www.mecsj.com

An overview of the visualization features in open source data mining tools

Sawsan Alodibat

Jordan University for Science and Technology

Irbid, Jordan

smalodibat15@cit.just.edu.jo

Abstract

Typically, data mining tends to predict future patterns and behaviors considering current repositories and warehouses of data [1]. The adoption of open source software provides more independence to the researchers and developers contrasted to the closed source licenses that limit rights [8]. In this research, the authors are going to explore and compare five tools amongst WEKA, Orange, RapidMiner, Tanagra, and KNIME. The main purpose is to distinguish the chief differences between open source software tools and formerly to classify them into different levels of visualization functions: high support, middle support, and low support. At last, the final assumption revealed that the best software amongst the investigated tools in terms of visualization features is RapidMiner.

Keywords—data mining, open source, data mining tool, data visualization.

I. INTRODUCTION

Typically, data mining tends to predict future patterns and behaviors considering current repositories and warehouses of data [10]. The main use of data mining tools enables decision makers to take knowledge driven from historical databases and formulate it into decisions [13]. Many data mining algorithms and techniques have been developed and integrated into data mining tools for a purpose of extracting information, evaluating their efficiency and accuracy, and using it for prediction [14].

The main advantage of open source software is the possibility of free copying, using, running, distributing, modifying, and improving. Thus, the adoption of open source software provides more independence to the researchers and developers contrasted to the closed source licenses that limit rights [15]. On the other hand, the license and constraints of open source software should be understood by the developer to save his/her right associated with the modification on software [17]. Commonly, the open sources licenses are: GNU, GPL, BSD, LGPL, MPL, and NPL [5].

Data visualization refers to the use of representation tools to show a dataset through figures, charts, diagrams, charts, etc. Different data mining tools accomplish data visualization to show the results into visualized charts.

This research is focused on discussing the presence of the mutual visualization techniques in different open source data mining tools. Particularly, different visualization techniques will be applied in these tools to extract information from a dataset through classification techniques, showing the supported techniques in every tool. The primary source of data of this research will be the documentation, tutorials, and user manuals provided by the developers of tools. The secondary source of data will be the actual traversing through each mentioned tool and practicing each supported visualization technique, and hence showing it into charts.

This line of research falls under comparison studies; it creates a discussion of some visualization techniques (supported by open source data mining tools) in order to present an analytical approach given the main specifications and determinations. The authors do not have an intention to compare the advantages or disadvantages of open source data mining tools, but rather the goal is to establish a comparison between the

supported visualization techniques in each tool. Lastly but not last, this research will emphasize the presence and absence of common visualization techniques, revealing the survival and the lack of the advanced visualization functionalities for these tools.

I. RELATED WORK

In this section, a set of similar works are introduced to review other ultimate remarks of other works and to discriminate this research. The period of the works summarized here ranges from 2009 to 2016.

In the paper of [4], an overview of different open source data mining tools along with their features including RapidMiner, KNIME, and WEKA. Thus, a comparison among these tools is involved in terms of their advantages and disadvantages to provide interested researchers an assistance in selecting the best tool. One revealing point attained by the paper is the efficiency of RapidMiner compared to other two tools, guiding researchers to choose the most relevant data mining tool to any specific area.

The authors in [5] surveyed different open source data mining systems free to download over internet. The study included a comparison of twelve open source systems in terms of general specifications, data mining functions, usability, and data source formats. Thus, the advantages and disadvantages were discussed.

The authors of [6] presented a comparative study of open source data mining software from professionality perspective. They also discussed the real-time challenges associated with available data mining tools and the factors affecting the choice of appropriate software. The paper showed the how the pre-processing tasks can improve the performance of open source data mining tool when it integrated with agent based framework. The results of the work have clarified the functional specifications of open source data mining tools that enable the developers to improve data management through accurate analysis and evaluation.

In [11], the author introduced an implementation of Tanagra and Weka data mining tools in a healthcare dataset in order to determine the most imperative variables that define the diabetes impact on patients using Kidney Function Tests (KFT). The application of C4.5 algorithm was based on decision trees to compare the results of the two selected data mining tools. The author concluded that Tanagra is less error prone and more

efficient against classification performance compared to Weka, whereas Weka has best performance when using the mode of use training set.

An expert paper presented in [12] that describes the main characteristics of some open source data mining tools mostly used by community, mainly Weka, RapidMiner, Orange, R, scikit-learn, and KNIME. The paper provides researchers general characteristics of most common open source data mining tools and the corresponding advantages and disadvantages, considering most functions of data mining (clustering, regression, association, classification, visualization, etc.). Therefore, more advanced data mining research areas are also involved such as text mining and big data. At result, the paper exposed that no single open source data mining tool at the top, but rather every tool has cons and pros.

The main purpose of [16] was to describe the open source data mining software that have received much interest in many useful works, considering the advantages and disadvantages of these software. Thus, a list of most relevant web addresses, textbooks, and several manuals was included to help researchers how to find, use and cite related references. The work revealed that the researcher should review the manual of an open source data mining tool to decide which good software relevant to students for educational purpose.

A review of classification algorithms applied on different open source data mining tools is presented in [18]. Weka, RapidMiner, Tanagra, Orange, and KNIME were conducted to elaborate the correlation between them in terms of the accuracy of classification algorithms such as decision tree, naive Bayes, decision stump, and k-nearest neighbour. A dataset named Indian Liver Patient (ILP) was tested with classification algorithm using some open source data mining software to carry out the accuracies generated by each tool and then evaluate their performance.

A comprehensive analysis of some open source data mining tools is provided in [19] from theoretical viewpoint. For each selected tool, several aspects and details such as general characteristics, technical specifications, area of specialization, applications and features were involved. The work helps researchers to gain an insight of the future development of data mining tools through extending more functions to cover more fields efficiently. The result can be an efficient data mining product although of the increasing complexity of the development procedures and mechanisms.

The authors of [23] conducted a comparative study for four open source data mining tools to reveal the tool suitable to apply classification function to support decision making process. The type of dataset affects the performance of classification task carried out by each investigated tool. For this reason, six classification algorithms (decision tree, naïve Bayes, k-nearest neighbour, support vector machine, zero rule, and one rule) were tested on nine different datasets to judge the best tool. The authors revealed that the results of classification task changes with the way of implementation and the dataset conducted, therapy no single tool is best.

The work in [25] explored and overviewed the most commonly used open source data mining tools with respect to the state of the art and knowledge discovery. The work implies that academics, practitioners, and data scientists should select the apt open source data mining tool suitable to the requirements of the specified area and field. A detailed information was presented to describe each investigated open source data mining tool including the history, development, and the capabilities. The available characteristics of the open source data mining tools were organized into a matrix to show their descriptions in detail.

The aim of [26] was to build a predictive model for the probability of default through applying six data mining functions. In the paper, an experiment was conducted to overcome the shortcoming in the accuracy of the predicted default probability based on binary classification by proposing more valuable classification technique based on neural network called “sorting smoothing method”. The proposed technique is based on two variables, X and Y, where X is the independent variable, Y is the response variable, and the simple linear regression is shown as $Y = A + B X$.

An overview of different approaches in data mining evolution and development is presented in [27], focusing on the user interface aspect. Therefore, a comparison between open source data mining tools was presented in terms of the degree of relevance to biomedical datasets. To achieve the goal, several representations were applied on open source data mining software to show their similarities and dissimilarities in terms of the availability of representative model. At result, many of open source data mining software have been used in multiple studies with the increasing interest in data science and knowledge discovery.

A. ***Discussion of related work***

The works [4] [11] [12] and [18] have discussed the common characteristics of open source data mining tools mainly Weka, Tanagra, and KNIME. [11] applied the C4.5 algorithm (classification) to analyze the results of decision trees compared to other software. [12] considered the most functions of data mining such as clustering, regression, association, classification, and visualization. [18] discussed the accuracy of classification algorithms: decision tree, naive Bayes, decision stump, and k-nearest neighbor.

Moreover, [23] applied classification algorithms: decision tree, naïve Bayes, k-nearest neighbor, support vector machine, zero rule, and one rule. [5] introduced a comparison of twelve open source systems in terms of general specifications, data mining functions, usability, and data source formats. [6] showed how the pre-processing tasks can improve the performance of open source data mining tool when it integrated with agent based framework. [27] established the degree of relevance of different data mining software to biomedical datasets, while [25] assisted researchers in selecting the open source data mining tool suitable to the requirements belonging to the specified area or field.

II. REVIEW OF DATA MINING TOOLS

Several open source data mining tools are listed below as follows: WEKA, Orange, RapidMiner, Tanagra, and KNIME. These tools have general characteristics and descriptions that make them at the top of the most popular and largest packages. For this reason, the selection of these tools leads to choose this small sample of open source data mining tools in addition to the space limitations.

For more information, the readers could be directed to the wide spread website of the most common data mining tools called KDnuggets (http://www.kdnuggets.com/). Therefore, the dataset tested in the experiment of this research is taken from UCI Machine Learning Repository of datasets, namely default of credit card clients. In the next subsection, a list of detailed discussion of the five open source data mining tools is shown providing the direct link for downloading.

A. WEKA (http://www.cs.waikato.ac.nz/ml/weka/)

WEKA, a short of Waikato Environment for Knowledge Analysis, was developed at University of Waikato in New Zealand [3]. It is a suitable software for general data mining tasks and machine learning algorithms, and well-matched to develop new machine learning algorithms [4]. WEKA was released for free under the General Public License (GNU). Primarily, the first origin of WEKA was non-Java version developed for analysing agricultural data [25]. It has an active community since it is widely used software in academic and business [20]. WEKA is a java based software involves a set of machine learning packages. Therefore, WEKA provides developers with API and add-in packages. Consequently, WEKA is built into java language and this provides the support of .jar files that permit custom programming more than inside WEKA environment [25].

B. Orange (http://www.ailab.si/orange)

Orange is a python-based software suite for data mining and machine learning, developed under General Public License (GNU) in 2009 [19]. It is a powerful open source data mining tool for beginners and professionals [7]. Hence, it is a python scripting and visual programming tool that supports add-ons for text mining, bioinformatics and machine learning components [18]. Orange is a full package of data analytics, scripting environment, and visualization features. The scripts can be written as an extension of C++ and python. However, it does not support big data processing [25].

C. RapidMiner(https://my.rapidminer.com/nexus/account/index.html#downloads): formerlyknown YALE

RapidMiner, previously known as Yet Another Learning Environment (YALE), is a data mining software was developed by Ralf Klinkenberg, Ingo Mierswa, and Simon fischer at the technical University of Dortmund. In 2001, the first original version was known YALE [9]. In 2007, the company RapidMiner in Germany developed it and changed its name to RapidMiner under AGPL open source license [4]. RapidMiner supports all steps of data mining, which make it applicable for research, training, education, application development, rapid prototyping, and industrial and business applications [19]. RapidMiner is more than a powerful data mining software; it provides the ability to integrate its learning models, algorithms, and schemes with R and WEKA scripts [18]. There is also a licensed software product of

RapidMiner in addition to the open source community version. It provides data transformation, modelling, import and export data, connecting repository, and evaluating results [25].

D. Tanagra (http://eric.univ-lyon2.fr/~ricco/tanagra/en/contenu_telechargement_logiciel_tanagra.html)

Tanagra is the successor of the SPINA software, which is a classification program including interactive and visualization techniques with several supervised learning algorithms, and it is full featured software of several algorithms implementations [25]. Tanagra was developed in 2004 in France. The main purpose of the development of Tanagra was for research and teaching. It is an open source developed in Delphi. Conversely, it also used for profit activities and commercial use according to the license agreement. Additionally, Tanagra does not support big data processing [25]. Multiple data mining algorithms are supported by Tanagra including data analysis, statistical learning, database area, and machine leaning [18].

E. KNIME (https://www.knime.org/downloads/overview)

KNIME, a short of Konstant Information Miner, was primarily developed in 2004 (formally released in 2006 under GNU General Public License) by software engineer team at University of Konstanz in Germany for pharmaceutical research purpose [4]. Later, it has been used in other areas such as business intelligence, customer relationship management, financial data analysis and customer data analysis. KNIME provides an integration of data manipulation, reporting and integration platform, data mining algorithms, and visualization models and methods [19]. It was developed through open API modular of the Eclipse platform that enables developers to extend its functionalities. It also has commercial licence extensions for open source software that can be downloaded [25]. Multiple versions of KNIME were emerged capable with different free and open functionalities. KNIME big data extension can be purchased that support big data processing, which is not supported by open source [18].

III. A COMPARISON OF GENERAL CHARACTERISTICS

The differences between open source data mining tools have to be determined through analysing their parameters such as product track record, compatibility with different environments, vendor viability, scope

of data mining algorithms, usability, and import and export data files. Therefore, general characteristics of open source data mining tools should be recognized such as license, programming language, data source, and operating system. Thus, the supporting of various data formats like Excel, SQL, Access, ARFF, .CSV, ODBC, and MySQL is also very essential to choose a proper data mining software. Moreover, another significant feature of open source data mining tools is the functionality aspect, which is dedicated to data mining problems and solutions.

Further, describing the usability aspect shows how an open source data mining tool support human interaction, extensibility, and interoperability, and whether the user can easily use it in solving real world problems. Additionally, Graphical User Interface (GUI) describes how an open source data mining tool is easy to use, user-friendly, and have meaningful labels to handle users' actions. In the following subsection, a discussion of the five open source data mining tools is presented considering their specifications.

A. WEKA

WEKA is free, extensible java-based open source software that offers various valuable features and can run in different platforms. WEKA provides an access to SQL databases, and provides the user two options of usability: command line and interface [20]. One of the main limitations is the data format constraints that do not accept any format of data [4]. The first version of WEKA was released in 1997, and the current version is WEKA 3.6.11. WEKA is independent to the platform. It supports a wide range of data mining algorithms such as pre-processing, clustering, classification, and associations [21]. Three graphical user interfaces are supported by WEKA: explorer, experimenter, and knowledge flow. Therefore, it supports a simple command line [19]. The main areas supported by WEKA are: association rules. WEKA can import data with binary, ARFF, C4.5, and CSV file formats. WEKA cannot be connected to non-java based databases and excel, whereas it can be integrated to other java packages [19].

B. Orange

Orange is an open source data mining software that supports graphical user interface. The framework of Orange is cross-platform, meaning that it can run on different platforms [7]. The current version of Orange software is 2.7. It is compatible with C, C++, and python languages [19]. The most common data mining and

machine learning features are supported by Orange including pre-processing, feature filtering, data modelling, exploration methods, and model evaluation [18]. Orange is easy to use for beginners and professionals, and it contains add-ons for machine learning algorithms. Orange is also specialized for data visualization. On the other hand, Orange has limited capabilities of machine learning algorithms, data models representation and reporting functionalities [7]. Major standard file types such as excel and CSV files are supported by Orange since it can read filed in native tab delimited format. Orange involves data widgets that enable users to manipulate data through different processes including discretization and concatenation [18]. Orange support major classification techniques such as decision trees, K-nearest neighbour, C4.5, Naïve Bayes, support vector machine, and CART. Therefore, orange provides regression trees, linear regression, and logistic regression [25].

C. RapidMiner

RapidMiner is compatible with different data files such as excel, SPSS, MySQL, SQL server, Oracle, etc. it has an independent, cross-platform, and language independent [4]. The first version of RapidMiner was released in 2006, whereas the current version is RapidMiner 7.2. RapidMiner supports about 22 data input and output file formats [19]. RapidMiner is specialized for business solutions including pre-processing, predictive analysis, statistical computing, predictive analysis, and visualization [9]. RapidMiner is one of the top data mining tools [25].

D. Tanagra

Tanagra or Tanagra Komercinis pasilymas is a simple, easy to understand and use. It is free open source machine learning software that can be used for research due to its simple user interface. The current version of Tanagra software is 1.4.50, while the most common used versions of Tanagra re 1.4 and 2.0 [11]. Three windows are provided by Tanagra including data mining components, diagram, and outputs. Drag and drop feature makes it usable. Advanced improvements have been presented to Tanagra including hierarchical agglomerative clustering and computing time [18].

E. KNIME

KNIME is compatible with WEKA, and it contains an embedded use of R through statistical methods [2]. The first version of KNIME has been released in 2004 and the current version is KNIME 2.9. It is compatible with windows, OS X, and Linux. KNIME is a java-based software that enables users to visualize data flows, pipelines, and data models [18]. KNIME provides pre-processing and cleaning, analysis, data mining, and modelling [19]. In addition, it was dedicated for chemical structures, business intelligence, enterprise reporting, and data mining. It can integrate different data sources and databases including CSV, ARFF, SDF, XML, etc. KNIME is written in Java and can be extended to add functionalities and plugins on the go [25]. Advanced functionalities involve univariate statistics, multivariate statistics, time series analysis, data mining, web analytics, image processing, text mining, social media analysis, and network analysis [18].

Table 1 shows the summary the general specifications of the five open source data mining tools. The characteristics can be divided into: Developer, Release Date, Programming language, License, Current version, GUI /command line, Main purpose, Areas, Portability, Usability, Compatibility with database, Operating system Platform, and Visualization.

IV. A COMPARISON OF VISUALIZATION TECHNIQUES

In this section, a comparison study amongst different data visualization techniques involved in open source data mining tools is introduced. Some of data visualization techniques such as histograms, boxplot, scatter plot, parallel coordinates, etc. are investigated and compared between the five open source data mining tools, WEKA, Orange, RapidMiner, Tanagra, and KNIME. In the following subsection, a description of the dataset selected from UCI website is presented to show dataset schema.

A. Dataset Description

The dataset downloaded from the UCI repository is named “default of credit card clients”**.** It purposed at the case of customers’ default payments in Taiwan in order to compare different data mining methods such as classification. The goal of using this dataset in this research is to compare the visualization functionalities

among five data mining tools [26]. The dataset consists of 30000 instances and 24 attributes, including the class attribute. Table (2) summarizes the information of this dataset.

Table 1: Dataset description

Data Set Characteristics	Multivariate	**Number of Attributes**	24
Number of Instances	30000	**Date Donated**	2016-01-26
Area	Business	**Associated Tasks**	Classification
Attribute Characteristics	Integer, Real	**Missing Values**	N/A

B. Attributes Information

The first attribute is the class variable named default payment, which is a binary variable (yes=1, no=0). The other 23 attributes are as following: 1) X1: the individual consumer credit and family credit (amount of the given credit). 2) X2: gender (1=male, 2=female). 3) X3: education (1=graduate, 2=university, 3=high school, 4=others). 4) X4: marital status: (1=married, 2=single, 3=others). X5: Age in years. X6-X11: history of past monthly payments (from April to September, 2005) on am measurement scale for the repayment status: -1=pay duly, 1=delayed payment for 1 month, 2=delayed payment for 2 months, and so on. X12-X17: amount of bill statement (from September to April, 2005). X18-X23: amount of previous payment (from September to April, 2005).

Table 2: General information of data mining tools

Criteria	Weka	Orange	RapidMiner	Tanagra	KNIME
Developer	University of Waikato	University of Ljubljana	University of Dortmund	Developed in France	University of Konstanz
Release Date	1997	2009	2006	2004	2006
Programming language	Java	C++ & Python	Java	Java	Java
License	GNU	GNU	AGPL	SDL	GNU
Current version	Weka 3.6.11	Orange 2.7	RapidMiner 7.2	1.4.50	KNIME 2.9
GUI/CL	Both	GUI	Both	GUI	GUI
Main purpose	Academic & business	Business solutions	Business and industrial	Academic & research	Pharmaceutical research
Portability	Can be Integrated into other java packages	Interface with other java packages	Flexible input and output file formats	Limited portability	Easy integration of new algorithms
Usability	Easy to use	User-friendly interface	Usable for beginner and experts	Easy drag and drop	Usable, learnable, and interactive
Compatibility with database	ARFF,CSV, C4.5	CSV and Excel	Most file format types	CSV, excel, and ARFF	Most file format types

Operating system Platform Supporting	Independent	Cross the platform	Independent	Cross platform	Independent

This dataset aimed to compare the predictive accuracy of probability of default among six data mining methods at the case of customers; default payments in Taiwan. The estimated probability is predicted to be more useful than simple classification result (credible or no credible) from the risk management perspective. The real portability of default is unknown so the real portability of default needs novel sorting and smoothing method for estimation. The predictive portability of default is X (independent variable), while the real portability of default is Y (response variable). The forecasting model produced by artificial neural network shows a simple linear regression result (Y=A+BX), which has the highest coefficient determination. Based on the results of six data mining methods, the only one algorithm that can accurately estimate the real probability of default is artificial neural network [26].

Data visualization refers to the use of representation tools to show a dataset through figures, charts, diagrams, charts, etc. Different data mining tools accomplish data visualization to show the results into visualized charts [22]. In the next subsection, a comparison between five open source data mining tools is shown in terms of data visualization techniques. Again, the goal is to show the presence of several data visualization techniques in each tool.

A. WEKA

WEKA contains a set of data analytics, predictive modelling, and visualization tools composed with an easy to use graphical user interface [19]. It provides many visualization techniques such as 1D single attribute, scatterplot 2D, scatterplot 3D, rotate 3D visualization, ROC curves, tree visualizer (decision tree), graph visualizer (Bayesian networks), boundary visualizer, and allows plugins. The graphs can be visualized into different formats: XML, DOT, and BIF [21]. Therefore, visualization of errors, visualization of attributes,

and visualization into x and y representation are displayed in Appendix: Figure 1, Figure 2, and Figure 4 respectively.

However**,** data visualization in WEKA is somewhat limited. Thus, it is not good in interfacing with other software [4]. Moreover, sequence modelling (more powerful technique) is not supported by the WEKA data mining algorithms [3].

B. Orange

Orange is a featured data analysis, visual programming frontend, and data visualization tool [19]. Orange provides visualization widgets to perform several graphing techniques such as linear projection, bar graphs, and plotting. Other widgets include standard evaluation including confusion matrices and ROC curves. Orange environment includes data, visualize, classify, regression, evaluate, associate, and unsupervised widgets [7]. In addition, association rule mining also includes widgets and unsupervised learning capabilities such as principle component analysis and k-means clustering [25].

Orange contains an array of widgets to support many data visualization techniques. Therefore, workflow creation is supported by Orange through a linkage between different widgets and block [18]. It also supports the following data visualization techniques: scatter plot, box plot (see appendix Figure 4), distributions (see appendix Figure 5), trees, heatmaps (see appendix Figure 6), linear projection (see appendix Figure 7), mosaic display (see appendix Figure 8), bar charts, networks, Pythagorean forest, Pythagorean tree, dendrograms, scatter map (see appendix Figure 9), sieve diagram (see appendix Figure 10), CN2 rule viewer, and silhouette [16].

C. RapidMiner

RapidMiner is a fully featured data mining tool that composes data mining methods and data visualization techniques through process flow visualization and program control structures visualization [9]. Further, different data modeling techniques are supported such as neural networks, support vector machine, decision trees, Naïve Bayes, clustering, and logistic and linear regression. RapidMiner enables to create well implementation of workflows [25]. Thus, it supports the following data visualization techniques: histogram

(see appendix Figure 15), parallel coordinates, bars, bar stacked, SOM, survey, blocks, Pareto, distribution, web, Pie chart, pie chart 3D, ring, box, box 3D, surface 3D [18], andrwes curves (see appendix Figure 11), bubble (see appendix Figure 12) density (see appendix Figure 13), deviation (see appendix Figure 14), quantiles (see appendix Figures 16, 17, and 18), scatter 3D (see appendix Figures 19, 20, 21, and 22), series (see appendix Figures 23 and 24), and Sticks (see appendix Figures 25 and 36).

D. Tanagra

Tanagra enables data source reading, descriptive statistics, supervised learning, learning assessment, instance selection, feature selection, feature construction, association rules, regression, clustering, meta supervised learning, and factorial analysis [25]. Tanagra supports the following data visualization techniques: correlation scatter plot, multiple scatter plot, with the ability to change the axis (see Figures 27 and 28). Tanagra can show the dataset in grid view, which is read only. It also provides the ability to export dataset into text files. It creates workflows through adding tasks via diagram menu to processed orderly [25].

E. KNIME

KNIME provides advanced visualization techniques through familiar rich GUI. The traditional diagrams supported by KNIME are: line plots (see appendix Figure 32), box plot (see appendix Figure 29), conditional box plot, histograms (see appendix Figure 30 and 31), scatter plot (see appendix Figure 36), scatter matrices (see appendix Figure 35), parallel coordinates (see appendix Figure 33), radar plot, spark line appender, pie charts (see appendix Figure 34), lift chart, heatmaps, and bubble charts [19]. It provides the ability to export data visualization result through reporting functions to HTML, PNG, SVG, and PDF. It has a convenient platform which make it a suitable of statistical, data mining and visualization tool [25]. It supports workflows [4]. It has a good interfacing with other data mining packages for visualization [19].

Table 3: A comparison of visualization techniques

Tool	Visualization techniques	Workflow	Category
Weka	1D single attribute, scatterplot 2D, scatterplot 3D, rotate 3D visualization, ROC curves, decision tree, association rules, data clustering, graph visualizer (Bayesian networks), boundary visualizer, visualization of errors, visualization of attributes, and visualization into x and y.	No workflow	High support
Orange	linear projection, bar graphs, plotting, confusion matrices, ROC curves, scatter plot, box plot, distributions, trees, heatmaps, linear projection, mosaic display, bar charts, networks, Pythagorean forest, Pythagorean tree, dendrograms, scatter map, sieve diagram, CN2 rule viewer, and silhouette.	Support workflow	Medium support
RapidMiner	histogram, parallel coordinates, bars, bar stacked, SOM, survey, blocks, Pareto, distribution, web, Pie chart, pie chart 3D, ring, box, box 3D, surface 3D, andrwes curves, bubble, density, deviation, quantiles, scatter 3D, series, and Sticks.	Support workflow	High support
Tanagra	Correlation scatter plot, multiple scatter plot, with the ability to change the axis, and grid view.	Support workflow	Low support
KNIME	Line plots, box plot, conditional box plot,	Support	Medium support

	histograms, scatter plot, scatter matrices, parallel coordinates, radar plot, spark line appender, pie charts, lift chart, heatmaps, and bubble charts.	workflow	

V. CONCLUSION

Based up on the analysis, the main concluding remark can be gained from this research is that most investigated tools have excellent data visualization functionality with some differences. They are in general powerful data mining tools, useful for research, education, training, and academic. More specifically, most of visualization techniques are provided by these tools. We recommend researchers to use the high support visualization tools presented in this research considering the following points: 1) Data sources and data file formats support. 2) The provision of visualized workflows. 3) The provision of more relevant-domain data visualization techniques. 4) The support of decision support and business applications.

According to above analysis, the open source data mining tools provide the proper environment to researchers and developers for data integration, data mining functions development, data and idea exchange, algorithms and methods implementation. The final assumption revealed from this research claims that the best software amongst the investigated tools in terms of visualization features are WEKA and RapidMiner that categorized as high support of visualization features.

Acknowledgment (HEADING 5)

We thank Dr. Hasan Njadat to his assistance, and we highly appreciate his efforts. A big thank to our parents for their support.

References

[1] Achtert, E., Kriegel, H. P., Schubert, E., & Zimek, A. (2013, June). Interactive data mining with 3D-parallel-coordinate-trees. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 1009-1012). ACM.

[2] Alcalá-Fdez, J., Sanchez, L., Garcia, S., del Jesus, M. J., Ventura, S., Garrell, J. M., ... & Fernandez, J. C. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307-318.

[3] Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). WEKA Manual for Version 3-7-8. *Hamilton, New Zealand*.

[4] Chauhan, N., & Gautam, N. (2015). PARAMETRIC COMPARISON OF DATA MINING TOOLS. 2nd International conference on recent innovations in science, engineering, and management.

[5] Chen, X., Ye, Y., Williams, G., & Xu, X. (2007, May). A survey of open source data mining systems. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 3-14). Springer Berlin Heidelberg.

[6] Christa, S., Madhuri, K. L., & Suma, V. (2012). A Comparative Analysis of Data Mining Tools in Agent Based Systems. *arXiv preprint arXiv:1210.1040*.

[7] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., ... & Štajdohar, M. (2013). Orange: data mining toolbox in Python. *Journal of Machine Learning Research*, *14*(1), 2349-2353.

[8] Grinstein, G., & Thuraisingham, B. (1995, October). Data mining and data visualization: Position paper for the second IEEE workshop on database issues for data visualization. In *Workshop on Database Issues for Data Visualization* (pp. 54-56). Springer Berlin Heidelberg.

[9] Hofmann, M., & Klinkenberg, R. (Eds.). (2013). RapidMiner: Data mining use cases and business analytics applications. CRC Press.

[10] Jacob, S. G., & Ramani, R. G. (2012). Evolving efficient clustering and classification patterns in lymphography data through data mining techniques. *International Journal on Soft Computing*, *3*(3), 119.

[11] Jain, D. (2014). A Comparison of Data Mining Tools using the implementation of C4. 5 Algorithm. *International Journal of Science and Research Vol3*, (8).

[12] Jovic, A., Brkic, K., & Bogunovic, N. (2014, May). An overview of free software tools for general data mining. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on* (pp. 1112-1117). IEEE.

[13] Junaini, S. N., Bolhassan, N. A., Mat, A. R., & Huspi, S. H. (2006). *A Framework for Data Mining Visualization using Cluster Analysis: An Experimental Study on HIV-AIDS Datasets* (pp. 1-4). Universiti Malaysia Sarawak.

[14] Karthikeyani, V., & Begum, I. P. (2013). Comparison a performance of data mining algorithms (CPDMA) in prediction of diabetes disease. *International Journal on Computer Science and Engineering*, *5*(3), 205.

[15] Keramati, A., & Yousefi, N. (2011, January). A proposed classification of data mining techniques in credit scoring. In the Proceeding of 2011 International Conference of Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, Jurnal (pp. 22-4).

[16] Konjevoda, P., & Štambuk, N. (2012). Open-Source Tools for Data Mining in Social Science. *Theoretical and Methodological Approaches to Social Sciences and Knowledge Management*, 163-176.

[17] Lakshmi, K. R., Krishna, M. V., & Kumar, S. P. (2013). Performance comparison of data mining techniques for predicting of heart disease survivability. *International Journal of Scientific and Research Publications*, *3*(6), 1-10.

[18] Naik, A., & Samant, L. (2016). Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. *Procedia Computer Science*, *85*, 662-668.

[19] Rangra, K. and Bansa, L. (2014). Comparative Study of Data Mining Tools. Volume 4, Issue 6, June 2014 ISSN: 2277 128X. International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com.

[20] Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *2*(6), 250-253.

[21] Srivastava, S. (2014). Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining. *International Journal of Computer Applications*, *88*(10).

[22] Uhlmann, E., Geisert, C., Hohwieler, E., & Altmann, I. (2013). Data mining and visualization of diagnostic messages for condition monitoring. *Procedia CIRP*, *11*, 225-228.

[23] Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa, E. M. (2011). A comparison study between data mining tools over some classification methods. *IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, 18-26.

[24] Walton, J. (1996). Data mining and visualization. *Database programming & design*, *9*(5).

[25] Wimmer, H., & Powell, L. M. (2016). A Comparison of Open Source Tools for Data Science. *Journal of Information Systems Applied Research*, *9*(2), 4.

[26] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473-2480.

[27] Zupan, B., & Demsar, J. (2008). Open-source tools for data mining. *Clinics in laboratory medicine*, *28*(1), 37-54.

APPENDIX

A. WEKA visualization techniques:

- Visualization errors

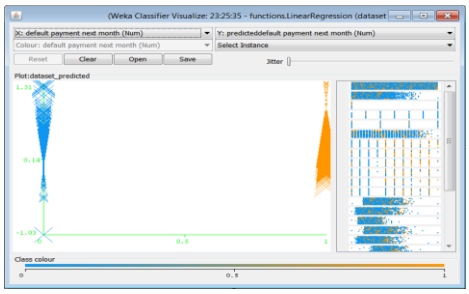

Figure 1: WEKA visualize errors

- Visualization attributes

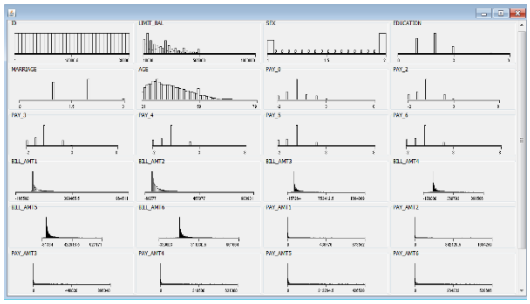

Figure 2: WEKA visualize attributes

- Visualization into x and y representation

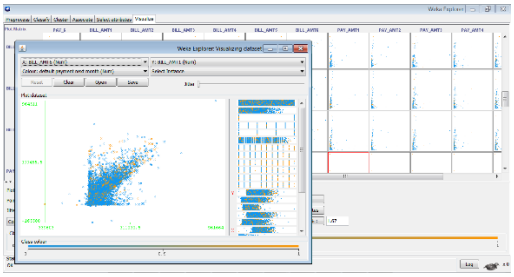

Figure 3: WEKA visualization (x and y)

B. Orange visualization techniques:

- Box plot

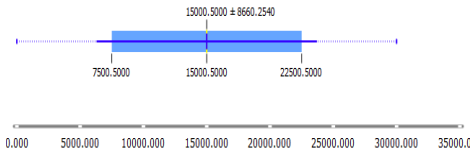

Figure 4: Orange box plot

- Distributions

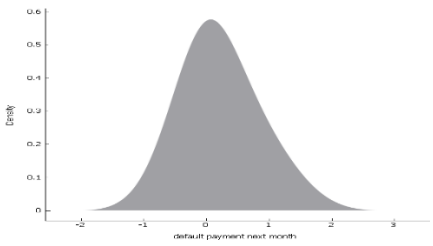

Figure 5: Orange distributions

- Heat map

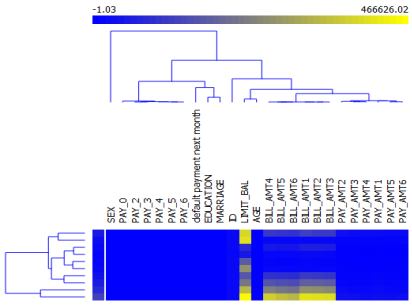

Figure 6: Orange heat map

- Linear projection

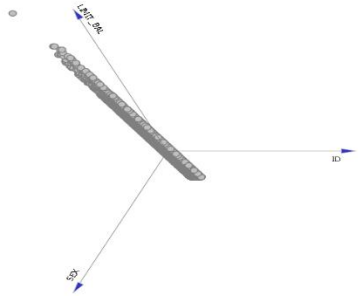

Figure 7: Orange linear projection

- Mosaic display

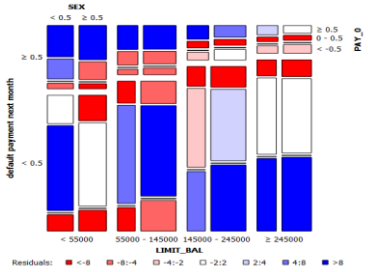

Figure 8: Orange mosaic display

- Scatter map

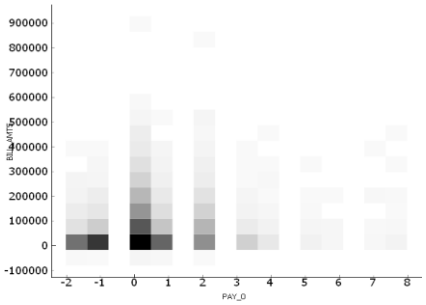

Figure 9: Orange scatter map

- Sieve diagram

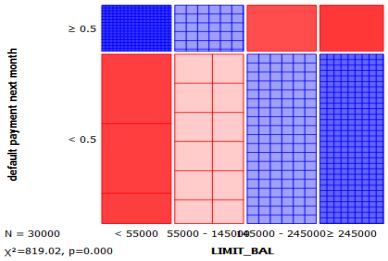

Figure 10: Orange sieve diagram

C. RapidMiner visualization techniques

- Andrwes curves

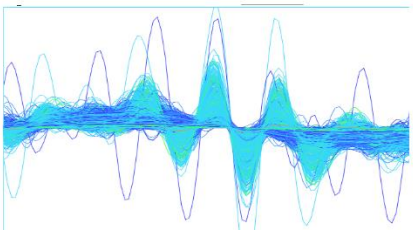

Figure 11: RapidMiner andrwes curves

- Bubble

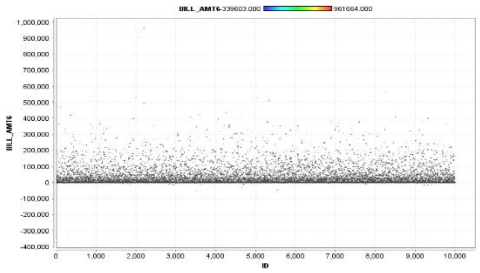

Figure 12: RapidMiner bubble

- Density

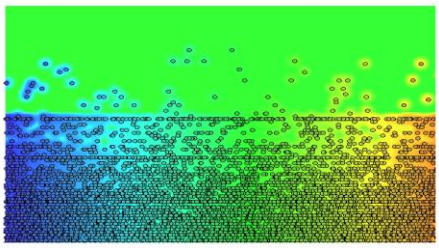

Figure 13: RapidMiner density

- Deviation

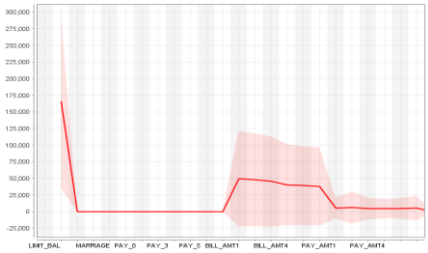

Figure 14: RapidMiner deviation

- Histograms

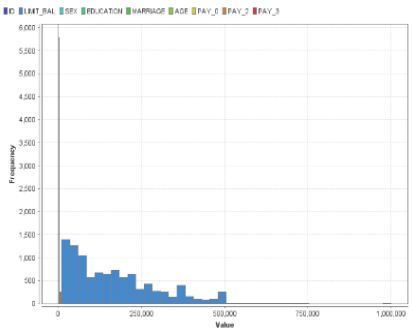

Figure 15: RapidMiner histograms

- Quantile colour matrix

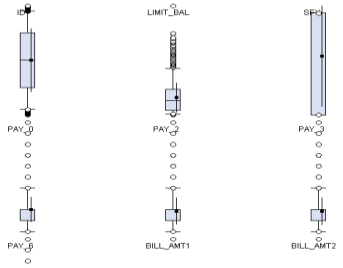

Figure 16: RapidMiner quantiles color marix

- Quantile colour

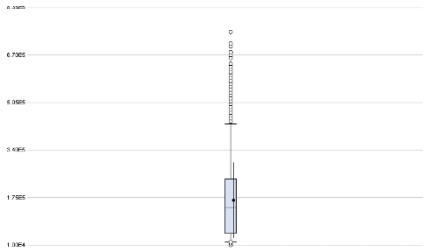

Figure 17: RapidMiner quantiles color

- Quantiles

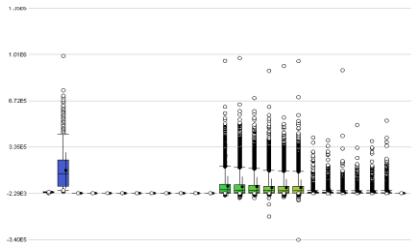

Figure 18: RapidMiner quantiles

- Scatter 3D colour

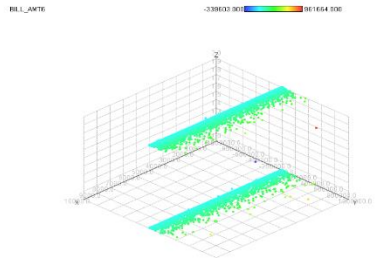

Figure 19: Rapidminer scatter 3D color

- Scatter 3D

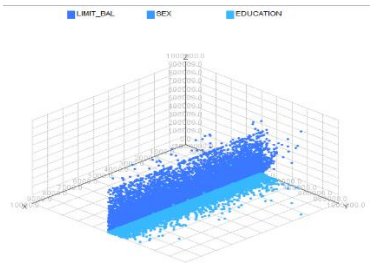

Figure 20: Rapidminer scatter 3D

- Scatter multiple

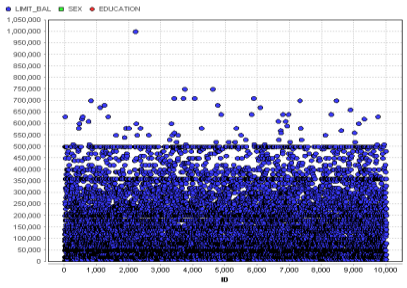

Figure 21: Rapidminer scatter multiple

- Scatter plot

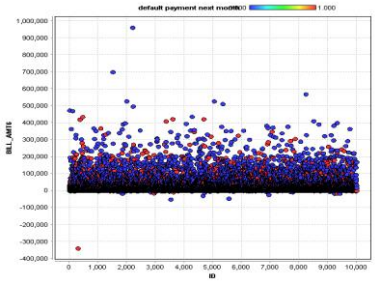

Figure 22: Rapidminer scatter plot

- Series multiple

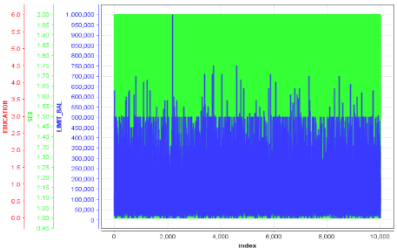

Figure 23: RapidMiner series multiple

- Series

Figure 24: RapidMiner series

- Sticks 3D

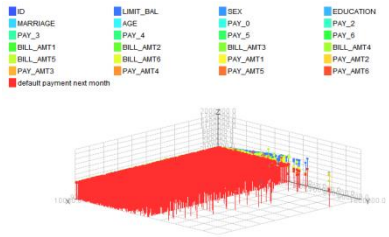

Figure 25: RapidMiner sticks 3D

- Sticks

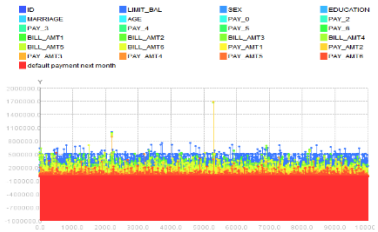

Figure 26: RapidMiner sticks

D. Tanagra visualization techniques

- Scatter plot

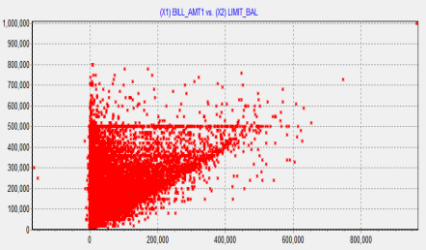

Figure 27: tanagra scatter plot

- Scatter plot with labels

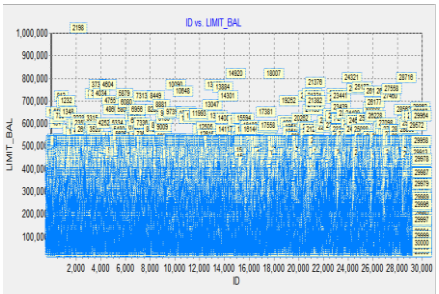

Figure 28: tanagra scatter plot with labels

E. KNIME visualization techniques

- Box plot

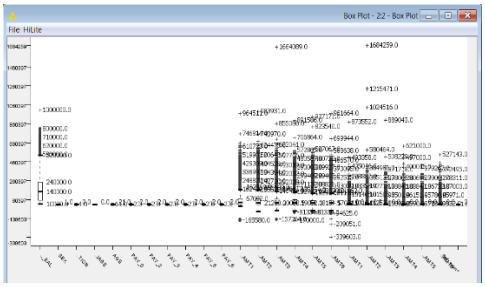

Figure 29: KNIME boxplot

- Histograms

Figure 30: KNIME histogram

- Interactive histograms

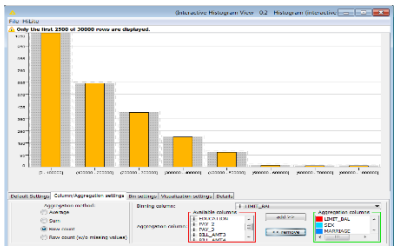

Figure 31: KNIME interactive histogram

- Line plot

Figure 32: KNIME line plot

- Parallel coordinates

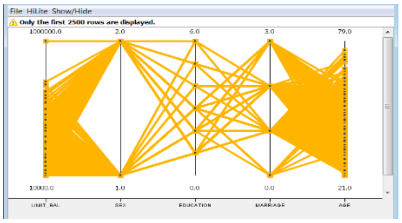

Figure 33: KNIME parallel coordinates

- Pie chart

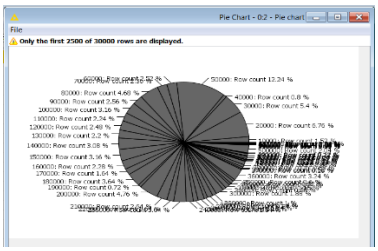

Figure 34: KNIME pie chart

- Scatter matrix

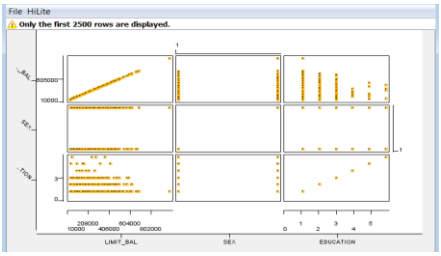

Figure 35: KNIME scatter matrix

- Scatter plot

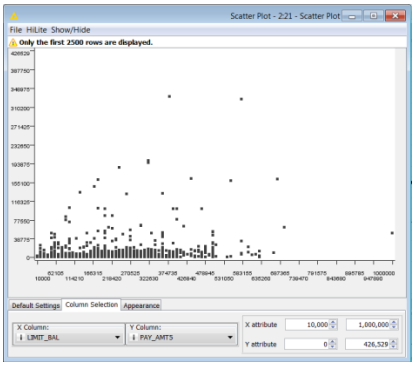

Figure 36: KNIME scatter plot