XVII International Scientific Conference on Industrial Systems (IS'17)
Novi Sad, Serbia, October 4. – 6. 2017.
University of Novi Sad, Faculty of Technical Sciences,
Department for Industrial Engineering and Management
Available online at http://www.iim.ftn.uns.ac.rs/is17

A Comparison of Contemporary Data Mining Tools

Dakić Dušanka
(Teaching Assistant, University of Novi Sad, Faculty of Technical Sciences, Serbia, dakic.dusanka@uns.ac.rs)
Stefanović Darko
(Assistant Professor, University of Novi Sad, Faculty of Technical Sciences, Serbia, darkoste@uns.ac.rs)
Sladojević Srdjan
(Assistant Professor, University of Novi Sad, Faculty of Technical Sciences, Serbia, Sladojevic@uns.ac.rs)
Arsenović Marko
(Teaching Assistant, University of Novi Sad, Faculty of Technical Sciences, Serbia, arsenovic@uns.ac.rc)
Lolić Teodora
(Teaching Assistant, University of Novi Sad, Faculty of Technical Sciences, Serbia, teodora.lolic@uns.ac.rs)

Abstract

Corresponding to the raising importance of data mining in today's highly competitive business environment, the number of available data mining tools continues to grow. Consequently, competition between data mining software developers increases as well and the choice of the most suitable tool becomes increasingly difficult. Therefore, comparison of data mining tools becomes important. In this paper, several, frequently used open-source data mining tools and tools with open-source algorithms implementations are selected and compared against user groups, data structures, algorithms included, visualization capabilities, platforms, programming languages, and import and export options. In addition, evaluation of publicly available datasets has been performed by using selected tools. By performing actions such as data format conversions, data input and output, data transformation, feature selection, classification tasks, simple data mining tool selection algorithm has been developed and presented.

Key words: *data mining, data mining tools, tool selection algorithm*

1. INTRODUCTION

Data mining is a core step in the knowledge discovery from databases (KDD) process, that consists of applying data analysis and discovery algorithms on data in order to discover useful patterns [1]. Increasing power of technology and complexity of datasets has lead data mining to evolve from static to more dynamic and proactive information deliveries. Over the time, the volume of data that needs to be managed became impossible to manually analyse in order to get valuable decision making information. Therefore, an urgent and constant need for new, improved generations of data mining tools appeared, in order to assist in extracting useful information from the rapidly growing amount of data [1].

As a result, the number of available data mining tools continues to grow, with an accent on open-source data mining software. Open-source tools represented a new trend in data mining, especially in small and medium enterprises in early 2000s [2]. Nowadays it is an established trend, as open-source data mining tools are constantly being developed and renewed, offering bigger flexibility and extensive development community [3].

A problem for managers, developers, data scientists and other stakeholders might occur while choosing suitable open-source tool, as there is a wide variety of tools on the market and many of them are not yet well-known in data mining community. A need for comparisons of different tools is rising, as decision making process becomes more complex.

Therefore, in this paper five contemporary, open-source data mining tools are presented and compared against their important features in order to aid scientists in decision making process. Among the others, contribution of this paper is a simple tool selection algorithm, which enables selection of data mining tool according to users needs.

The rest of the paper is organized as follows: Section 2 deals with the related work, applied methodology is described in Section 3, Section 4 deals with the results and appropriate discussion and finally Section 5 addresses the conclusions and further research.

2. RELATED WORK

Over the years, a number of authors have written about comparison of data mining tools, all suggesting similair criteria for comparison and important features of data mining systems. Selecting software is a complex problem, due to many criteria and frequent technology changes, as new tools or new versions of existing tools are being released rapidly [4]. Current literature might be useful in terms of defining criteria of comparison, but not useful in regard to technology (tools) described in them, as they quickly become outdated. Nevertheless, in this chapter several papers will be mentioned, with accent on comparison criteria used in this paper.

Wang et al. (2008) in their comparison of leading data mining software packages, compared them against several software quality criterias, such as portability, reliability, efficiency, human engineering, understanding, modifiability, price, training and support [4]. From their reserach, several criterias were selected as relevant, such as portability, in therms of supported platforms and software architectures, human engineering, in therms of ease of use and understanding, with focus on graphical user interface (GUI).

Chen et al. (2007) in a survey of open-source data mining systems claim that each data mining process is composed of a sequence of data mining operations, each implementing a data mining function or algorithm, where these operations (data understanding, data preprocessing, data modeling, evaluation and deployment) are base for data mining tool's comparison [2]. To understand the characteristics of these diverse open-source data mining systems and to evaluate them, they looked into the following important features: ability to access various data sources, data pre-processing capability, integration of different techniques, ability to operate on large datasets and good data and model visualization [2]. Among their criteria, data sources, data pre-processing and operation on large datasets were chosen as significant and applied in this research.

Similairly, Goebel and Gruenwald, in 1999, proposed feature classification scheme, divided into following sections: general characteristics, database conectivity and data mining characteristics [5]. These categories can be used as sections in which all other criteria can be distributed.

Mikut and Reishl, 2011, singled out nine significant criteria for comparing data mining software. These criteria are based on user groups, data structures, data mining tasks and methods, import and export options and license models, where every criteria is further explained [6]. In this research, user groups and data mining tasks are singled out as criteria relevant for comparison.

According to above mentioned several criteria can be separated as important and repeating, such as possible data sources, data mining tasks (pre-processing, prediction, classification, etc.) and human engineering. Selected criterias are further grouped into three categories: general characteristics, data management and functionality.

3. METHODOLOGY

3.1 Open-source data mining tools

Data mining tools and tools with open-source algorithms implementations compared in this paper are WEKA, Microsoft Azure ML Studio, RapidMiner, H2O and Apache Spark.

WEKA, a well-known open-source Java application produced by the University of Waikato in New Zealand, is a collection of machine learning algorithms for data mining tasks [7].

Microsoft Azure Machine Learning Studio is critical part of the Cortana Analytics suite and provides an interactive visual workspace that enables data scientists and developers to easily build, test and deploy predictive analytic models [8]. It does not require installation, as it is used from web based GUI.

RapidMiner is a powerful visual workflow designer for building predictive analytic workflows, with a user-friendly GUI [9]. RapidMiner consists of RapidMiner Studio, where users can build and edit analytic processes, RapidMiner Cloud, where users can store their data, RapidMiner Server, which connects and interacts with other services and RapidMiner RapidMiner Radoop, which is a code-free environment for designing advanced analytic processes that push computations down to Hadoop cluster [10]. RapidMiner Studio is free, altough it has some limitations regarding logical processors and the amount of data used.

H2O in is an open-source tool for big data analyis, and it enables in-memory, distributed, fast, and scalable machine learning support. It offers H2O Flow, a web based, open-source user interface [11]. One of it's most regarding cappabilities is training a model on complete datasets and in-memory distributed parallel processing.

Apache Spark is a popular open-source platform for large-scale data processing that is well-suited for iterative machine learning tasks [12].

3.2 Criteria for comparison

Criteria for comparison of open-source data mining tools can be divided into general characteristics, data management, functionality, usability and classification.

3.2.1 *General characteristics*

By general characteristics, some general product features are considered, including product name, architecture, operating systems and programming languages. As far as architecture is considered, data mining tools can initially be subdivided into standalone and client/server architecture. Tools that have standalone architecture do not require any software other than the operating system to run, unlike client/server architecture. One of the emerging trends is an increasing number of web interfaces providing data mining as SaaS, therefore web based, cloud service architecture will also be included as possible value for architecture of data mining tools [6].

3.2.2 *Data management*

It is unrealistic to expect one data mining system to mine all kinds of data, given the diversity of data types and data mining agendas [13]. Data mining tools require integration with database systems or data warehouses for data selection, pre-processing, transformation, etc. Not all tools have the same database characteristics in terms of data model, possible data size, data format and import options.

This comparison includes possible data sources for data mining tools, such as connections to databases. Also, possible data formats are determined, considering if data can be stored in ARFF, CSV or Excel. Last feature important for data management is the size of the dataset. For every data mining tool it will be determined whether it is suited for small or large datasets (Big data).

3.2.3 *Functionality*

At the core of the KDD process are data mining methods for extracting patterns from data. These methods can be very diverse and have different goals [5]. In the study presented, open-source data mining tools will be compared in regard to the following methods/tasks: Data Pre-processing [14], Prediction, Regression, Classification, Clustering, Text Analytics and Model Visualization [5].

3.2.4 *Usability*

Usability is, generally, ease of use and learnability and can refer to human interaction with the tool. Data mining process should be highly interactive. Thus, it is important to build flexible user interfaces and an exploratory mining environment, facilitating user's interaction with the system [13]. Hence, in this paper, usability features include existence of GUI or CLI and user groups. User groups are divided into Business application, Applied research and Education. Business application users use data mining as a tool for mining large volumes of business data. Applied research users apply data mining to research problems and are mainly interested in tools with well proven methods, a graphical user interface (GUI), and interfaces to relevant databases [6]. Education users use data mining tools for education at universities. These data mining tools should be very intuitive, with a comfortable interactive graphical user interface.

3.2.5 *Classification*

In order to perform simple classification in each data mining tool, decision tree and decision forest models are build using well-known Iris dataset [15]. Iris is multivariate dataset, with four real attributes: sepal length in cm, sepal width in cm, petal length in cm, petal width in cm. The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant: Iris Setosa, Iris Versicolor, and Iris Virginica. Initial Iris dataset has 4KB size. Weka, RapidMiner and Apache Spark have implementation of J48 (C4.5) decision tree, which is used in classification process in this tools. On the other hand, Microsoft Azure ML Studio and H2O Flow do not offer multiclass decision trees for classification, but decision forests instead.

After classification is performed, in order to test performance of data mining tools on larger dataset, Iris dataset was enlarged up to 20MB (around 1000000 records) and the same procedure was repeated.

3.3 Tool selection algorithm

After performing the evaluation process and reviewing the results of comparison, a simple algorithm is developed with the aim to help decision-making process when it comes to choosing open-source data mining tools.

Finally, algorithm included only attributes relevant for decision making process and excluded attributes that have same values for each tool.

4. RESULTS AND DISCUSSION

4.1 **General characteristics**

General characteristics of selected data mining tools are presented in Table 1.

WEKA's architecture can be standalone or client/server, and it can operate on Windows, MAC and Linux OS. WEKA is developed in Java and the algorithms can either be applied directly to a dataset or called from Java code, because of the Java API.

Microsoft Azure Machine Learning Studio has web-based architecture. Operating systems on which Azure ML operates are Windows and Linux. Azure Machine Learning Studio includes hundreds of built-in packages and support for custom R and Python code, although Python can only be used for pre-processing tasks.

RapidMiner has client/server, standalone and cloud architecture, as it consists of RapidMiner Studio, Server and Cloud. It can operate on Windows, MAC and Linux, and is written in Java, but can use R and Python scripts.

H2O uses web-based open-source user interface. It can run on Windows, OS X and Ubuntu operating systems and allows programming in R, Python, Scala, Java and JSON. There is a possibility of running H2O on Hadoop cluster. H2O is supported on a number of cloud environments.

Apache Spark is a fast and general-purpose cluster computing system, meaning that Spark applications run as independent sets of processes on a cluster (standalone, Apache Mesos or Hadoop YARN). It is also possible to submit applications from a distant computer, so Apache Spark supports client/server architecture, or to run applications on cloud platform. It provides high-level APIs in Java, Scala, Python and R.

Apache Spark can be run on Windows, Linux Ubuntu and MAC operating systems.

Table 1. General characteristics

Product	***Architecture***	***OS***	***Language***
Weka	Client/Server, Standalone	Windows, MAC, Linux	Java
Azure ML Studio	Web based, Cloud service	Windows, Linux	R, Python
RapidMiner	Standalone, Client/Server, Cloud service	Windows, MAC, Linux	Java, R, Python
H2O	Web based, client/server, Cloud service	Windows, OS X, Linux Ubuntu	R, Python, Scala, Java, JSON
Apache Spark	Standalone (on a cluster), Client/Server, Cloud	Windows, Linux Ubuntu, MAC	R, Python, Scala, Java

4.2 **Data management results**

Table 2. presents which data sources and formats tools support. In the last column is determined if certain data mining tool is suitable for processing big data. In the last column of the Table 2 (Column S), it is marked that every tool can be used for processing of big data, although there is a certain limitation considering WEKA and RapidMiner. WEKA can only process big data if it is used through CLI and not GUI and it is necessary to manually enlarge heap size. RapidMiner can also process big data, but not free of charge.

Table 2. Data Management

Product	***J/O***	***S***	***E***	***OR***	***P***	***C***	***A***	***U***	***S***
Weka	✓	✓		✓	✓	✓	✓	✓	✓*
Azure ML Studio	✓	✓	✓	✓	✓	✓	✓	✓	✓
RapidMiner	✓	✓	✓	✓	✓	✓			✓*
H2O	✓		✓		✓	✓	✓	✓	✓
Apache Spark	✓	✓	✓	✓	✓	✓	✓	✓	✓

J – JDBC O – ODBC S - MS SQL SERVER E - MS EXCEL OR – Oracle P – PostgreSQL C – CSV A – ARFF U - WEB URL S – SIZE

4.3 **Functionality results**

As Table 3 presents, each tool enables every data mining task listed, increasing amount of acceptable and appropriate tools and the list of their capabilities, making it difficult to choose the most suitable tool.

Table 3. Functionality

Product	***DP***	***P***	***R***	***CS***	***C***	***T***	***L***	***M***	***E***
Weka	✓	✓	✓	✓	✓	✓	✓	✓	✓
Azure ML Studio	✓	✓	✓	✓	✓	✓	✓	✓	✓
RapidMiner	✓	✓	✓	✓	✓	✓	✓	✓	✓
H2O	✓	✓	✓	✓	✓	✓	✓	✓	✓
Apache Spark	✓	✓	✓	✓	✓	✓	✓	✓	✓

DP – Data pre-processing, P – Prediction, R - Regression, CS - Classification, C - Clustering, T – Text Mining, L – Link Analysis, M – Model visualization, E - Exploratory Data Analysis

4.4 **Usability results**

The result of usability comparison is presented in Table 4. As it is presented, every tool except Apache Spark has GUI. Therefore only Apache Spark is not suitable for educational purposes. As business application and applied research is considered, every tool can be used. CLI is available in WEKA, H2O, Apache Spark and RapidMiner.

Table 4. Usability

Product	***GUI***	***CLI***	***BA***	***AR***	***E***
Weka	✓	✓	✓	✓	✓
Azure ML Studio	✓		✓	✓	✓
RapidMiner	✓	✓	✓	✓	✓
H2O	✓	✓	✓	✓	✓
Apache Spark		✓	✓	✓	

GUI – Graphic User Interface, CLI – Command Line Interface, BA – Business Application, AR – Applied Research, E – Education

4.5 **Classification results**

In Table 5. are presented classification results performed in WEKA, RapidMiner and Apache Spark, using C4.5 algorithm (or J48 implementation). In Table 6. are results of classification performed in Microsoft Azure ML Studio and H2O, using random decision forest algorithm. Results are including area under the curve (AUC), percentage of correctly classified instances, true positive ratio (sensitivity) and true negative ratio (specificity).

Classification in WEKA is performed through simple GUI where user imports data, trains and tests learning schemes and performs model metrics visualization.

Performing classification in Microsoft Azure ML Studio is simple, because of intuitive, drag and drop graphic user interface. Once iris.arff dataset is uploaded, it can easily be converted to other file formats, such as CSV. Next step is splitting the dataset and training model with machine learning algorithm, then scoring and evaluating the model.

RapidMiner has user-friendly GUI which enables user to perform classification by choosing blocks that represent data and operations performed on data. For building classification model, it is necessary to retrieve Iris dataset, set role for attribute the model should predict, split data into training and test set, train data using decision tree and then apply model on the rest of the dataset. Finally, it was necessary to include performance block to inspect performance of the model.

In H2O Flow, classification process is realised as a flow of actions, where the flow starts with importing and parsing iris dataset (to ARFF, CSV, XLS, etc.), then splitting the frame, building and evaluating the model.

Apache Spark classification was performed in spark-shell in CLI, using Scala programming language. Iris dataset is loaded and split into training set and test set. Then a model is created by generating decision tree instance, following with model evaluation on test instances.

Table 5. Decision tree classification results

Product	***AUC***	***Correctly classified***	***TPR***	***TNR***	***L***
Weka	0.97	98%	0.98	0.99	✓*
RapidMiner	x	95%	0.95	0.96	✓*
Apache Spark	0.97	0.98	0.97	0.98	✓

Table 6. Decision forest classification results

Product	***AUC***	***Correctly classified***	***TPR***	***TNR***	***L***
Azure ML Studio	x	97%	0.97	0.98	✓
H2O	x	97%	0.97	0.98	✓

AUC – Area under the curve TPR – True positive rate (sensitivity) TNR – True negative rate (specificity) L – Enlarged Iris dataset

After observing tables 5. and 6. it can be concluded that there is a minor difference in classification results among tools presented. However, it is prominent that RapidMiner, Microsoft Azure ML Studio and H2O do not offer AUC for multiclass classification. Moreover, Weka and RapidMiner have an accent (*) in Column L, with reference to their limitations in ability to process enlarged dataset. When Iris dataset was enlarged up to 20 MB, WEKA was not comfortably able to process it, unless the heap size was enlarged and CLI is used. Furthermore, open-source version of RapidMiner cannot process that volume of data free of charge, as its limit is 10000 records and one logic processor.

On the contrary, Apache Spark, H2O and Microsoft Azure ML Studio were able to process enlarged data set in seconds.

4.6 **Developed tool selection algorithm**

According to the presented research results, several characteristics are common for every evaluated tool and they will not be included in the algorithm. The tool selection algorithm is build from differences between the tools and has several different starting nods, in order to include all of them. The algorithm is presented in Figure 1.

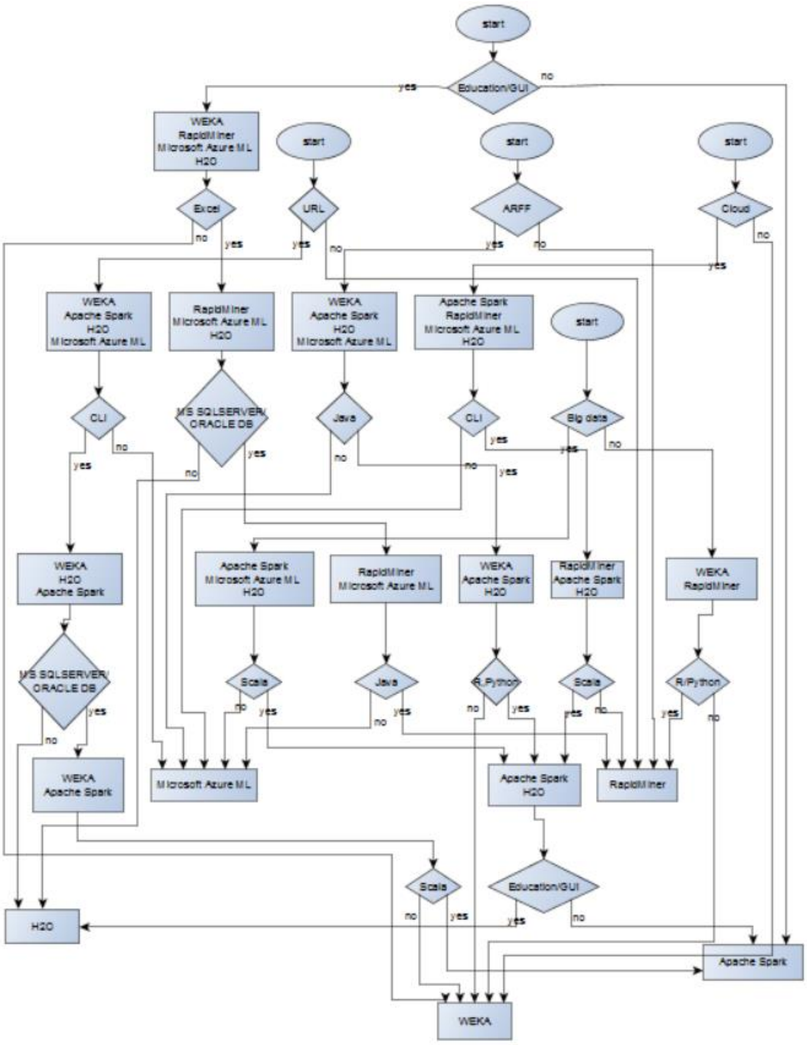

Figure 1. Tool selection algorithm

5. CONCLUSION

In the study presented, several contemporary open-source data mining tools were presented, quick evaluation process performed using well-known dataset, tools performance compared and a simple tool selection algorithm is developed. Data mining software developers are answering to users paramount needs, resulting in minor differences between tools. Each tool offers standard platforms, data mining tasks and data sources. Evaluation of classification process showed that there are minor differences when processing small datasets, compared to classification of larger volumes of data. Apache Spark, Microsoft Azure ML Studio and H2O performed classification of larger dataset in seconds, while WEKA and RapidMiner had certain limitations. Therefore, big data processing remains most distinctive quality of these tools, as certain open-source tools have in-memory data processing, parallel programming and iterative machine learning tasks.

Future work should focus on including more contemporary open-source data mining tools in the research, such as R, KNIME and Orange, as well as closer examinating differences in their functionality.

6. REFERENCES

[1] Fayyad U, Piatetsky-Shapiro G, Smyth P. (1996) "*From data mining to knowledge discovery in databases*" AI Mag 1996, 17:37–54.

[2] Chen X, Ye Y, Williams G, Xu X. (2007) "*A survey of open source data mining systems*", Lecture Notes in Computer Science 2007, 4819:3–14.

[3] Almeida, P., & Bernardino, J. (2016) "*A survey on open source Data Mining Tools for SMEs*", In *New Advances in Information Systems and Technologies* (pp. 253-262). Springer, Cham.

[4] Wang J, Hu X, Hollister K, Zhu D. (2008) "*A comparison and scenario analysis of leading data mining software*". Int J Knowl Manage 2008, 4:17–34.

[5] Goebel, M., Gruenwald, L. (1999) "*A survey of data mining and knowledge discovery software tools*“, In: SIGKDD Explorations. Volume 1., ACM SIGKDD (1999) 20–33

[6] Mikut, R., & Reischl, M. (2011) "*Data mining tools*", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(5), 431-443.

[7] Hall, Mark, et al. (2009) "*The WEKA data mining software: an update*." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.

[8] Barga, Roger, Valentine Fontama, and Wee Hyong To (2015) "*Introducing Microsoft Azure Machine Learning*." Predictive Analytics with Microsoft Azure Machine Learning. Apress, 2015. 21-43.

[9] Jovic A., Brkic K., Bogunovic N. (2014) "*An overview of free software tools for general data mining*." Information and Communication Technology, Electronics and Microelectronics (MIPRO), 37th International Convention on. IEEE, 2014.

[10] Rapid Miner user manual available at: https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf (accessed: 02.06.2017.)

[11] H2O documentation available at: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html (accessed: 01.06.2017.)

[12] Meng, Xiangrui, et al. (2016) "*Mllib: Machine learning in apache spark*." The Journal of Machine Learning Research 17.1: 1235-1241.

[13] Han, Jiawei, Jian Pei, and Micheline Kamber.(2011) *Data mining: concepts and techniques.* Elsevier, 2011.

[14] Kotsiantis, S. B., D. Kanellopoulos, and P. E. Pintelas.(2006) "*Data preprocessing for supervised leaning*." *International Journal of Computer Science* 1.2 2006: 111-117.

[15] Iris dataset is available at: https://archive.ics.uci.edu/ml/datasets/iris (accessed: 10.06.2017.)