Comparative Study of Data Analytics Open Source Tools for Educational Data Analytics

Bharati Kawade[1], Dr. Aruna Deoskar[2]
[1] Research Scholar IICMR,Pune, India
[2] Principal, ATSS CBSCA College, Pune, India

Abstract:Educational data analytics is useful and important in data driven decision making to achieve better educational outcomes.Open source data mining tools are extendable, and may offer substantial flexibility in handling various types of data. Comparative study of different data mining tools come as comprehensive, integrated suites featuring a wide range of data analysis components.The comparative study provides the specific details along with description of various open source data mining tools enlisting the area of specialization.Also the tools are compared based on the use and applicability of these tools by different researchers for their research purpose. WEKA is the most popular tool used by many researchers.

Index Terms*-Data Analytics, Comparison, Open Source Tools*

I. INTRODUCTION

Data driven decision making in education is important to achieve better educational outcomes to meet requirements of accountability to regulatory standards, continuous self-evaluation and improvement. Data driven decision making in education is the systematic collection, analysis, evaluation and interpretation of the data to improve the processes and educational outcomes. Educational data analytics refers to methods and tools for analysing large sets of data from sources for sustainable and effective data-driven decision making in teaching and learning [1].

Data mining, statistics, machine learning, data visualization, and knowledge engineering contribute their methodsin research fields for data exploration and model inference. Withindata mining, there is a group of tools that have been developed by a research community and data analysis enthusiasts; they are offered free of charge using one of the existing open-source licenses. Open-source data mining suites may not be as stable and visually finished as their commercialcounterparts but instead may offer high usefulness through alternative,exciting, and cutting-edge interfaces and prototype implementations of themost recent techniques. Being open source they are by definition extendable,and may offer substantial flexibility in handling various types of data. Most open-source data mining tools today come as comprehensive, integrated suites featuring a wide range of data analysis components [2].

II. COMPARATIVE STUDY

The comparative study of different data mining open source tools has been done based on their features, specifications, advantages and limitations. Also the tools are compared based on the use and applicability of these tools by different researchers for their research purpose. The comparative study provides the specific details along with description of various open source data mining tools enlisting the area of specialization.

The following tools are compared: Rapid Miner, Orange, KNIME, WEKA, KEEL and R Programing [3].

Table 1: Comparison of Data Analytics Open Source Tools

Sr. No.	Name of Tool	Features	Specifications	Advantages	Limitations
1	Rapid Miner	• Functions for data analysis, handling including multiple new aggregation functions • File operators to operate directly from Rapid Miner • A macro viewer shows macros and their values in real time during process execution • Intuitive GUI	Released on 2006 Latest version available is Rapid miner 6. Licensed by AGPL Proprietary Cross platform i.e. can be installed on any operating system Language Independent Website: www.rapidminer.com.	Visualization, Statistical, Attribute Selection, Outlier detection, parameter optimization	Requires prominent knowledge of database handling. The software requires the ability to manipulate SQL statements and files.
2	Orange	• Visual Programming, • Visualization, • Interaction And Data Analytics • Large toolbox, • Scripting interface • Extendable Documentation	Developed in 2009. Latest version available is Orange 2.7 Licensed by GNU General Public License Compatible with Python, C++,C. Website: www.orange.biolab.si	Better debugger, Shortest scripts, poor statistics, suitable for novoice Experts	Limited list of machine learning algorithms. Machine learning is not handled uniformly between the different libraries. Weak in classical statistics; It provides no widgets for statistical testing. Reporting capabilities are limited to exporting visual representations of data models.
3	KNIME	• Scalability , Intuitive user interface ,High extensibility • Well-defined API for plugin extensions • Sophisticated data handling, intelligent automatic caching of data, Data visualization • Import/export of workflows, Parallel execution on multi-core systems • Command line version	Released on 2004. Latest version available is KNIME2.9 Licensed By GNU General Public License Compatible with Linux ,OS X, Windows Written in java Website: www.knime.org	Molecular analysis, Mass spectrometry. Chemistry Development kit	Limited error measurements, no wrapper methods for descriptor selection, poor parameter optimization
4	WEKA	• Forty nine data pre-processing tools, • Seventy six classification/regression algorithms, • Eight clustering algorithms, • Fifteen attribute/subset evaluators, • Ten search algorithms for feature selection. • Three algorithms for finding association rules • Three graphical user interfaces – "The Explorer" (exploratory data analysis) – "The Experimenter" (experimental environment) – "The Knowledge Flow" (new process model inspired • Poor documentation	First released in 1997. Latest version available is WEKA 3.6.11. Has GNU general public license. Platform independent software. Supported by Java Website: www.cs.waikato.ac.	Ease of use, It is also suitable for developing new machine learning schemes. Weka loads data file in formats of ARFF, CSV, C4.5, binary. Though it is open source, Free, Extensible, Can be integrated into other java packages	Poor documentation, weak classical statistics, Worse connectivity to Excel spreadsheet and non-Java based databases. CSV reader not as robust as in Rapid Miner Does not have the facility to save parameters for scaling to apply to future datasets. Does not have automatic facility for Parameter optimization of machine learning /statistical methods
5	KEEL	• Classification Discovery, • Cluster Discovery, • Regression Discovery, • Association Discovery, • Data Visualization • Discovery Visualization, • A user-friendly graphical interface,	First released in 2004. Latest version available is KEEL 2.0. Licensed by GNU, general public license. Can run on any platform. Supported by java language.	Evolutionary algorithms, fuzzy systems	Limited algorithms Efficiency is restricted by the number of algorithms it support as compared to other tools.

		• Evolutionary learning	Website: www.sci2s.ugr.es/keel.		
6	R	• Data Exploration, • Outlier detection, • Clustering , • Text Mining, • Time Series Analysis , • Social Network Analysis , • Parallel Computing, • Graphics, • Visualization of geo spatial data, • Web Application Big data • Data and error handling, • Requires array language, • Poor mining	First released in 1997 Latest version Available is 3.1.0 Licensed by GNU General Public License Cross Platform C, Fortran and R Website: www.r-project.org 2	Purely statistical	Less specialized for data mining, requires knowledge of array language

III. LITERATURE REVIEW

The paper by Priti S. Patel, Dr. S.G. Desai [4] discusses about various available data mining tools and compares their utilities. All Data mining tools have their own pros and cons with respect to efficiency and accuracy. Most of the researchers work on R, WEKA and Rapid miner for their research work.

Harshvardhan Solanki [5] has discussed different data analytics tools with their specification, techniques and the algorithms used along with features. The author has discussed the tools on the basis of analysis and processing capabilities, database system support, graphical representation and issues. The author used WEKA, KNIME and TNAGRA data mining tools for four different classification algorithms namely ZeroR, OneR, C4.5 and KNN over two datasets. The results showed that WEKA is the best among the three tool in terms of accuracy.

The author did comparative study of WEKA, RapidMiner, Tableau and R programming tools along with their features such as usability, speed, visualization, different algorithms supported, data set size, memory usage and primary usage [6].

Sharon Christa, K. LaxmiMadhuri, V.Suma studied [7] the popular data mining tools Orange, WEKA, RapidMiner, KNIME based on their features, specification and capabilities. According to the authors, there is need of comparative study for selection of the most suitable tool for particular data domain. It is important to select most appropriate algorithm for mining from particular data domain. They also mentioned there is need of general framework for a methodology to deal with manual implementation of mining algorithm with the data mining tools. It is concluded that WEKA is user friendly with greater accuracy.

The study conducted by the authors represents efforts taken for this categorization of tools based on several factors like target users, data organization and supported data and data structures, tools and services for data exploration, visualization and interface design and divided among nine different categories. This categorization has also helped in the determination of the tool considered best for application of a particular data mining task [8].

WEKA is the most popular tool used by many researchers because of the following benefits of WEKA tool [9].

1. Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms for data mining tasks.
2. These algorithms can either be applied directly to a data set or can be called from your own Java code.
3. The WEKA workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modelling, together with graphical user interfaces for easy access to this functionality.
4. WEKA would be considered the best tool because of its many built-in features.

IV. CONCLUSION

The tools compared are WEKA, Rapid Miner, Orange, Tanagra, R and KNIME. Educational data analytics refers to methods and tools for analysing large sets of data from sources for sustainable and effective data-driven decision making in teaching and learning. The comparative study of different data analytics open source tools has been done based on their features, specifications, advantages and limitations. From the review of different research papers, it is found that WEKA is the popular open source tool among the researchers due to its many built-in features.

REFERENCES

[1] Demetrios G. Sampson, Educational Data Analytics Technologies for Data-Driven Decision Making in Schools, https://elearningindustry.com/educational-data-analytics-technologies, Oct 2016

[2] BlazZupan, JanezDemsar,Open-Source Tools for Data Mining, Clinics in Laboratory Medicine, Elsevier, 2008, 37–54

[3] KalpanaRangra, Dr. K. L. Bansal, Comparative study of Data Mining Tools, International Journal of Advanced Research in Computer Science & Software Engineering, Volume 4, Issue 6, June 2014, ISSN: 2277 128X

[4] Priti S. Patel, Dr. S.G. Desai, A Comparative Study on Data Mining Tools, International Journal of Advanced Trends in Computer Science and Engineering, Volume 4 No.2, March - April 2015, ISSN 2278-3091

[5] Harshvardhan Solanki, Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory, International Journal of Computer Applications (0975 – 8887) Volume 75 – No.16, August 2013, https://research.ijcaonline.org/volume75/number16/pxc3890862.pdf

[6] AkshayVishwanathBhinge, A comparative analysis of data mining tools in agent based systems, Master of Science Project, The faculty of the Department of Computer Science California State University, Sacramento

[7] Sharon Christa, K. LaxmiMadhuri, V.Suma, A comparative analysis of data mining tools in agent based systems

[8] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa, A Comparison study between data mining tools with some classification methods”

[9] WEKA, the University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/ downloading.htm