www.ijemr.net ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Volume-7, Issue-2, March-April 2017

International Journal of Engineering and Management Research

Page Number: 356-358

Comparative Study of Various Open Source Data Mining Tools

Dileep Kumar Singh
Department of CSE, SOET, Jagran Lakecity University, Bhopal INDIA

ABSTRACT

Data mining is one of the most wildly required fields in today's world where huge data are stored and available in digital form. Data mining can be used for extracting the useful information from the heap of the available data. It will be very difficult to extract useful information without using the computer software/tools. Fortunately there are various commercial as well as open source data mining tools are available with its own pro & cons. This paper provides the comparative study of various available open source data mining tools such as Weka, Orange, RapidMiner, KNIME etc.

Keywords— open source data mining tools, Weka, Orange, RapidMiner, KNIME.

I. INTRODUCTION

Data is representing the raw facts. Data mining is one of the most widely used methods for data exploration and fetching the useful information that can be used as an asset in the today's world [1, 2, 5]. Although there are various commercial tools are available with the great power of mining the data, but they cost. Fortunately there are various open source data mining tools are also available and some of having equal competency.

From the available various open source data mining tools, six well known and most frequently used tools are selected for our purpose: 1) RapidMiner, 2) Weka, 3) Orange, 4) Knime, 5) Keel and 6) R [4, 6, 7]. This paper provides the general discussion and comparative study of various features of these selected open source data mining and classification tools.

II. DATA MINING

Data mining is the process of extracting useful information or the required pattern from the big set of data. There are many applications of data mining where the information/ knowledge extracted can be used such as exploration for the new ideas, analysis, decision making, policy making, controlling etc.

Data mining task mainly divided into two categories: Descriptive and classification & prediction [3]. Data mining descriptive function focused on Knowledge Discovery in Databases (KDD) to extract the human understandable data pattern from the given huge set of data. Whereas the classification & prediction function of data mining are used for the prediction of class labels or some values [3].

III. OPEN SOURCE DATA MINING TOOLS

There are various open source data mining tools are available, some of them are even equally or more powerful then the some commercial IDS tools, such as RapidMiner, Weka, Orange, Knime, Keel, R [7] etc.

RapidMiner is a java based open source data mining tool for with various data analysis features. It requires very less coding to perform the data mining functions and having a vast functionality apart from data mining provides visualization, preprocessing statistical analysis etc [2, 7, 8].

Weka is a java based open source data mining tool. This tool is having many data mining functionality with various algorithms for the data analysis, prediction and visualization [9].

ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

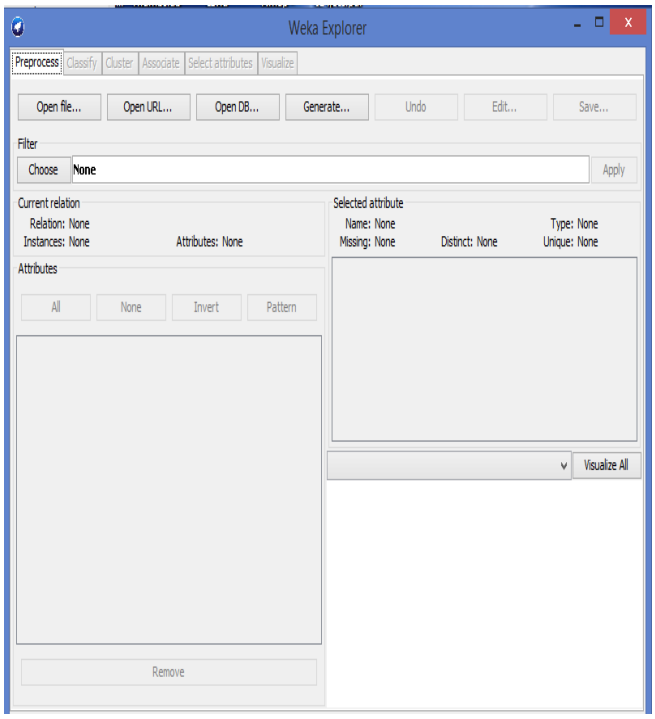

Figure 1: Weka GUI

ORANGE is a python based power full data mining tool with the capability of machine learning features and functionality [10].

KNIME is a Java data mining tool with the various components of machine learning and data mining are integrated. This tools is easily extensible with a good user friendly graphical user interface (GUI) [11].

KEEL is a java based data mining tool with the various features such as clustering, regression, pattern mining, classification etc. It contain the large collection of data mining techniques, preprocessing algorithms, statistical analysis techniques etc [7,12].

R programming languages is written in C, FORTRAN and R itself. R is one of the widely used data mining and analysis tool. R provides the functionality of statistical analysis and graphical techniques, classification, clustering and various linear & non-linear modeling techniques [7, 13].

IV. COMPRATIVE STUDY

Comparative study based on the technical specifications and Characteristics of various open source data mining tools are presented in the table below:

Parameters	**RAPID MINER**	**WEKA**	**ORANGE**	**KNIME**	**KEEL**	**R**
Latest version	7.4	Weka 3.8	3.4.2	3.3.2	3.0	3.4.0
Operating System Supported	Windows, macOS, Linux	Windows, macOS, Linux	Windows, macOS, Linux	Windows, macOS, Linux	Platform Independent	Windows, macOS, Linux
License	AGPL-3.0	GNU General Public License	GNU General Public License	GNU General Public License	GPLv3	GNU General Public License
Programming Language	Java	Java	C++ Python	Java	Java	C, Fortran and R Programming Language
Availability	Open Source	Open Source	Open Source	Open Source	Open Source	Open Source
Portability	Cross Platform	Cross Platform	Cross Platform	Cross Platform	Cross Platform	Cross Platform
Compatibility with database (Ex MySQL)	Yes	Yes	Yes	Yes	Yes	Yes
GUI	Yes	Yes	Yes	Yes	Yes	Yes (Ex – RStudio, R

						Commander)
Big Data Processing	No	Yes	No	No	Yes	Yes
Key Features	Freely available User-friendly GUI Large Number of functions for analysis and data handling A macro viewer	Platform-independent Easy to use freely available Support very large collection of data mining algorithms	Freely available Large toolbox and Scripting interface User-friendly interface Visual Programming	Freely available Well-defined API for plug-in extensions Import/export of workflows Scalability and High extensible User-friendly interface	User-friendly interface Has an Excellent statistical analysis library Reduces programming work Contains a big collection of algorithms Freely available	Having a wide variety of statistical and graphical techniques Highly extensible freely available Import and export of data from spreadsheet is easier Easier to combine with other statistical calculations

Table – 1: Comparative study based on the technical specifications and Characteristics

V. CONCLUSION

Huge digital data is available in today's world. But this data will be of no use if we cannot be able to extract the useful information or pattern. As digital data are increasing day by day, it is required that there should available an efficient & cost effective data mining tools. Open source data mining tools are freely available. The General discussion and the comparative study based on the technical specifications and Characteristics of various open source data mining tools are presented and discussed in this paper.

REFERENCES

[1] Hayden Wimmer and Loreen M. Powell "A Comparison of Open Source Tools for Data Science", Conference on Information Systems Applied Research Wilmington, North Carolina USA, ISSN: 2167-1508, 2015
[2] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa, "A Comparison Study between Data Mining Tools over some Classification Methods", International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, 2011
[3] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, ISSN: 0738-4602-1996, 1997
[4] Pritam H. Patil, Suvarna Thube, Bhakti Ratnaparkhi and K. Rajeswari, "Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining", International Journal of Computer Applications, ISSN: 0975 – 8887, 2014
[5] S. Usharani and K. Kungumaraj, "A Survey on Data Mining with Big data - Applications, Techniques, Tools, Challenges and Visualization", International Journal of Advanced Research in Computer and Communication Engineering, ISSN: 2278-1021, 2015
[6] Neha Chauhan and Nisha Gautam, "Parametric Comparison Of Data Mining Tools", International Conference on recent innovations in science, Engineering and management, ISBN: 978-81-931039-9-9, 2015
[7] Kalpana Rangra and Dr. K. L. Bansal, "Comparative Study of Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, 2014
[8] https://rapidminer.com/
[9] http://www.cs.waikato.ac.nz/ml/weka/
[10] https://orange.biolab.si/
[11] https://www.knime.org/
[12] http://www.keel.es/
[13] https://www.r-project.org/