

Evaluating four of the most popular Open Source and Free Data Mining Tools

Ahmad Al-Khoder¹, Hazar Harmouch²

¹(Department of IT, Faculty of computer and information system / Islamic University of Medina, KSA)

²(Department of Software engineering and information Systems, Faculty of informatics Engineering/ AL-Baath University, Syria)

ABSTRACT The ability of DM⁽¹⁾ to provide predictive information derived from huge datasets became an effective tool for companies and individuals. Along with the increasing importance of this science, there was rapid increase in the number of free and open source tools developed to implement its concepts. It wouldn't be easy to decide which tool performs the desired task better, plus we cannot rely solely on description provided by the vendor. This paper aims to evaluate four of the most popular open source and free DM tools, namely: R, RapidMiner, WEKA and KNIME to help user, developer, and researcher in choosing his preferred tool in terms of platform in use, format of data to be mined and desired output format, needed data visualization form, performance, and the intent to develop unexciting functionality. As a result, All tools under study are modular, easy to extend, and can run on cross-platforms. R is the leading in terms of range of input/output formats, and visualization types, followed by RapidMiner, KNIME, and finally WEKA. Based on the results yielded it can be conducted that WEKA outperformed the highest accuracy level and subsequently the best performance.

KEYWORDS Data Mining, KNIME, R, RapidMiner, Tool, WEKA.

I. INTRODUCTION

Recent advances in the field of information technology have made usage of DM increasingly simple and affordable. However, this led towards availability of a large number of open source and free DM tools and still growing. The user of those tools may be specialist in the field of data mining or a beginner just needs a simple, easy to use tool. The platform in which they may be used generally consists of computers varying in operating systems and hardware connected via networks of different types (Internet, local or wireless). Furthermore, there will also be databases or files storing data which might be centralized or distributed using server-client model or distributed systems aspects. Resulting in choice to be made by the user, which tool to select from all the available open source and free tools to fit his needs? Clearly, each user looks for different factors to be available in his preferred tool. From the user point of view, his best tool will visualize data and apply desired DM task on user available data while running on user current platform efficiently. In addition, it may be popular which effect on the availability of support and solutions to the problems that may appear while using the tool. Also, if the user is advanced he may need to extend and add functionality to the tool. This paper aims to evaluate four most popular open source and free DM tools, namely: R, RapidMiner, WEKA and KNIME to help user, developer, and researcher in choosing his preferred tool evaluated DM tools in terms of platform in use, format of data to be mined and desired output format, needed data visualization form, performance, and the intent to develop unexciting functionality. The selection of these four most popular free

¹ DM: Data Mining.

and open source DM tools based on the results of the 15th annual poll (at 2014) on KD Nuggets(1) asking the voters to answer: “what analytics, Big Data, Data Mining, and Data Science software you used in the past 12 months for a real project?”. RapidMiner holds first place in the top ten data mining list followed by R, WEKA, and KNIME on the second, sixth, and seventh places respectively. Excel and SQL are excluded because they are not free. Python also excluded because WEKA and KNIME traded in the field of DM since a longer time. Rest of paper is organized as follows; section 2 introduces the concept of data mining in particular and DM techniques in brief. Section 3 provides evaluation of the tools under study done by other researchers according to different criteria. Later, section 4 presents the details of information about the tools analyzed / compared in the research study. Further, the comparative analysis of the selected DM tools is presented in section 5. Section 6 then concludes the research study.

II. DATA MINING

Data mining is a highly application-driven domain and it has incorporated technologies from various domains such as; statistics, machine learning, database, data warehouse systems, and information retrieval [1]. Data mining is a process of discovering useful or actionable knowledge from large-scale raw data, also known as knowledge discovery from data (KDD). KDD typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation [2]. The given datasets provided as input to DM algorithms are divided into two kinds: training datasets and testing datasets with known class labels. The models are built from the training data to be used for prediction. Data mining algorithms are broadly classified into: supervised (Classification), unsupervised (Clustering), and semi-supervised learning algorithms (Co-training) [1]. In supervised learning algorithms, the class label of each training tuple is provided to the algorithm to learn from. Evaluation of the trained model is then obtained by applying the model to available test dataset. Typical supervised learning methods include decision tree induction, naive Bayes classification, support vector machines etc. Unsupervised learning algorithms are designed for data in which the class label of each training tuple is unknown, and the number or set of classes to be learned is not known in advance. So the models are built based on similarity or dissimilarity between data objects using proximity measures including Euclidean distance, Jaccard coefficient, cosine similarity, Pearson's correlation etc. Typical examples of unsupervised learning include K-means, hierarchical clustering, density-based clustering etc. When large amount of unlabeled data exist, the Semi-supervised learning algorithms are more applicable. The semi-supervised classification builds a classifier using both labeled and unlabeled data. Examples of semi-supervised classification include self-training and co-training. The semi-supervised clustering uses labeled data to guide clustering [2].

III. RELATED WORKS

[3] describe the technical specification, features, and specialization for six open source data mining tools: WEKA, KEEL, R, KNIME, RapidMiner, and Orange along with their advantages and limitations. They recommend KNIME for people who are novices and consider WEKA a very close second to KNIME because of its many built-in features that require no programming or coding knowledge. On contrast, they consider that Rapid Miner and Orange appropriate for advanced users because of the additional programming skills that are needed, and the limited visualization support that is provided. The study concluded Rapid miner is the only tool which is independent of language limitation and has statistical and predictive analysis capabilities, so it can be easily used and implemented on any system; moreover it integrates maximum algorithms of other mentioned tools.

[4] provide the interested researchers with multidimensional overview of pros and cons for six free data mining tools: RapidMiner, R, WEKA, KNIME, Orange, and scikit-learn. They compare the tools according to

¹ <http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-software-used.html> (Retrieved 27, Sep, 2014).





the implemented algorithms, specialized topics like big data, text mining, etc. and finally with respect to the community support. They conclude that RapidMiner, R, WEKA, and KNIME have most of the desired characteristics for a fully-functional data mining platform and therefore their use can be recommended for most of the data mining tasks.

[5] Have conducted a performance comparative study between four of the free available DM tools and software packages: WEKA, Tanagra, KNIME, and Orange. Results have shown that performance of the tools for the classification task is affected by the kind of dataset used and by the way the classification algorithms were implemented within the toolkits. For the applicability matter, the WEKA toolkit has achieved the highest applicability followed by Orange, Tanagra, and KNIME respectively. Finally; WEKA toolkit has achieved the highest improvement in classification performance; when moving from the percentage split test mode to the Cross Validation test mode, followed by Orange, KNIME and finally Tanagra respectively.[3] evaluate R, RapidMiner, WEKA, and KNIME in terms of (1)features and, (2)Advantages and limitation. [4] compare them according to (3) implemented algorithms, (4)specialized topics in DM ,and (5) community support. In this paper we will add additional comparison criteria to evaluate the studied tools namely, (6) Input/output formats, (7) Platforms, (8) Structure and development, (9) Visualization. And we will also extend the WEKA and KNIME (10) performance evaluation of [5] to involve R and RapidMiner by attending their same methodology. A cumulative review of our work and the mentioned researches will provide the user with ten factors evolution of R, RapidMiner, WEKA, and KNIME which make choice and selection of one tool easy.

IV. Tools Description

Table 1 briefly summarizes the important information about the tools analyzed in this study. Data in Tables (1) to (4) are from R⁽¹⁾⁽²⁾ [6], [7] , RapidMiner⁽³⁾⁽⁴⁾⁽⁵⁾, WEKA⁽⁶⁾⁽⁷⁾⁽⁸⁾ [8], and KNIME⁽⁹⁾.

Table (1): Open source and free DM Tools analyzed in this study

TOOL	R	RapidMiner	WEKA	KNIME
Logo				
Description	software programming language and software environment for statistical computing and graphics.	software platform provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics	popular suite of machine learning software	Konstanz Information Miner, is an open source data analytics, reporting and integration platform.
Launch Date	1997	2001	2002	2006
Current Version	3.1.1 10-07-2014	6 02-05-2014	3.7.11 24-04-2014	2.10 10-08-2014
Development Team	R Foundation	Rapid-I company	University of Waikato	KNIME.com AG

¹ <http://rforge.net/Rserve/> (Retrieved 27, Sep, 2014).

² <http://cran.r-project.org/doc/manuals/r-release/R-data.html> (Retrieved 27, Sep, 2014).

³ http://www.rapid-i.com/downloads/brochures/RapidMiner_Fact_Sheet.pdf (Retrieved 27, Sep, 2014).

⁴ https://rapidminer.com/Fwp-content/Fuploads/F2013/F10/FRapidMiner_OperatorReference_en.pdf (Retrieved 27, Sep, 2014).

⁵ <https://rapidminer.com/> (Retrieved 27, Sep, 2014).

⁶ <http://www.cs.waikato.ac.nz/~ml/WEK> (Retrieved 27, Sep, 2014).

⁷ <http://wiki.pentaho.com/display/DATAMINING/Pentaho+Data+Mining+Community+Documentation> (Retrieved 27, Sep, 2014)

⁸ <http://edouard-lopez.com/fac/ICPS%20-%20S7/Data%20Mining/ComparingWekaAndR.pdf> (Last Modified on December 17, 2007).

⁹ <http://www.knime.org> (Retrieved 27, Sep, 2014).

Price	Open Source	Free Community Edition Commercial Enterprise Edition	Open Source	Open Source Commercial Extensions
License	GNU General Public License	AGPL (Community Edition) Closed (Enterprise Edition)	GNU General Public License	GNU General Public License
Programing Language	R interpreted language	JAVA	JAVA	JAVA

V. COMPARATIVE STUDY

Selected open source and free DM tools in this study are evaluated using following measures;

- Input/output formats: file formats that can be imported and exported by each tool.
- Performance of the selected DM tools tested over some classification algorithms.
- Platforms: Software includes operating system and middleware .Hardware includes: architectures, multi-core, distributed computing, and Client-Server.
- Structure and development: available options to develop and extend the tool functionality according to its software design and structure.
- Visualization: plots used by the tool to help understand the data and results better.

1. Platform

All four tools can run under the popular Operating systems: windows, Mac, UNIX and Linux. While R needs no additional prerequisites whereas, all other tools need Java Run Time environment (JRE) to be installed, prior to the use, each tool require specific version of JRE. All tools show ability to run on 32 and 64 bit machines, with multi-core support. Distributed experiments are supported and server versions of all the tools are also available.

Table (2): Comparison based on Platform

	R	Rapid Miner	WEKA	KNIME
windows	✓	✓	✓	XP to Windows 8
Mac	Support MacClassic ended with R 1.7.1	✓	✓	Mac OS X
Unix/Linux	✓	✓	✓	Linux(SUSE, Salaries, RedHat, Ubuntu)
Min. JRE	X	1.5	1.3	1.1
X86-x64	✓	✓	✓	x32 on XP , Vista x64 on Vista ,Win 7
Multi-Cores	started with 2.14.0	Since 4.3 Enterprise Edition	✓	✓
Distributed Computing	✓	✓	✓	✓
Client- server	✓	✓	✓	✓

2. Input / output Formats:

The DM tools analyzed in this study accept different format of the data to be given as input for training and test. Quoting the same reason of accepting different format of data, makes it difficult for users to share the same data file between two different DM tools. This inspired the DM tool developers to come up with exporting

the data files into various desired formats according to the needs of the users. Details of data file format for import and export is provided in Table 3.

Table (3): Comparison based on supported input/ output files format

format	R		RapidMiner		WEKA		KNIME	
	in	out	in	out	in	out	in	out
text file(ASCII,.dat)	✓	✓	✓	✓	✓	✓	✓	✓
Binary Files	✓	✓	X	X	X	X	✓	✓
Excel spreadsheet and ODS(.csv,.delim,.DIF)	✓	✓	.csv	.csv	✓	✓	X	✓
Network Connection(Socket)	✓	✓	Webpages RSSfeeds, web services	Web based reports	X	X	✓	X
SPSS	✓	✓	X	✓	X	X	X	X
SAS(.ssd or.sas7bdat)	✓	✓	JDB	✓	X	X	X	X
Stata	✓	✓	X	X	X	X	X	X
EpiInfo(.REC)	✓	✓	X	X	X	X	X	X
Minitab	✓	✓	X	X	X	X	X	X
S-PLUS	✓	✓	X	X	X	X	X	X
Systat(.sys ,.syd)	✓	✓	X	X	X	X	X	X
Octave	✓	✓	X	X	X	X	X	X
DBMSs	✓	✓	✓	✓	✓	✓	✓	✓
ODBC(.dbf,.xls)	✓	✓	JDBC	JDBC	JDBC	JDBC	JDBC	JDBC
DBF	✓	✓	X	X	X	X	X	X
Xml	✓	✓	✓	✓	X	X	✓	✓
SAP	X	X	✓	✓	X	X	X	X
Pdf, html	✓	✓	✓	✓	X	X	✓	X
Audio	X	X	✓	✓	X	X	X	X
WEKA	✓	✓	✓	✓	✓	✓	✓	✓
images	✓	✓	✓	✓	X	X	✓	✓

3. Visualization:

The best way to manage modern huge datasets is to use Visualization and interact with data through visual drill-down capabilities and dashboards. Data visualizations allow users to gain insight into the data and come up with new hypotheses.

Table (4): Comparison based visualization technique

Plot	R	RapidMiner	WEKA	KNIME
Bar chart	✓	✓	✓	✓
Line	✓	✓	✓	✓
Bubble	X	✓	X	✓
Deviation	X	✓	X	X
Density	X	✓	X	✓
Survey plots	X	✓	X	X
Pie chart	✓	✓	X	✓
Histogram	✓	✓	✓	✓
Box	✓	✓	X	✓

Scatter	✓	X	✓	✓
Cleveland dot	✓	X	X	✓
QQ (quantile-quantile)	✓	X	X	X
Parallel	✓	✓	X	✓
Conditioning plot	✓	X	X	✓
Scatterplot matrix	✓	✓	✓	✓
Kernel density	✓	X	X	X
Contour	✓	✓	X	X
Association	✓	X	X	X
Mosaic	✓	X	X	X
Perspective	✓	X	X	X
Surfaces	✓	✓	X	X
3d scatter plots	✓	✓	X	X
Two-way interaction	✓	X	X	X
Google Visualisation API	✓	X	X	X
Maps	✓	✓	X	X
Andrews curves	X	✓	X	X
Quartile	X	✓	X	X

4. Performance

Performance of the selected DM tools is carried out by comparing the accuracy of the Naive base, Decision tree, and K-NN classification models built by each tool against a group of datasets varies in their area, number of instances, attributes, and class labels. The classifiers accuracy compared between two test modes i.e. 10-FCV and hold-out (66% training, 34% testing) to ensure the evaluation.

This methodology is detailed in [5] because we extend their comparison to involve R and RapidMiner with the following modification:

- As [5] stated that performance of the tools they studied affected by the way the classification algorithms were implemented, we re-calculate the accuracy of KNIME and WEKA on the current software versions and on same test machine to normalize the numbers over the four tools.
- We limit our test to the following three classification algorithms: Naïve Bayes (NB) [9] [10], Decision Tree Classifier (DT) [11], and the K Nearest Neighbor (K-NN) [12] [13] [10].
- In order to test the tools on dataset from multiple disciplines, we kept datasets ([Spambase](#)(SB):computer, [Breast Cancer Wisconsin](#)(BC):life, [Car Evaluation](#) (CE): industrial, [Nursery \(N\)](#):Social) and add two additional dataset ([Wine](#) (W): Physical and [Bank Marketing](#) (BM): Business). All Datasets are selected from UCI-repository⁽¹⁾. The selected datasets vary in the area, number of instances that ranged from 178 to 45211, and the number of attributes that varied between 6 and 57, as some of them has binary class label while other has multiple class labels.

The experiment was conducted on a laptop computer with:

- Hardware: Intel core i7 processor 2GH 64-bit with 8 GB of RAM.
- OS: Windows 7 Home premium 64-bit.
- RapidMiner studio 6 x64.
- Rstudio 0.98.1028 and R x64 3.1.1.
- WEKA 3.6.11 x64.
- KNIME Analytics Platform and KNIME SDK 2.10.1 x64.
- All tools used with their default setting with no pre-processing stage.

¹ <http://archive.ics.uci.edu/ml/> (Retrieved 27, Sep, 2014).

Tables from 5 to 8 shows results yielded using the performance evaluation methodology and also the accuracy measure ranges with the indexation of improvement as a result of using 10-FCV.

Table (5): Results yielded using RapidMiner.

BM	W	N	CE	BC	SB	Dataset/technique	
82.35	100.0	76.47	76.47	97.06	97.06	Hold-out	NB
▲87.64	▼98.30	▲90.25	▲85.99	▼95.84	▼79.68	10-FCV	
79.41	94.12	76.47	73.53	97.06	76.47	Hold-out	DT
▲88.30	▼91.08	▲97.12	▲92.42	▼94.27	▲90.74	10-FCV	
76.65	64.71	67.65	79.41	55.88	50.00	Hold-out	K-NN
▲88.79	▲68.01	▲85.72	▼76.04	▲65.52	▲74.09	10-FCV	
▲ accuracy increased using 10-FCV ▼ accuracy decreased using 10-FCV ■ Min. Accuracy ■ Max. accuracy							

Table (6): Results yielded using R.

BM	W	N	CE	BC	SB	Dataset/technique	
87.90	098.4	90.00	85.10	94.10	98.01	Hold-out	NB
0▼87.8	0▼97.7	▲90.30	▲85.70	▲96.60	▼81.25	10-FCV	
90.30	090.1	97.30	94.80	95.80	90.90	Hold-out	DT
▼90.20	0▼89.2	▲98.10	▼94.30	▼95.60	▲92.10	10-FCV	
89.10	80.10	75.90	91.05	81.09	80.2	Hold-out	K-NN
—89.10	▼74.90	▲86.90	▼77.50	▼70.00	▲80.50	10-FCV	
▲ accuracy increased using 10-FCV ▼ accuracy decreased using 10-FCV ■ Min. Accuracy ■ Max. accuracy							

Table (7): Results yielded using WEKA.

BM	W	N	CE	BC	SB	Dataset/technique	
88.04	98.36	90.67	87.58	96.63	78.00	Hold-out	NB
▼88.00	▼96.62	▼90.32	▼85.53	▲97.42	▲79.28	10-FCV	
88.94	98.36	95.64	95.40	94.95	93.15	Hold-out	DT
▲89.22	▲98.87	▲96.25	▼95.08	▼94.56	▲93.24	10-FCV	
86.41	95.08	97.52	90.64	95.79	89.00	Hold-out	K-NN
▲86.96	▼94.94	▲98.37	▲93.51	▼94.84	▲90.76	10-FCV	
▲ accuracy increased using 10-FCV ▼ accuracy decreased using 10-FCV ■ Min. Accuracy ■ Max. accuracy							

Table (8): Results yielded using KNIME.

BM	W	N	CE	BC	SB	Dataset/technique	
86.05	98.36	90.08	84.69	94.95	89.07	Hold-out	NB
▼85.01	▼96.06	▲90.24	▲86.57	▼94.70	▲89.91	10-FCV	
90.01	86.88	94.39	78.23	94.53	90.47	Hold-out	DT
▼88.91	▲89.32	▲94.78	▲81.59	▼93.41	▲91.58	10-FCV	
88.65	62.29	34.94	78.28	68.53	76.10	Hold-out	K-NN
▼88.60	▲65.73	▼34.35	▼77.54	▼65.73	▲78.85	10-FCV	
▲ accuracy increased using 10-FCV ▼ accuracy decreased using 10-FCV ■ Min. Accuracy ■ Max. accuracy							

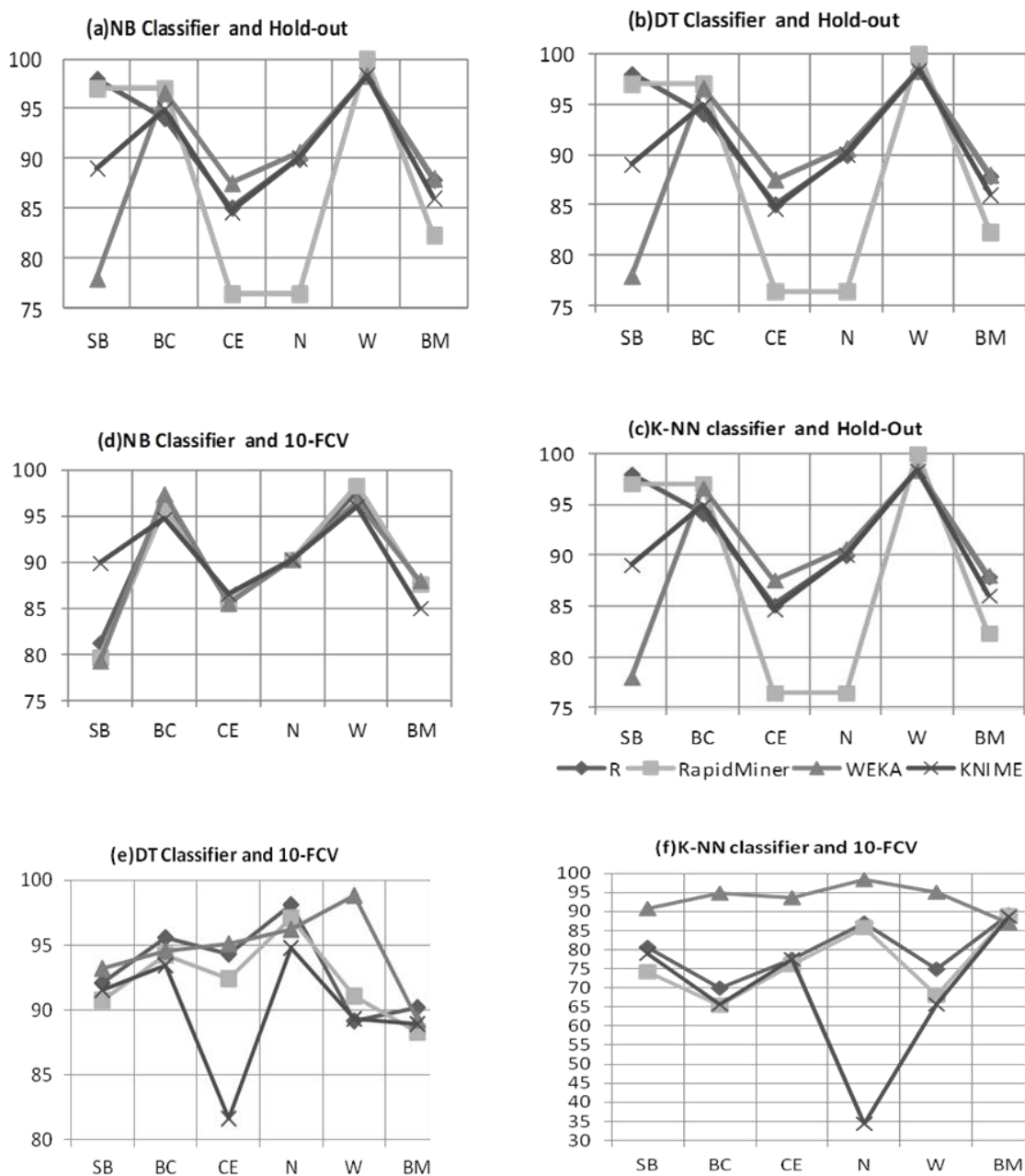


Figure (1): Performance Evaluation.

Fig. 1 illustrates the Hold-out and 10-FCV test results of naive bias, decision tree, and K-NN classifiers of the studied DM tools on six different datasets where SB, BC, and BM have binary class labels while CE, N, and W have multi class labels. Each of SB and W datasets are continuous, in contrast with BC, CE, N, and BM are categorical. SB and BC have a missing values.

Table (9): Evaluation Summary.

Tool with best accuracy							
BM	W	N	CE	BC	SB	Dataset/technique	
WEKA (88.04)	RapidMiner (100)	WEKA (90.67)	WEKA (87.58)	RapidMiner (97.06)	R (98.01)	Hold-out	NB
WEKA (88.00)	RapidMiner (98.30)	WEKA (90.32)	KNIME (86.57)	WEKA (97.42)	KNIME (89.91)	10-FCV	
R (90.30)	WEKA (98.36)	R (97.30)	WEKA (95.40)	RapidMiner (97.06)	WEKA (93.15)	Hold-out	DT
R (90.20)	WEKA (98.87)	R (98.10)	WEKA (95.08)	R (95.60)	WEKA (93.24)	10-FCV	
R (89.10)	WEKA (95.08)	WEKA (97.52)	R (91.05)	WEKA (95.79)	WEKA (89.00)	Hold-out	K-NN
R (89.10)	WEKA (94.94)	WEKA (98.37)	WEKA (93.51)	WEKA (94.84)	WEKA (90.76)	10-FCV	
Hold-Out: WEKA (10), R (5), RapidMiner (3), KNIME (0).							
10-FCV: WEKA (11), R (4), KNIME (2), RapidMiner (1).							
Tool best deals with multi-class data (CE, N, and W)	Tool best deals with missing values (SB and BC)	Tool best deals with continues data (SB and W)	Tool with best accuracy on average		Dataset/technique		
WEKA	RapidMiner	R	WEKA	WEKA	Hold-out	NB	
RapidMiner	WEKA	RapidMiner	WEKA	WEKA	10-FCV		
WEKA	R	RapidMiner	WEKA	WEKA	Hold-out	DT	
WEKA	R	WEKA	WEKA	WEKA	10-FCV		
WEKA	R	RapidMiner	WEKA	WEKA	Hold-out	K-NN	
WEKA	WEKA	WEKA	WEKA	WEKA	10-FCV		

- For all classifier and test modes, WEKA gives higher accuracy compared to other tools.
- R gives the best accuracy levels with datasets contains missing values.
- RapidMiner is the best to handle continues data type.
- WEKA is the best to classify Multi class labels datasets.
- Counting the cases where each tool achieved the best result among the others on same dataset and classifier results the order from lager to smaller number of cases:
 - WEKA, R, RapidMiner, and KNIME in hold-out test mode.
 - WEKA, R, KNIME, and RapidMiner in 10-FCV test mode.

5. Structure and Development

Most of the R functionality is provided through built-in and user-created functions, and all data objects are kept in memory during an interactive session [6]. R can be extended by creating R packages, writing R documentation files, tidying and profiling R code, interface functions .C and .Fortran, and adding new generics⁽¹⁾. RapidMiner structure consists of two types of operators. Normal operator which contains one or more sub processes. Super operator relies on the execution of other operators or should be user defined. RapidMiner can be extended, either by using the built-in scripting operator to write a quick hack, or by building an extension providing new operators and new data objects with all the functionality of RapidMiner⁽²⁾. WEKA framework is a set basically in one big plugin framework. With WEKA's automatic discovery of classes on the class-path, adding new features that may be either additional machine learning algorithms and tools for data

¹ <http://cran.r-project.org/doc/manuals/R-exts.pdf> (Retrieved 27, Sep, 2014).

² <http://rapidminer.com/wp-content/uploads/2013/10/How-to-Extend-RapidMiner-5.pdf> (Retrieved 27, Sep, 2014).

visualization, or even extensions of the Graphical User Interface (GUI) in order to support different workflows. You basically choose a superclass to derive your new algorithm from and then implement additional interfaces, if necessary [8]. The architecture of KNIME is designed with three main objectives i.e. visual, interactive framework, modularity, and expandability. In order to achieve the objectives the data analysis process consists of a pipeline of nodes, connected by edges that transport either data or models. Each node processes the arriving data and/or model(s) and produces results on its outputs when requested. KNIME already includes plug-ins to incorporate existing data analysis tools. It is usually straightforward to create wrappers for external tools without having to modify these executables themselves. One needs to extend abstract classes to add new nodes to KNIME for native new operations. A wizard integrated in the Eclipse-based development environment enables convenient generation of all required class bodies for a new node [14].

VI. CONCLUSION AND FUTURE WORK

This research conducts a comparison between four DM tools against five criteria namely: platform, input/output formats, visualization, popularity, structure and development and finally performance. Six different datasets were used to evaluate the performance of three classification algorithms namely; Naïve Bayes (NB), Decision Tree (DT), and K Nearest Neighbor (KNN). R seems to support wider range of input/output formats, and visualization types. Nevertheless, R is holding the second place in popularity behind RapidMiner, which is leading the market. In terms of classifiers' applicability, we conclude that WEKA was the best tool to run the selected classifiers followed by R, RapidMiner, and finally KNIME respectively. All the tools have modular structure and extendibility nature. As a future research, we are planning to test the selected DM tools for other machine learning tasks, such as clustering, using test datasets designed for such tasks and known clustering and association algorithms.

REFERENCES

- [1] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., San Francisco: Morgan Kaufmann, 2011.
- [2] P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Boston: Pearson Addison Wesley, 2006.
- [3] K. Rangra and K. L. Bansal, "Comparative Study of Data Mining Tools," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 6, JUNE 2014.
- [4] a. jovic, k. brkic and n. bogunovic, "An overview of free software tools for general data mining," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, Opatija, 2014.
- [5] A. Wahbeh, Q. Al-Radaideh, M. Al-Kabi and E. Al-Shawakfa, "A Comparison Study between Data Mining Tools over some Classification Methods," *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Artificial Intelligence*, vol. 2, no. 8, pp. 18-26, 2011.
- [6] R. KABACOFF, *R in Action Data Analysis and Graphics with R*, Manning Publications, 2011, p. 472 p.
- [7] R. Ihaka, "R: Past and Future History," in *the 30th Symposium on the Interface*, S. Weisberg Ed., 1998.
- [8] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewal and D. Scuse, *WEKA Manual for Version 3-6-10*, Hamilton, New Zealand: University of Waikato, 2013.
- [9] A. HEB, P. DOPICHAJ and C. MAAB, "Multi-value Classification of Very Short Texts," in *the 31st annual German conference on Advances in Artificial Intelligence*, 2008.
- [10] S. Zhou, T. W. Ling, J. Guan, J. Hu and A. Zhou, "Fast Text Classification: A Training-Corpus Pruning Based Approach," in *Eighth International Conference on Database Systems for Advanced Applications*, Kyoto, Japan, 2003.

- [11] Q. Al-Radaideh, "The Impact of Classification Evaluation Methods on Rough Sets Based Classifiers," in *the 2008 International Arab Conference on Information Technology*, 2008.
- [12] Y. Li and K. Bontcheva, "adapting Support Vector Machines for F-term-based Classification of Patents," *Journal ACM Transactions on Asian Language Information Processing*, vol. 7, no. 2, 2008.
- [13] A. Pathak, M. Sehgal and D. Christopher, "A Study on Selective Data Mining Algorithms," *IJCSI International Journal of Computer Science*, vol. 8, no. 2, 2011.
- [14] M. Berthold, N. Cebron, F. Dill, T. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel and B. Wiswedel, "KNIME: The Konstanz Information Miner, Data Analysis, Machine Learning and Applications," in *31st Annual Conference of the Gesellschaft für Klassifikation*, Albert-Ludwigs-Universität Freiburg, New York., 2007.