

Experimental Evaluation of Open Source Data Mining Tools (WEKA and Orange)

Ritu Ratra¹, Preeti Gulia²

¹ Research Scholar, Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, Haryana, India

² Assistant Professor, Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, Haryana, India

¹ rituharjai25@gmail.com_ ² preeti.gulia81@gmail.com

Abstract: Nowadays, it is possible for every organisation to manage the large dataset at minimum cost. But in order to collect the fruitful information, it is mandatory to utilize the large volume of stored data. Data mining is an on-going process of searching pattern and collecting useful information from large datasets for future use. There is no doubt that Data mining is very important in various areas like education, military, e-business, healthcare etc. The main objective of data mining process is to supervise the data from various sources in different manner then assemble it to collect the useful information. It can be done by the help of various tools and techniques. There are a number of data mining tools available in the digital world that can help the researchers for the evaluation of the data. These tools work as an interface to receive the data and to extract some meaningful patterns out of large dataset. Selection of best tool according to requirement is not an easy task. In order to find out the best data mining tool for classification problem, comparison of various tools is necessary on the basis of different parameters. In this paper, data mining tools **WEKA and Orange** are analysed on the basis of implementation of parameters. The main objective of this comparison is to help the researchers to select the suitable tool from these two.

Keywords: Classification, Naïve Bayes, Random Forest tree, WEKA, Orange, Precision, Recall.

I. INTRODUCTION

In present scenario, data is increasing day by day according to different parameters. It is very difficult for a person to analyse the large volume of data for perfect decision making. Hence, there is need of data mining to extract valuable and useful data from the available data. Data mining is the process of finding the most useful knowledge from the large volume of data available in databases or data repositories. Classification is one of the most important problems in data mining, which is a collection of finding rules that divides the given data into different classes. These classes are predefined. There is trillions of data available in the form of different types in digital world. Manually, it very

time consuming task to execute that data. So, there is a requirement of automated tools that can help the researcher to convert that messy data into useful information. Few years ago, there are so many data mining software tools have been developed to overcome this problem. Some of them are freely available as open-source tools. The affirmation of open source tools of information sharing for implementations of different machine learning algorithms can be most beneficial for the complete field [10].

In this paper, a comparative study is conducted among various classification algorithms like Random Forest tree, K-Nearest Neighbour and Naïve bayes algorithm using WEKA and Orange tool. The evaluation metrics Precision and Recall are used to analyze the performance of the both the tools with the help of various classification algorithms. The following Classification Algorithms have been used for the experimentation:

- Naïve Bayes: Naive Bayes classifier is a group of simple probabilistic algorithms. These are based on Bayes' theorem. In it the algorithm is applied with powerful assumptions between the various features.
- K-Nearest Neighbour: It is a simple classifier that saves all cases that are available and then generates new cases based on a similar measurement e.g., distance functions.
- Random forest: It is almost same as Decision tree classifier. But it adds some randomness to the model at the time of making the tree. It can produce great results without the help of hyper parameter. It builds different decision trees and then combines them to generate more stable prediction.

To handle huge volume of data, there are several tools available for the user. Moreover it is not easy to include all the features in single tool. That's why a number of different varieties of tools have been introduced [8][10]. In this paper, two data mining tools i.e. WEKA and Orange will be compared. These tools have different characteristics, functionality and capabilities. Researchers can use

these according to their research activities requirements. These tools are continuously upgraded with new features as per the needs of the user which are changing day by day. It is very typical to deal with the complexity of huge data.

The rest of the paper flow is as follows: section II describes open source software, section III describes the comparative study of WEKA and Orange tool on the basis of parametric comparison and experimental analyses and conclusions and future scope is discussed in Section IV.

II. OPEN SOURCE SOFTWARE

Open source software is computer software in which the source code publically available for user under a license. In this license copyright holder permit the users to use it. They can inspect and update it and can also distribute it to anyone for use. Open source software is cheap and flexible because it is developed by group of company rather than a single programmer. The common open-source licenses are GPL, general people consent (GNU.org, 2015a), GNU (GNU.org, 2015b), Mozilla Public License (MPL), Berkeley Software Distribution (BSD), Netscape Public License (NPL) and Lesser General Public License (LGPL) [10]

There are lot of open-source data mining tools are available for data mining process such as the KNIME, RapidMiner, Orange, WEKA, R-Programming etc. These data mining tools are assembled with a set of techniques and algorithms that are very helpful in better data analytics. Researcher can take help in classification, clustering and visualization of data. These tools are also useful for regression analysis, Predictive analytics etc. These tools are present with their own functionalities to help the user with their work. In this paper, WEKA and Orange tool are described.

A. WEKA: WEKA is a popular toolkit for learning the machine learning algorithm. It was originally

developed at the University of Waikato in New Zealand. WEKA is a data mining tool that allows data pre-processing process. Attribute selection is very interesting feature of WEKA. It enhances the effectiveness and accuracy of selected data. WEKA comes with these functionalities: command-line interface (CLI), Explorer, Experimenter and Knowledge flow and weka workbench. Explorer is used to define the data source, preparation, selection of algorithms, and visualization. The Experimenter is helpful for the comparison of the different algorithms on same dataset.

In WEKA software,secondary data can be used to analyse. Researcher can apply algorithm to a data set and can analyse the results to make decision about the data, various predictions can also generate to predict the new instances. Even though, this tool support a lot of model evaluation metrics, but there is absence of many data survey and visualization methods [6]. WEKA is more towards the classification and regression and less towards the descriptive statistics and clustering methods. There is less support for big data and semi-supervised learning in WEKA [11]. WEKA is a tool that available freely for download. Popular features of WEKA are shown in Figure 1.

As shown in the figure most famous feature of WEKA are as: It is an open source data mining tool that is based on JAVA language. It is very easy to understand and use for the beginners and it has the capability of running and comparing several algorithms. It is able to perform different data mining activities including: Data preprocessing, clustering, Classification, Association Rule, knowledge discovery etc. There are a number of built in features in WEKA that makes easy for the users. Without the knowledge of programming and coding, researcher can use it for analyses.

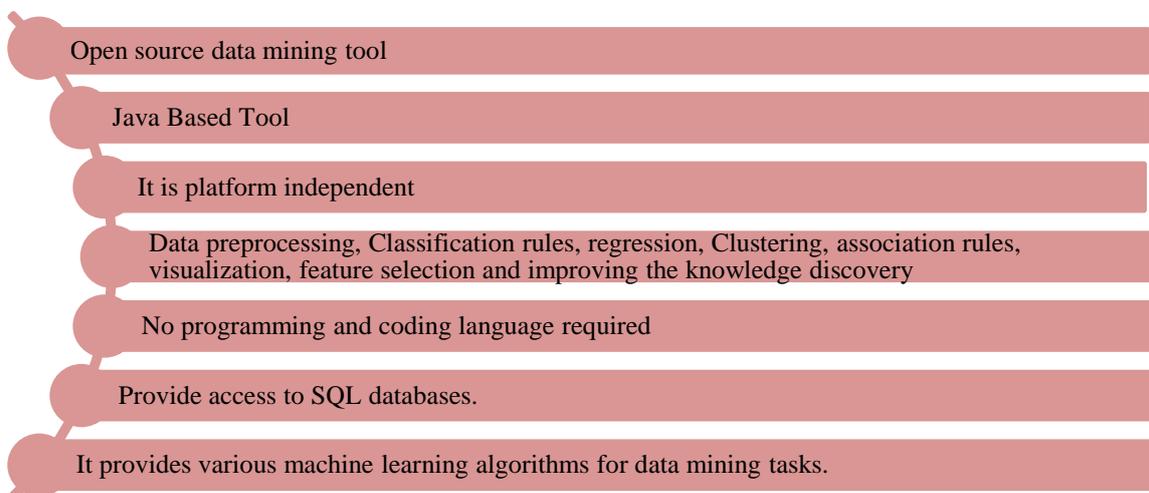


Figure 1: Features of WEKA

B. Orange

Orange is also freely available open source data mining software. It can be useful for explorative data analytics and visualization. It provides a platform for different experiment selection. Orange is very effective when the concept of innovation, reliability

or quality is involved[12],[13]. Basically Orange Canvas is quite useful for visual programming interface. It provides a well-structured view of different features. These features are depicting in figure 2.

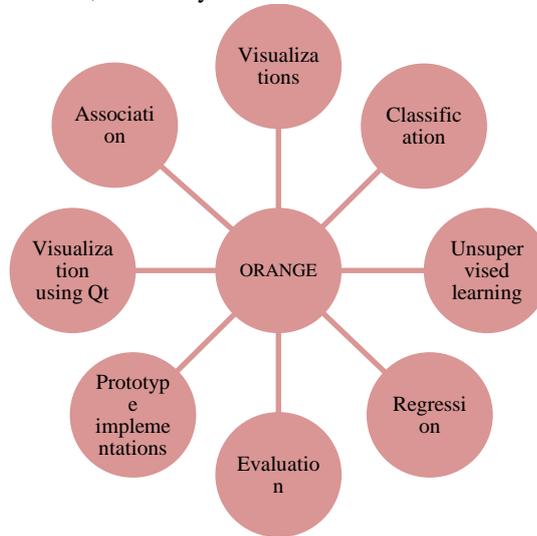


Figure2: Features of Orange tool

This figure depicts different features of Orange Data Mining. Visualization of data, classification, evaluation,

unsupervised learning, association, visualization using Qt, and prototype implementations are some famous features of Orange. The cross-platform application of orange is QT and developers can use UI framework for applications. It can be done by using C++. CSS & JavaScript like language. Orange tool’s working is visually represented by using different widgets for example reading file, training SVM classifier etc. Every widget is self-explained i.e. has a short description about itself is within the interface. To program, first of all widgets are placed on the canvas and then inputs and outputs are connected. The widgets available are limited in Orange in counting as compared to other tools.

The activity is measured by the frequency of updates and time of latest update. Whenever there is comparison between two tools, then it becomes necessary to compare them both parametrically and experimentally. After then reliable results could be achieved. So in this manner, let us start with parametrical comparison and then analysed the experimental results.

A. Parametric Comparison: In parametric comparison, all the characteristics of tools are taken from previous available sources. These characteristics were listed in Table I. Some characteristics are common in both tools for example Graphical User Interface (GUI) functionalities, command line of are in both tool [18], [19].

III. COMPARATIVE ANALYSIS

Table I: General Characteristics of Open-Source DM Tool WEKA and Orange

Parameters	WEKA	ORANGE
Company Name	University of Waikato New Zealand	University of Ljubljana Switzerland
Source	http://www.cs.waikato.ac.nz/ml/weka/	http://orange.biolab.si
Programming language	JAVA	C++, Python
Released date`	1993	1996

License	GNU General Public License	Open-source, GNU GPLv3
Availability	Open Source	Open Source
Current Version	3.8	3.24.1
Areas	Machine learning, Data visualization, time series and analysis, text mining, fraud detection	Marketing, Direct Mail Financial Service, Manufacturing, Health Care, Military
Portability	Cross Platform	Cross Platform
Logo		
GUI/Command line	Both	Both

B. Technical comparison of WEKA and Orange

To make technical comparison between these tools, first of all these free data mining and knowledge discovery tools are to be downloaded. After then specified the datasets to be used and selecting some classification algorithm to test the performance of tools. Precision and Recall are most popular evaluation metrics of model. To make comparison these are used in this paper.

1) Precision: Precision is positive predictive value. It is defined as the average probability of relevant retrieval.

Precision = Number of true positives/(Number of true positives + False positives).

2) Recall: Recall is the average probability of complete retrieval.

Recall= True positives/True positives + False negative

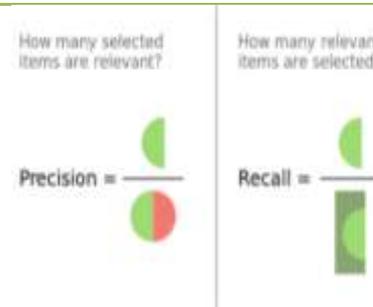


Figure 3: Precision and Recall [15]

Data set: The dataset Heart Disease is used for the work. It is taken from UCI Machine Learning repository and Cleveland heart disease dataset is selected for the study. It has 303 instance and 76 attributes.

The comparison between these tool are well shown through the table II and Table III

Table II: Comparative study of WEKA and Orange tool Precision Metric

Classifier	WEKA(%)	Orange(%)
Naïve bays	83.7	82.4
Random	81.8	77.9
Forest		
k-nearest	75.3	58.0

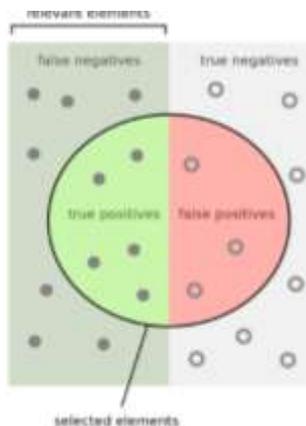


Table III: Comparative study of WEKA and Orange tool Recall Metric

Classifier	WEKA(%)	Orange(%)
Naïve bays	83.7	80.6
Random Forest	81.9	73.4
k-nearest	75.2	54.7

When the dimension of the input data is high, then Naïve Bayes Classifier algorithm is most suited. Naive Bayes is particularly applicable in artificial intelligence. When comparative study is made, the analysis of precision and recall is analysing for heart disease data sets precision in Orange 82.4% and Recall 80.6%. In WEKA the value of precision is 83.7% and Recall 83.7%. WEKA tool is best in precision and Recall as compare to Orange tool in Naïve bayes classifier. Same is happened with Random forest and k-nearest classifiers. In Random Forest, precision value in Orange is 77.9% and Recall value is 73.4%. In WEKA the value of precision is 81.8% and Recall 81.9%. And in k-nearest algorithm, precision value in Orange is 58% and Recall value is 54.7%. In WEKA the value of precision is 75.3% and Recall 75.2%.

IV. CONCLUSION AND FUTURE STUDY

This paper presents the study of two different open source Data mining tools along with their features-WEKA and Orange. Both tools have their own merits and demerits This paper specifies the comparison between these tools by experimental analysis and by using their parameters. This comparative study is based on datasets and algorithms. It may be possible that the results may vary with different datasets or algorithms. The comparative analysis is helpful in learning and selection of the data mining tools as per the areas. By employing experimental study, it is to be concluded that WEKA tool is better than Orange. It can be stated that WEKA has most desired features for a fully-functional and user friendly platform for classification problems. So, WEKA can be recommended for Classification problems of data mining. In the future work, different data sets and different problems like clustering, association rule mining will be taken and applied using these tools.

ACKNOWLEDGEMENTS

The authors are thankful to the <http://archive.ics.uci.edu/ml/datasets/heart+Diseasef> or providing the dataset.

REFERENCES

- [1] Jović, A., Brkić, K., & Bogunović, N. (2014). "An overview of free software tools for general data mining".

- Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention, (May), 26–30. Retrieved from http://www.zemris.fer.hr/~ajovic/articles/MIPRO_2014_final.pdf
- [2] Alcalá-Fdez, J., Sánchez, L., & García, S. (2009). "KEEL: a software tool to assess evolutionary algorithms for data mining problems". *Soft Computing*. Retrieved from <http://link.springer.com/article/10.1007/s00500-008-0323-y>
- [3] Collier, K., Ph, D., Carey, B., & Marjaniemi, C. (1999). "A Methodology for Evaluating and Selecting Data Mining Software" Keywords: Data Mining, Tool Evaluation, Knowledge Discovery, 00(c), 1–11.
- [4] Sonnenburg, S., Braun, M., & Ong, C. (2007). "The need for open source software in machine learning", 8, 2443–2466. Retrieved from <http://researchcommons.waikato.ac.nz/handle/10289/3928>
- [5] Chen, X., Ye, Y., Williams, G., & Xu, X. (2007). "A survey of open source data mining systems". *Emerging Technologies in Knowledge Discovery and Data Mining*, (60603066), 3–14. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-770183_2
- [6] Jović, A., Brkić, K., & Bogunović, N. (2014). "An overview of free software tools for general data mining". *Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014 37th International Convention, (May), 26–30. Retrieved from http://www.zemris.fer.hr/~ajovic/articles/MIPRO_2014_final.pdf
- [7] Kalpana Rangra, Dr. K. L. Bansal. "Comparative Study of Data Mining Tools", presented at International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, 2014.
- [8] Dr. Anil Sharma, Balrajpreet Kaur, "A RESEARCH REVIEW ON COMPARATIVE ANALYSIS OF DATA MINING TOOLS, TECHNIQUES AND PARAMETERS", ISSN No. 0976-5697, International Journal of Advanced Research in Computer Science, volume 8, No. 7, July – August 2017.
- [9] H.Witten, E. Frank, M. A.Hall, "Data Mining practiced machine learning tools and techniques", 3rd ed., Morgan Kaufmann Elsevier: USA, 2011.
- [10] Predictive Analytics [Online]. Available from: <http://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics/>
- [11] Jović, A., Brkić, K., & Bogunović, N. "An overview of free software tools for general data mining. *Information and Communication Technology*", *Information and Microelectronics (MIPRO)*, 2014 37th International Convention, (May), 26–30. Retrieved from http://www.zemris.fer.hr/~ajovic/articles/MIPRO_2014_final.pdf
- [12] <http://www.kdnuggets.com/2015/12/top-7-new-features-orange-3.html>
- [13] Orange Data Mining, "Orange Data Mining Library Documentation Release 3". <http://orange.biolab.si/>
- [14] <http://Precision%20and%20recall%20-%20Wikipedia.PDF>
- [15] M.Hall, E.Frank, G.Holmes, B.Reutemann, IH Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, 2009.
- [16] A.Wahbeh, "A Comparison Study between Data Mining Tools over some Classification Methods," *International Journal of Artificial Intelligence*, 2012.
- [17] Swasti Singhal, Monika Jena. "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering" presented at International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-2, Issue-6, 2013.
- [18] <http://www.ionos.com>digitalguide>
- [19] <http://www.google.com>
- [20] Venkateswarlu Pynam, R Roje Spanadna, Kolli Srikanth, "An Extensive Study of Data Analysis Tools (Rapid Miner,

- Weka, R Tool, Knime, Orange)*”, SSRG International Journal of Computer Science and Engineering (SSRG – IJCSE) – Volume 5 Issue 9 – September 2018, ISSN: 2348 – 8387,pp 4-11.
- [22] [http://opensourceforu.com/2017/03/top-10-open-source-datamining-t ools/](http://opensourceforu.com/2017/03/top-10-open-source-datamining-tools/)
- [23] Nurdatillah Hasim, Norhaidah Abu Haris, “A Study of Open-Source Data Mining Tools for Forecasting”, IMCOM '15, January 08 - 10 2015, BALI, Indonesia.
- [24] Witten, I. H., & Eibe, F. (2005), “Data Mining: Practical Machine Learning Tools and Techniques”, (2nd ed., p. 525).
- [25] Sonnenburg, S., Braun, M., & Ong, C., “The need for open source software in machine learning”, 8, 2443–2466. 2007. Retrieved from <http://researchcommons.waikato.ac.nz/handle/10289/3928>.
- [26] 12 data mining tools and techniques [Online]. Available: <https://www.invensis.net/blog/data-processing/12-datamining-tools-techniques>.
- [27] A. kumar, et al., “ Data mining: various issues and challenges for future,” IJETA,2014
- [28] H. Nasereddin, “ NEW TECHNIQUE TO DEAL WITH DYNAMIC DATA MINING IN THE DATABASE,” IJRRAS,,December 2012.
- [29] J.Demšar and B.Zupan, “Orange: Data Mining Fruitful and Fun - A Historical Perspective”, 2012.
- [30] C.Shah, A.Jivani, ”Comparison of data mining classification algorithms for breast cancer prediction”, 4th ICCCNT ,IEEE,2013.
- [31] P.Kakkar, A.Parashar, “Comparison of different clustering Algorithm using WEKA tool”, International Journal of Advanced Research in Technology, Engineering and Science, 2014.
- [32] N.Chauhan and N.Gautam, “Parametric comparison of data mining tools,” IJATES, 2015.
- [33] A.Gupta, N.Chetty , S.Shukla, “A classification method to classify High Dimensional data”,IEEE,2015.