An overview of free software tools for general data mining

A. Jović*, K. Brkić*[1] and N. Bogunović*
* Faculty of Electrical Engineering and Computing, University of Zagreb / Department of Electronics, Microelectronics, Computer and Intelligent Systems, Unska 3, 10 000 Zagreb, Croatia
{alan.jovic, karla.brkic, nikola.bogunovic}@fer.hr

Abstract - This expert paper describes the characteristics of six most used free software tools for general data mining that are available today: RapidMiner, R, Weka, KNIME, Orange, and scikit-learn. The goal is to provide the interested researcher with all the important pros and cons regarding the use of a particular tool. A comparison of the implemented algorithms covering all areas of data mining (classification, regression, clustering, associative rules, feature selection, evaluation criteria, visualization, etc.) is provided. In addition, the tools' support for the more advanced and specialized research topics (big data, data streams, text mining, etc.) is outlined, where applicable. The tools are also compared with respect to the community support, based on the available sources. This multidimensional overview in the form of expert paper on data mining tools emphasizes the quality of RapidMiner, R, Weka, and KNIME platforms, but also acknowledges the significant advancements made in the other tools.

I. INTRODUCTION

Data mining (DM) is the core step in knowledge discovery in datasets. It integrates all the analysis procedures that are required in order to reveal new and relevant information to an interested user. DM includes data preparation and data modeling. Datasets may be obtained from a variety of sources, including: traditional relational databases, data warehouses, web documents, or simple local textual files. It is important to prepare data in the most efficient way in order to extract as much information as possible [1]. After preparation, various models can be constructed, depending on the research goal. In order to properly interpret the models, standard evaluation and statistical procedures are pursued. The visualization of the results is also encouraged.

Free and publicly available software tools for DM have been in development for the past 20 years. The goal of these tools is to facilitate the rather complicated data analysis process and to offer all interested researchers a free alternative to commercial data analysis platforms. They do so mainly by proposing integrated environments or specialized packages on top of standard programming languages, which are often open source.

This paper focuses on the review of several such tools that have grown more efficient and useful over the years, some even comparable or better in certain aspects than their commercial counterparts. In particular, the main characteristics of RapidMiner [2], R [3], Weka [4], Orange [5], KNIME [6], and scikit-learn [7] will be outlined and compared. All of these free tools have implementations of general DM tasks. Other, more specialized tools such as Elki (clustering) or Anatella (big data), and tools with small DM community support (e.g. F#, GNU Octave) are not considered here.

The tools are compared in terms of general properties (language, license, etc.), core DM tasks (supported input, data preparation, data analysis, output), and support for some recent more advanced DM topics (big data, data streams, text mining, deep learning).

II. FREE DATA MINING TOOLS OVERVIEW

In a recent, 2013, poll published on the influential KDnuggets portal [8], regarding the use of DM tools in a real project, it is interesting to observe that in the top 5 tools there is only one commercial tool: Excel. The domination of free tools, mainly RapidMiner and R, probably stems from the maturity and availability of a large number of machine learning algorithm implementations. In Fig. 1, an adapted version of the poll is shown, with only free tools listed, for comparison.

Most of the modern DM tools are effectively software-based dataflow architectures. Some of the tools (e.g. RapidMiner, KNIME) are graphical integrated environments that enable visual component placement, connection, and dragging. The most common paradigm for such components is: pull the data in, transform, and push it further down the pipeline. The component only pulls the data in once all of its prerequisites are met. The under-the-hood implementation of such components is irrelevant for the average user, but more advanced users usually have the possibility of tackling with the code in order to improve it. Other tools (e.g. R) are just plain extensions of the underlying language in the form of specialized packages and/or GUI add-ons.

General characteristics of the six DM tools are listed in Table I. All of the tools have implementations for Windows, Linux, and Mac OS X operating systems. In the large Table II, the supported machine learning algorithms for each DM tool are summarized. The tools either implement an algorithm (+), use an external add-on (A) to support it, show some degree of support for the procedure (S), or do not implement it (-) at all. It must be noted here that the data in Table II should be considered temporary, because most of the tools are in constant state of upgrades. Nevertheless, it is important and useful to summarize their

[1]We acknowledge the support of the Research Centre for Advanced Cooperative Systems ACROSS (EU FP7 #285939).

TABLE I. GENERAL CHARACTERISTICS OF THE FREE DATA MINING TOOLS

Characteristic	RapidMiner	R	Weka	Orange	KNIME	scikit-learn
Developer:	RapidMiner, Germany	worldwide development	Univ. of Waikato, New Zealand	Univ. of Ljubljana, Slovenia	KNIME.com AG,Switzerland	multiple; support: INRIA, Google
Programming language:	Java	C, Fortran, R	Java	C++, Python, Qt framew.	Java	Python+NumPy+ SciPy+matplotlib
License:	open s. (v.5 or lower); closed s., free Starter ed. (v.6)	free software, GNU GPL 2+	open source, GNU GPL 3	open source, GNU GPL 3	open source, GNU GPL 3	FreeBSD
Current version:	6	3.02	3.6.10	2.7	2.9.1	0.14.1
GUI / command line:	GUI	both; (GUI for DM = Rattle)	both	both	GUI	command line
Main purpose:	general data mining	sci. computation and statistics	general data mining	general data mining	general data mining	machine learning package add-on
Community support (est.):	large (~200 000 users)	very large (~ 2 M users)	large	moderate	moderate (~ 15 000 users)	moderate

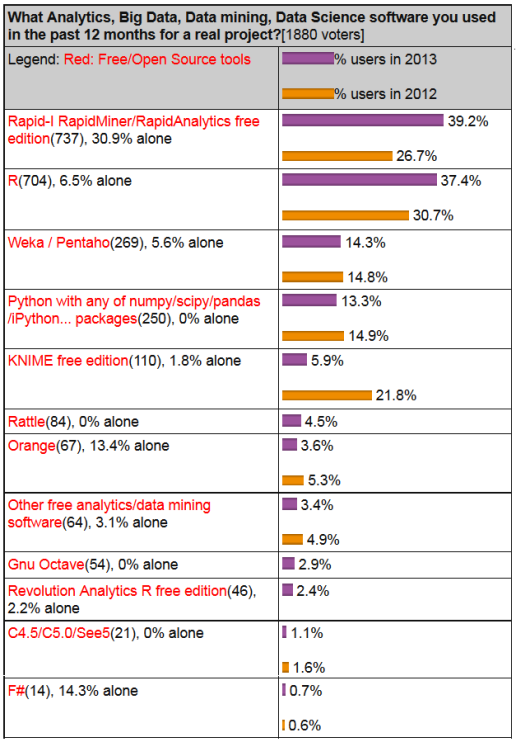

Figure 1. The community use of free DM tools, adapted from [8]

capabilities so that interested users can choose the appropriate environment for handling their problem. The algorithms shown in Table II were selected based on their significance and presence in most of the tools. Although the tools may contain some additional algorithms, not all of them could have been listed. Some of these were put in the appropriate rows with label "others". References to the work that describe the algorithms are not provided.

Table III lists the support for some of the more specialized topics in DM such as big data, text mining, etc.

III. TOOLS DESCRIPTION

A. *RapidMiner*

RapidMiner (previously: Rapid-I, YALE) is a mature, Java-based, general DM tool currently in development by the company RapidMiner, Germany. Previous versions (v. 5 or lower) were open source. The latest one (v. 6) is proprietary for now, with several license options (Starter, Personal, Professional, Enterprise). The Starter version is free with limitations only in respect to maximum allocated memory size (1 GB) and input files (.csv, Excel). The tool has become very popular in several recent years and has a large community support.

RapidMiner offers an integrating environment with visually appealing and user-friendly GUI. Everything in RapidMiner is focused on processes that may contain sub-processes. Processes contain operators in the form of visual components. Operators are implementations of DM algorithms, data sources, and data sinks. The dataflow is constructed by drag-and-drop of operators and by connecting the inputs and outputs of corresponding operators. RapidMiner also offers the option of application wizards that construct the process automatically based on the required project goals (e.g. direct marketing, churn analysis, sentiment analysis). There are tutorials available for many specific tasks so the tool has a stable learning curve.

Although RapidMiner is quite powerful with its basic set of operators, it is the extensions that make it even more useful. Popular extensions include sets of operators for text mining, web mining, time series analysis, etc. Most of the operators from Weka are also available through extension, which increases the number of implemented DM methods (Table II). The tool has very few shortcomings. The most important one is the transition to a novel model of business. It remains to be seen whether the transition to proprietary license will limit the number of its users, but it may not be helpful. The support for deep learning methods and some of the more advanced specific machine learning algorithms (e.g. extremely randomized trees, various inductive logic programming algorithms) is currently limited. However, big data analysis via Hadoop cluster (Radoop) is supported.

B. *Weka*

Weka is a Java-based, open-source DM platform developed at the University of Waikato, New Zealand. The software is free under GNU GPL 3 for non-commercial purposes. Weka has had mostly stable

TABLE II. DATA MINING ALGORITHMS AND PROCEDURES SUPPORTED BY THE TOOLS

Category	Name	RapidMiner	R	Weka	Orange	KNIME
Data import	textual files (.txt, .csv)	+	+	+	+	+
	specific input format files	+ (e.g..arff, .xrff)	A (foreign)	+ (.arff, .libsvm)	+	+
	Excel/ spreadsheet	+	A (xlsx)	-	-	A
	database table	+	A (RODBC)	+	+ (prototype)	+
	data from an URL	+	+	+	-	+
Feature selection	filters	+	A (FSelector)	+	+	+
	wrappers	+	A (FSelector)	+	+	+
Feature transformation	discretization	+	A (RWeka)	+	+	+
	normalization	+	A (RWeka)	+	+	+
	PCA	+	+	+	+	+
	ICA	+	+ (fastICA)	-	-	-
	MDS	-	+	+	+	+
	SVD	+	+	+	A	-
	random projections	-	+	-	-	+
Decision tree learner	ID3	A (Weka)	-	+	+	A (Weka)
	C4.5	A (Weka)	A (RWeka)	+	+	-
	CART	A (Weka)	A (RWeka)	+	+	A (Weka)
	others	+, A (own*, dec. stump)	+, A (own*, RWeka)	+ (dec. stump)	+ (own*)	+ (own*)
Classification rules	1Rule	+	A (RWeka)	+	-	A (Weka)
	PART	A (Weka)	A (RWeka)	+	-	A (Weka)
	RIPPER	+	A (RWeka)	+	-	A (Weka)
	others	+ (subgroup discovery)	A (RWeka)	+ (Ridor)	+ (CN2, subgroup discovery)	A (Weka)
Bayesian networks	Naive Bayes	+	+	+	+	+
	full bayesian network	A (Weka)	-	+	-	A (Weka)
	AODE	A (Weka)	A (AnDE)	+	-	A (Weka)
	others	A (Weka)	-	+ (DMNB, WAODE)	-	A (Weka)
Instance based learning	kNN	+	A (class, RWeka)	+	+	+
	LWL	A (Weka)	A (stats)	+	-	A (Weka)
	others	A (Weka) e.g. LBR	A (RWeka) e.g. LBR	+ (LBR)	+ (ML-kNN)	A (Weka)
Function based learning	regression analysis	+ (log., lin., polynomial)	+ (lin., nonlin.)	+ (log, lin.)	+ (log., lin., Lasso, PLS, trees, mean)	+(log., lin., poly, trees)
	ANN	+ (MLP)	A (nnet, RSNNS)	+ (MLP)	MLP	MLP, PNN
	SVM	+ (linear, evolut., PSO)	A (e1071, RWeka)	+ (SMO), A (libsvm)	+ (libsvm, liblinear)	+ (integr. libSVM)
	others	+ (Gaussian proc., RVM)	-	-	-	-
Hybrid learning methods	logistic model trees	A (Weka)	A (RWeka)	+	-	A (Weka)
	function trees	A (Weka)	-	+	-	A (Weka)
	Bayesian trees	A (Weka)	-	+	-	A (Weka)
	others	-	-	+ (DTNB), A (CBA)	-	A (Weka)
Ensemble	bagging	+	A (RWeka)	+	+	A (Weka)

Category	Name	RapidMiner	R	Weka	Orange	KNIME
learning	AdaBoost	+	A (RWeka)	+	+	A (Weka)
	random forest	+	A (random Forest)	+	+	+
	extremely randomized trees	-	+ (extra Trees)	-	-	-
	rotation forest	A (Weka)	S	+	-	A (Weka)
	LogitBoost	+	A (RWeka)	+	-	A (Weka)
	option tree	A (Weka)	-	+	-	A (Weka)
	stacking	+	A (RWeka)	+	-	A (Weka)
	other	+ (Bayesian boosting)	+ (boosted regression)	+ (Multi Boost)	-	A (Weka)
Hierarchical clustering	linkage-based	+	+	+	+	+
	AGNES	-	A (cluster)	-	-	-
	BIRCH	-	A (birch)	-	-	-
Centroid (partition) based clustering	*k*-means	+	+	+	+	+
	X-means	+	A (RWeka)	+	-	A (Weka)
	Partition Around Medoids	-	A (cluster)	-	A	A
	fuzzy *c*-means clustering	-	A (e1071)	-	A	+
Distribution based clustering	EM clustering	+	A (mclust)	+	-	A (Weka)
Density based clustering	DBSCAN	+	A (fpc)	+	-	A (Weka)
	OPTICS	A (Weka)	-	+	-	A (Weka)
ANN based clustering	Self-organising map	+	A(RSNNS)	A	+	A (Weka)
Association rules (unsupervised)	GSP	+	-	+	-	A (Weka)
	Apriori	A (Weka)	A (arules, RWeka)	+	+ (sparse, attr.-value)	A (Weka)
	FP-growth	+	-	+	-	A (Weka)
	Eclat	-	A (arules)	-	-	-
	Tertius	A (Weka)	A (RWeka)	+	-	A (Weka)
	others	-	A (arules Sequences)	+ (predictive Apriori)	+ (Apriori-SD)	A (Weka)
Evaluation methods and metrics	holdout	+	+	+	+	+
	cross-validation	+	A (cvTools)	+	+	+
	classif. accur., mean abs.err., MSE, κ	+	A (ROCR)	+	+	+
	TP, FP, FN, TN conf. matrix, precis., recall	+	A (ROCR)	+	+	+
	ROC, Lift chart, Cost-benefit	+	A (pROC)	+	+	+
Data visualization	histograms	+	+	+	+	+
	scatterplots	+	+	-	+	+
	other plots (e.g. box-whisker, mean/error)	+	+	-	+	+
	3D graphs (e.g. bivar. hist., surface plots)	S	A (lattice)	-	-	S (scatterplot3D)

* Own implementation of decision tree, mostly similar to C4.5 and CART

popularity over the years, which is mainly due to its user friendliness and the availability of a large number of implemented DM algorithms. It is still not as popular as RapidMiner or R, both in business and academic circles, mostly because of some slow and more resource demanding implementations of DM algorithms. Although it is not a single tool of choice in DM, Weka is still quite powerful and versatile, and has a large community support.

Weka offers four options for DM: command-line interface (CLI), Explorer, Experimenter, and Knowledge flow. The preferred option is the Explorer which allows

TABLE III. SUPPORT FOR SPECIALIZED AND ADVANCED DATA MINING TASKS

Name	RapidMiner	R	Weka	Orange	KNIME	scikit-learn
Big data	S (not free: Radoop)	A (ff, ffbase)	S (CLI, knowl. flow, distributedWekaHadoop)	-	A	S
Link, graph mining	-	A (igraph, sna)	A	-	A	-
Spatial data analysis	-	A (ggmap)	-	-	A	S
Time-series analysis	A	+, A(forecast)	S (several time series filters)	-	+	S (timeseries module has bugs)
Semi-super-vised learning	S	A (upclass)	S	-	S	+ (label propagation)
Data streams	+	A (stream)	A (massiveOnlineAnalysis)	-	+	S
Text mining	A	A (tm, RTextTools, qdap)	S	A	A	+
Paralelization	S (enterprise ed.)	A (snow, multicore)	S	-	+	A (joblib)
Deep learning	-	S (darch: incomplete)	-	-	-	S (Restricted Boltzmann Mach.)

the definition of data source, data preparation, machine learning algorithms, and visualization. The Experimenter is used mainly for comparison of the performance of different algorithms on the same dataset. Knowledge flow is akin to RapidMiner's operator paradigm in a sense that it allows one to specify the dataflow using appropriately connected visual components. Although not as visually appealing and extendable as RapidMiner, Weka's Knowledge flow still does the job.

Weka supports many model evaluation procedures and metrics, but lacks many data survey and visualization methods. It is also more oriented towards classification and regression problems and less towards descriptive statistics and clustering methods, although some improvements were made recently with respect to clustering. The support for big data, text mining, and semi-supervised learning is currently limited, while deep learning methods are still not considered.

C. R

The open-source tool and programming language of choice for statisticians, R, is also a strong option for DM tasks. R has been in development for the last 15 years and is the successor of S, a statistical language originally developed by Bell Labs in 1970s. The source code of R is written in C++, Fortran, and in R itself. It is an interpreted language and is mostly optimized for matrix based calculations, comparable in performance to commercially available Matlab and freely available GNU Octave.

The main language is extended by a myriad collection of packages for all sorts of computational tasks. Some of the packages are listed in Table II in parentheses. The tool offers only a simple GUI with command-line shell for input. It is certainly not a user-friendly environment because all commands need to be entered in the R language. The learning curve is steep, and although simple tasks such as drawing graphs and descriptive statistics can be learned rather easy, the language's full potential is difficult to master.

Some advanced users sometimes offer helpful reference cards with the list of the most significant functions [9]. From DM user's perspective, R offers very fast implementations of many machine learning algorithms, comparable in number to RapidMiner and Weka (from which a large number of algorithms is borrowed), and also the full prospect of statistical data visualizations methods. It has specific data types for handling big data, supports parallelization, data streams, web mining, graph mining, spatial mining, and many other advanced tasks, including a limited support for deep learning methods.

R's main problem is its language, which, although highly extendable, is also a difficult one to learn thoroughly enough to become productive in DM. Advancement for DM tasks in that direction is the Rattle package (author Dr. Graham Williams and other contributors) that offers a decent GUI for R. Rattle, which is in development from 2006, is similar to Weka's Explorer in the sense of user-friendliness. It loads separate packages from R upon request for a specific analysis. Rattle uses some of the R's standard implementations of DM methods and also additional packages. The only problem with Rattle is that it cannot use all of the R's algorithms or Weka's DM implementations. Nevertheless, Rattle is user-friendly and quite popular in DM community.

D. Orange

Orange is a Python-based tool for DM being developed at the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana. It can be used either through Python scripting as a Python plug-in, or through visual programming. Its visual programming interface, Orange Canvas, offers a structured view of supported functionalities grouped into nine categories: data operations, visualization, classification, regression, evaluation, unsupervised learning, association, visualization using Qt, and prototype implementations.

Functionalities are visually represented by different widgets (e.g. read file, discretize, train SVM classifier etc). A short description of each widget is available within the interface. Programming is performed by placing

widgets on the canvas and connecting their inputs and outputs. The interface is very polished and visually appealing, offering a pleasant user experience.

One apparent downside of Orange is that the number of available widgets seems limited when compared to other tools such as RapidMiner or KNIME, especially because of the lack of integration with Weka. Still, the coverage of standard data mining techniques is quite good, as can be seen from Table II. Furthermore, there are a number of interesting widgets currently in development that can be found in the "Prototype" category, so it is reasonable to expect that the feature set will be expanded in the future.

E. scikit-learn

scikit-learn is a free package in Python that extends the functionality of NumPy and SciPy packages with numerous DM algorithms. It also uses the matplotlib package for plotting charts. The package keeps improving by accepting valuable contributions from many contributors and is supported by both INRIA and Google Summer of Code. One of its main strong points is a well-written online documentation for all of the implemented algorithms. Well-written documentation is a requirement for any contributor and is valued more than a lot of poorly documented algorithm implementations.

The package supports most of the core DM algorithms. However, several significant DM algorithm groups have been omitted currently, including classification rules and association rules. On the other hand, the package is strong in function-based methods including many general linear models and various types of SVM implementations. It is also quite fast despite being written in an interpreted language. This is mainly because the contributors are asked to optimize the code in various aspects, such as calling array based NumPy number crunching algorithm or writing wrappers for existing C/C++ implementations in Cython.

Despite its advantages, the use of scikit-learn still requires that one is a skilled programmer in Python because of its command-line interface. This will detract almost anyone not versed in this language because there are other tools that do not have this assumption.

F. KNIME

KNIME (Konstanz Information Miner) is a general-purpose DM tool based on the Eclipse platform, developed and maintained by the Swiss company KNIME.com AG. Its development started in 2004 at the University of Konstanz, Germany, and the initial version was released in 2006. KNIME is open-source, though commercial licenses exist for companies requiring professional technical support. According to the official website, KNIME is used by over 3000 organizations in more than 60 countries, and there seems to be a considerable community support.

The tool adheres to the visual programming paradigm present in most DM tools, where building blocks are placed on a canvas and connected to obtain a visual program. In KNIME, these building blocks are called nodes, and according to the official website, more than 1000 nodes are available through the core installation and various extensions. Nodes are organized in a hierarchy and can be searched by name within an intuitive interface. Each node is documented in detail, and the documentation is automatically shown within the interface once the node is selected.

A large repository of example workflows is available to facilitate quicker learning of the tool. One of the greatest strengths of KNIME is the integration with Weka and R. Although extensions have to be installed to enable the integration, the installation itself is trivial. Weka integration enables using almost all the functionality available in Weka as KNIME nodes, while R integration enables running R code as a step in the workflow, opening R views and learning models within R. Several other interesting free extensions are also available, e.g. JFreeChart extension that enables advanced charting, OpenStreetMap extension that enables working with geographical data, etc. There are also commercial extensions for more specific functionalities. Overall, KNIME seems to be one of the best choices for a user interested in a purely visual programming paradigm with a need for a large variety of nodes.

IV. CONCLUSION

Several DM tools were presented in this work. Overall conclusion is that there is no single best tool. Each tool has its strong points and weaknesses. Nevertheless, RapidMiner, R, Weka, and KNIME have most of the desired characteristics for a fully-functional DM platform and therefore their use can be recommended for most of the DM tasks.

REFERENCES

[1] D. Pyle, Data Preparation for Data Mining, San Diego: Academic Press, 1999.

[2] M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, Boca Raton: CRC Press, 2013.

[3] Y. Zhao, R and Data Mining: Examples and Case Studies, San Diego: Academic Press, 2012.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," SIGKDD Explorations, vol. 11, no. 1, pp. 10–18, 2009.

[5] J. Demšar, T. Curk, and A. Erjavec, "Orange: Data Mining Toolbox in Python," Journal of Machine Learning Research, vol. 14, pp. 2349−2353, 2013.

[6] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, et al., "KNIME: The Konstanz Information Miner", in Data Analysis, Machine Learning and Applications (Studies in Classification, Data Analysis, and Knowledge Organization), Springer Berlin Heidelberg, pp. 319–326, 2008.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[8] G. Piatetsky, KDnuggets Annual Software Poll: RapidMiner and R vie for first place, 2013, Available at [last accessed 2014-02-22]: http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html

[9] Y. Zhao, R Reference Card for Data Mining, Available at: [last accessed 2014-02-23] http://www.rdatamining.com/docs/R-refcard-data-mining.pdf