nd addition, Morgan Kaufmann, San Francisco (2005).

9. Alcala-Fdez, J.,L., del Jesus, M.J., Ventura, s., Garrell, J.M, Otero, J., Romero C., bacardit, j., Rivas, V.M., Fernandez, J.C., Herrera., F., : "KEEL: A software tool to Assess Evolutionary Algorithms to Data mining Problems", Soft computing 13:3,pp 307-318(2009).

10. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler,T. "YALE: Rapid Prototyping for Complex Data Mining tasks", in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-06), pp. 935-940, 2006.

11. http://orange.biolab.si/features/

12. https://github.com/Dans-labs/recommender-systems/blob/.../datamining.r

13. http://www.r-project.org/

14. http://www.knime.org/

15. http://rapidminer.com/