Evaluating Open Source Data Mining Tools for Business

Pedro Almeida[1], Le Gruenwald[2] and Jorge Bernardino[3]
[1]*ISEC, Polytechnic of Coimbra, Rua Pedro Nunes, Quinta da Nora, 3030-190 Coimbra, Portugal*
[2]*University of Oklahoma, School of Computer Science,*
110 W. Boyd St., Room 150 DEH, 73019 Norman, Oklahoma, U.S.A.
[3]*Centre of Informatics and Systems, University of Coimbra, Pinhal de Marrocos, 3030-290 Coimbra, Portugal*

Keywords: Data Mining, Data Mining Tools, Open Source.

Abstract: Businesses are struggling to stay ahead of competition in a globalized economy where there are more and stronger competitors. Managers are constantly looking for advantages that can generate benefits at low costs. One way to have such advantage is using the data about customers, demographic data, purchase history, customer behavior and preferences that can help to take better business decisions. Data Mining addresses the challenges of collecting value inside data and the ways to put that value to use for virtually any area of our lives, including business. In this paper, we address the interest of Data Mining for business and analyze three popular Open Source Data Mining Tools – KNIME, Orange and RapidMiner – considered as a good starting point for enterprises to begin exploring the power of Data Mining and its benefits.

1 INTRODUCTION

Data Mining is a field of computer science that brings together many different disciplines with the goal of performing different tasks with data such as pattern finding, organization of information about hidden relationships, association rules structuring, classification of objects based on the value of unknown items, creation of clusters of similar objects and unveiling of other relevant findings that cannot be found with classic techniques of data analysis (Witten et al., 2011). Data Mining is useful for many areas of our lives such as gaming, science, engineering, human rights, medicine, security and many others. It is also of great interest for the area of business. In this area, Data Mining can be used for many different purposes that share the same objective. That objective is helping the business thrive, be one step ahead of competitive enterprises and provide solutions for operating problems, social concerns and economic issues the enterprises face every day. The basic use of Data Mining in business is analyzing stored data about past business activities and transactions and extracting unknown patterns and trends. Advances in technology have allowed for the cross-analysis of stored data with data that is streamed live in order to provide a more accurate and faster response to client demands. Concrete examples of businesses that use Data Mining include market analysis for product bundle identification, prevention of customer attrition, customer acquisition, cross-selling to existing customers and also more accurate profiling of existing customers (O'Brien and Marakas, 2011). The vast majority of businesses can generate a lot of data from their work. But this does not mean they will have a spare budget to spend on Data Mining Tools. This should not be a reason for them to abdicate the value of Data Mining and this is where Open Source tools become important. Open Source Data Mining Tools provide small businesses the opportunity to tap into the potential of data with minimal costs or even no costs at all. They are also more flexible than proprietary solutions and have a faster renovation process that makes them updated to answer new challenges. They provide all these capabilities with no consequence to the robustness necessary for business environments. In this paper, based on the previous works from (Borges et al., 2013); (Hasim and Haris, 2015); (Jović et al., 2014) and (Fernández et al., 2014) we choose to analyze three open source Data Mining platforms: KNIME, Orange and RapidMiner. The analysis of these platforms will be performed by running tests on several classifications algorithms.

The remainder of this paper is structured as follows. Section 2 describes the classification area of

Almeida, P., Gruenwald, L. and Bernardino, J.
Evaluating Open Source Data Mining Tools for Business.
DOI: 10.5220/0005939900870094
In *Proceedings of the 5th International Conference on Data Management Technologies and Applications (DATA 2016)*, pages 87-94
ISBN: 978-989-758-193-9

Data Mining analyzed in our experiments. Section 3 gives an overview of the three popular open source Data Mining platforms. Section 4 shows the results of the experiments. Section 5 details some related work on the subject of Open Source Data Mining Tools. Finally, Section 6 presents conclusions and points out some future work.

2 DATA MINING TECHNIQUES FOR BUSINESS

This section gives a short description of the Data Mining area. We explain in detail the classifications tasks that we explore in our experiments and highlight how businesses can use such tasks to gather knowledge from their data. Data Mining techniques can be subdivided into six main groups – Change and Deviation Detection (also known as Outlier Detection), Dependency Modeling (also known as Association Rules Learning), Summarization, Classification, Clustering and Regression (Fayyad et al., 1996). For our study we focus on Classification because it is the one area that has the most use for business environment (Petre, 2013); (Rajagopal, 2011). We now explain how the classification tasks work in detail, state some of its generic uses and also discuss some of the use cases for the business environments.

Classification is the task of classifying data under a generic structure. This structure is usually loaded before the data analysis is done rather than being built over the outputs of the analysis. Generic classification applications include classification of trends in the financial markets or classification of images in large image databases. In business one of the most common uses for classification is the categorization of items available in electronic commerce stores with a large and diverse inventory. The more correct the item categorization system is, the easier it will be for clients to find products that are similar to the ones they have previously bought or looked for and thus maybe of their interest. This type of classification works as a process of two steps: building a model composed of a number of preliminary groups that have similar items and then making a second round of classification that will find the definitive category each item belongs to (Shen et al., 2012). Data today is being not only generated but also collected by companies at an exponential rate. Without technological investment and techniques such as the ones we mentioned above, this data would be stored forever in data warehouses without any valuable use. If a company works with devices such as sensors, mobile devices, or RFID tags, it can easily generate great amounts of data in short periods of time. It is due to Data Mining tools and techniques that these companies are able to use the data to improve their business in any ways necessary such as determining sales trends, developing marketing campaigns and better profiling customers (Alexander, 2015); (Medri, 2013).

3 OPEN SOURCE DATA MINING TOOLS

Based on the works of (Borges et al., 2013), (Hasim and Haris, 2015); (Jović et al., 2014) and (Fernández et al., 2014), we choose to analyze the three open source Data Mining platforms: KNIME, Orange and RapidMiner. In this section we give an overview of the three tools and describe some of their main functionalities.

3.1 KNIME

KNIME is an enterprise level data analytics platform aiming to help organizations stay one step ahead of change through the use of data knowledge (KNIME, 2015). With a high level of customization it provides an adaptive learning curve according to the time and effort each user wants to spend with the tool. It is completely visual and free of code so that the user can focus on working with the data and not waste time with implementation details. Based on the Eclipse IDE, it has a modular and extensible API that is ideal to use at both commercial, research and educational settings. By providing hundreds of different processing nodes it offers powerful capabilities for tasks such as pre-processing, cleansing, modeling, analysis and mining. When it comes to Data Mining, KNIME has 13 groups of algorithms – Bayes, Clustering, Rule Induction, Association Rules, Neural Network, Decision Tree, Miscellaneous Classifiers (such as the K-Nearest Neighbor), Ensemble Learning, Multi-Dimensional Scaling (MDS), Principal Component Analysis (PCA), Predictive Model Markup Language (PMML), Support Vector Machine (SVM) and Feature Selection. One of its strong points is the high level or integration with other Data Mining tools such as Weka and R. Licensed under the GNU General Public License version 3, KNIME Analytics Platform can be extended with the KNIME

Commercial Software (that has to be paid for) for additional professional support. KNIME strong points are:

- High level of customization that adapts the learning curve to the user.
- Being based on the Eclipse IDE that is familiar to a lot of programmers.
- Easy integration with other Data Mining Tools.

3.2 Orange

Orange is an open source tool for data analysis and visualization developed at the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana (Demšar et al., 2013). It provides tools for the performing Data Mining tasks through a graphical interface and also through the use of Python script coding. Packed with numerous features for analytics and components for machine learning, its main feature is essentially the high level of possible expansion through the use of add-ons that give the core bundle extra possibilities for tasks such as text mining, bioinformatics among others. The visual interface, known as Orange Canvas, is very easy to use as it provides an understandable division of functionalities through nine different groups – data, visualization, classification, regression, evaluation, unsupervised learning, association, bioinformatics and prototypes. To further increase such ease of use, widgets were introduced as representation of the functionalities that can be placed and connected between them in the visualization area in the most intuitive way. The downside of Orange is that it is not as complete as other existing open source Data Mining tools.

Its latest stable version is 2.7., licensed under the GNU General Public License version 3. Orange main features are:

- Containing both GUI and Command-Line for Python script coding.
- Providing an easy-to-understand division of functionalities through nine different groups.

3.3 RapidMiner

RapidMiner is a tool for machine learning, data mining, text mining, predictive analytics and business analytics built not just for data scientists but also for business managers, developers and anyone with interest in this area (RapidMiner, 2015). RapidMiner Studio (the open source version which we focus on) has an easy-to-use visual environment that takes the user directly to the execution of data management tasks without requiring any coding.

Not only it is intuitive to use, but also it grants access to the help of a huge community of about 250,000 users. This community brings advantages such as speedy renovation of the tool, and fast and quality assistance for new users. The aforementioned qualities make RapidMiner very appealing for people who cannot use much time going through the learning curve. RapidMiner provides hundreds of existing methods for data transformation, modeling and visualization. It also gives the users a powerful and extensible API that can be used to upgrade the tool to include their own algorithms. It is highly versatile in terms of configurations and sizes of datasets since all its methods can run in-memory, in-database or in clusters that work with Hadoop (RapidMiner, 2015).

Other strong points of this tool is that it provides different visualization outputs such as 3D graphs, scattered matrices or maps, the multiple interfaces such as the GUI or the batch processing unit, the accuracy of pre-processing methods and the complete toolbox with over 1500 operations available. RapidMiner divides its Data Mining tasks in 7 groups – Classification and Regression, Attribute Weighting, Clustering and Segmentation, Association and Item Set Mining, Correlation and Dependency Computation, Similarity Computation and finally Model Application. Inside these groups we can find many different algorithms. Developed in Java, RapidMiner runs in every major platform and operating system. The open source version is very complete, with limits only on the size of memory that can be allocated (1GB) and types of accepted data sources (.csv and Excel only). To sum up, RapidMiner main advantages are the following:

- Support for all computer environments.
- Visual interface that abstracts the user from implementation details.
- API that provides extension capabilities and versatility of configuration.
- Support for in-memory, in-database and cluster processing.
- Variety of visualization outputs.

4 DATA MINING TOOLS TESTS

Our experimental work consists of testing a number of different algorithms in the area of Classification. All algorithms tested are present on at least two of the examined platforms. All algorithms are tested and compared for execution time and for a set of specific performance metrics that will be explained in the following subsection. All the results are based on four executions of each algorithm. These experiments were all run in a laptop computer with Windows 7^{TM} Home Premium 64 bits Operating System, Intel® Core^{TM} i5-3210M CPU @ 2.50 GHZ and 6 GB of RAM. We test three Open Source Data Mining Platforms – KNIME, Orange and RapidMiner – because they are the most suitable for out test setup in terms of hardware requirements and because, they are referenced in most of the literature on this area and are also featured in the list of tools most used for real Big Data projects provided by (Jović et al., 2014).

4.1 Selected Classification Algorithms and Performance Metrics

For our experiments on Classification, we chose to test five common classification algorithms – Decision Tree (known as Classification Tree on Orange), K-Nearest Neighbors (KNN), Neural Network, Naïve Bayes and Support Vector Machine (SVM). We made our tests using the Adult dataset (Lichman, 2013) that has 48,842 instances, 15 attributes (14 variables and 1 target class) and a size of 3.88 Megabytes. The goal of our classification task is to use the variables to predict the annual income of an individual. The instances of the dataset are classified in two classes that divide the subjects by their annual income – greater than 50K and less than or equal to 50K. Examples of the dataset instances are shown in *Table 1*.

Table 1: Examples of Dataset Instances.

Attribute Name	Instance 1	Instance 2
Age	39	31
Work class	State-gov	Private
Final Weight	77516	45781
Education	Bachelors	Masters
Education Number	13	14
Marital Status	Never-Married	Never-Married
Occupation	Adm-clerical	Prof-speciality
Relationship	Not-in-family	Not-in-family
Race	White	White
Gender	Male	Female
Capital Gain	2174	14048
Capital Loss	0	0
Work Hours p/Week	40	50
Native Country	United-States	United-States
Class	<=50	>50

Predicting the annual income of an individual (or an household) has many applications for different businesses. For banks it is important for tasks such as predicting clients that are likely to need and accept a loan offer, predicting the viability of conceding a loan and calculating the risk of payment failure once the loan is given. For sales businesses predicting the income of an individual is important to know exactly what clients to market for based on your product types. For example a company that sells luxury cars will not take much advantage from targeting customers with low income. For this company,given that it has a list of customer data, it is important to try and predict the annual income of such customers to eliminate the ones with low income and therefore lower the costs of the marketing operation.

This dataset tries to make the annual income prediction based on 14 different variables that we explain next. First is the Age of the individual that is a continuous variable. Also continuous variables are Final Weight (numeric value used to distinguish individuals with different demographic characteristics), Education Number (numeric value used to represent each type of education), Capital Gain (numeric value used to represented profits obtained from the sale of an investment or real estate), Capital Loss (numeric value used to represented losses generated from the sale of an investment or real estate) and Work Hours per Week. Other variables that are not continuous but assume one of a given set of values are – Work Class (represents the work class of the individual, can assume one of 8 different values), Education (represents the education level of the individual, can assume one of 16 different values), Marital Status (represents the marital status of the individual, can assume one of 7 values), Occupation (represents the actual job of the individual, can assume one of 14 different values), Relationship (represents the role of the individual in his/her relationship, can assume one of 6 values), Race (represents the race of the individual, can assume one of 5 different values), Gender (represents the gender of the individual, can assume one of two different values) and Native Country (represents the country of origin of the individual, can assume one of 41 different values). Our first experimental results are based on the execution times the best results of which are represented in *Table 2*. From this table it is easy to see that RapidMiner is the best platform when it comes to algorithm execution time because it is the fastest platform for four out of five algorithms tested.

Besides comparing the execution times of the algorithms, we also compared the results on seven other performance metrics – Precision, Recall, F-Measure, ROC, Accuracy, Specificity and Sensitivity. These metrics are the ones that are shown by the performance measuring nodes of the platforms. We will explain each of these metrics in this section. Classification algorithms work by predicting whether an instance belongs to one class or the other of the two existing classes in the model.

Table 2: Execution time of classification algorithms.

	Best Time	Platform
Decision Tree	106ms	RapidMiner
KNN	2s	KNIME
Neural Network	2s	RapidMiner
Naïve Bayes	72ms	RapidMiner
SVM	24m33s	RapidMiner

One class is considered as being the positive class (the target) and the other class is the negative. This generates four values – **True Negatives** (TN), **True Positives** (TP), **False Negatives** (FN) and **False Positives** (FP). **TNs** are the instances of the negative class that are predicted as being so. **TPs** are the instances that are of the target class and are predicted as being of that class. **FNs** are the instances that are of the target class but are predicted as being of the negative class, and finally **FPs** are instances that are of the negative class but are predicted as being of the target class. These four values are the basis of the calculations of the performance metrics we are analyzing.

Precision is the result of dividing TPs by the sum of TPs and FPs.

$$\textbf{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \qquad (1)$$

Recall is the result of dividing TPs by the sum of TPs and FPs.

$$\textbf{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad (2)$$

The F-Measure is the weighted average of the precision and the recall. It is calculated by dividing the multiplication of precision and recall by the sum of precision and recall. It can spawn values from 0 (worst value) to 1 (best value) (Powers, 2007).

$$\textbf{F-Measure} = \frac{\boldsymbol{Precision \times Recall}}{\boldsymbol{Precision + Recall}} \qquad (3)$$

The **ROC or ROC curve** is a graphical plot that plots the **true positive rate** (TPR) against the **false positive rate** (FPR, sum of TNs and FPs) for the total of instances analyzed (T). In our list of results, a numeric value is seen instead of a plot. This value is the Area Under Curve (AUC) that represents the area under the curve of the ROC graph (Fawcett, 2006).

$$\textbf{AUC} = \int_{\infty}^{-\infty} \boldsymbol{TPR(T) \times FPR'(T)dT} \qquad (4)$$

AUC has though been recently questioned as a poor metric for algorithm comparison since it is considered a noisy measure for classification (Hanczar et al, 2010); (Hand, 2009). **Accuracy** is the number of correctly predicted classifications. It is the result of the division of the sum of TPs and TNs by the total number of instances analyzed. The higher the accuracy given, the better the algorithm is (Dogan and Tankrikulu, 2012).

$$\textbf{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{T}} \qquad (5)$$

The **specificity** is the result of the division of TNs by the sum of TNs and FPs.

$$\textbf{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \qquad (6)$$

Finally the **sensitivity** is the result of the division of TPs by the sum of TPs and FNs (Grzymala-Busse and Marepally, 2010). Both sensitivity and specificity calculate correctly predicted instances (sensitivity for the positives and specificity for the negatives) and just like in the other metrics, it is important to have higher values.

4.2 Performance Metrics

The complete list of the best results for each performance metric on each algorithm is shown in *Table 3* and *Table 4*. Each result shown is the best result out of all executions in all platforms. In some of the metrics two results are shown because the platforms generate not one result for all the classified classes but one result for each classified class. For the dataset analyzed, the two classified classes are **<=50** (i.e. income <= 50K) and **>50** (i.e. income > 50K). In the situations where different platforms have the best results the table shows the notation of the class next to the platform. If it does not show the class notation this means the same platform has the best result for both classes.

Analyzing both *Table 3* and *Table 4* we can conclude that Orange is the platform that gives the best results for most of the metrics. RapidMiner also gives a significant number of best results. KNIME never gives best results. As we mentioned before the tools will generate a precision result for each of the two classified classes.

Table 3: Best results for classification algorithms, part I.

	Parameter	Result	Platform
Decision Tree	Precision	0.878 0.890	Orange (<=50) RapidMiner (>50)
	Recall	0.988 0.606	RapidMiner (<=50) Orange (>50)
	F Measure	0.888 0.629	Orange
	ROC	0.792	Orange
	Accuracy	0.828	Orange
	Specificity	0.989	RapidMiner
	Sensitivity	0.898 0.606	Orange
KNN	Precision	0.862 0.596	Orange
	Recall	0.881 0.555	Orange
	F Measure	0.871 0.575	Orange
	ROC	0.830	Orange
	Accuracy	0.802	Orange
	Specificity	0.555 0.881	Orange
	Sensitivity	0.881 0.555	Orange

Table 4: Best results for classification algorithms, part II.

Neural Network	Precision	0.855 0.743	RapidMiner (<=50) Orange (>50)
	Recall	0.935 0.639	Orange (<=50) RapidMiner (>50)
	F Measure	0.906 0.659	Orange
	ROC	0.908	Orange
	Accuracy	0.853	Orange
	Specificity	0.936	RapidMiner
	Sensitivity	0.935 0.592	Orange
Naïve Bayes	Precision	0.922 0.733	Orange (<=50) RapidMiner (>50)
	Recall	0.939 0.777	RapidMiner (<=50) Orange (>50)
	F Measure	0.876 0.676	Orange
	ROC	0.900	Orange
	Accuracy	0.837	RapidMiner
	Specificity	0.939	RapidMiner
	Sensitivity	0.835 0.777	Orange
SVM	Precision	0.860 0.716	Orange
	Recall	0.990 0.519	RapidMiner (<=50) Orange (>50)
	F Measure	0.896 0.602	Orange
	ROC	0.892	Orange
	Accuracy	0.835	Orange
	Specificity	0.990	RapidMiner
	Sensitivity	1.000	KNIME

Out of ten precision results analyzed, Orange is the best tool for seven of them and RapidMiner is the best one for the remaining three. The same thing happens for Recall. Orange gives all the best results in the metrics of F-Measure and ROC.

Orange also has four out of five best results in Accuracy with RapidMiner having the best result for the Naïve Bayes algorithm. RapidMiner has four out of five best results in Specificity, with Orange holding the best result for the KNN algorithm and lastly Orange holds four best results for Sensitivity with KNIME holding the best result for Sensitivity for the SVM algorithm, which is the only best result of KNIME from all the tests performed. All things considered, we can conclude that RapidMiner is the best platform for classification algorithms if the speed of execution is the most important feature and Orange is the best platform if speed is not so important but better performance metrics other than speed are preferred.

5 RELATED WORK

There are papers studying, analyzing and comparing Open Source Data Mining Tools. However many of them are not very useful because they do not provide any conclusion on which tools are the best based on their analytical or experimental evaluation.

(Abbott and Elder, 1998) made a very complete study on seventeen different Data Mining tools existing at the time. They listed algorithms, general properties, qualities and distinctive features between them. However, all the tools analyzed were proprietary which meant all of them require a great investment to be acquired and used. (Goebel and Gruenwald, 1999) created a scheme based on tools general characteristics, database connectivity and data mining tasks available to perform their comparison on Data Mining Tools. They studied a total of forty-three different tools. This made their study very broad but incomplete and difficult to understand specially if we consider the fact that they gave no indication of what platforms are the best. In addition, a few of the tools they analyzed were just research prototypes that are not available for users or enterprises to test and use. (Wahbeh et al., 2011) analyzed four open source Data Mining tools and tested their various classification algorithms with many datasets with different characteristics but their study is very generic and not focused on a specific area of interest such as business, industry or other. (Borges et al., 2013) tested four tools with many different datasets to see how the algorithms behave when the numbers of instances and attributes change but only the results of the accuracy metric were analyzed.

(Fernández, 2014) makes an actual complete listing of technologies and tools for Data Mining. It goes all the way from the programming paradigms, the processing frameworks to the Data Mining Tools. However it does not provide though any kind of information that may be useful to help make choices of what tools to use. (Jović, 2014) provides an appropriate complete description of six tools for generic Data Mining, lists generic characteristics, and makes a comparison of what tools are more widely used in real scenarios and projects. Also the paper lists the tools features very thoroughly, both for basic and advanced Data Mining tasks. However, the paper does not give any final considerations about which ones are the best. (Hasim and Haris, 2015) also provides a very complete description of five tools for forecasting and a list of real life scenarios where each tool has application and use. However like (Jović, 2014), (Hasim and Haris, 2015) does not give any opinion or advice on what is the best platform among the ones analyzed.

Our paper improves on the aforementioned articles because it narrows its focus on the benefits of Open Source Data Mining Tools for businesses which generate great amounts of data but do not have a budget to acquire commercial products for Data Mining. We also provide detailed results on diverse performance metrics of the algorithms based on multiple tests run on the tools. Most importantly, we use those results to give recommendations about the platform we consider the best to use in Small and Medium Enterprises environments based on our experiments. This is a major extension over our previously published papers, (Almeida and Bernardino, 2015) and (Almeida and Bernardino, 2016), that give only a general overview of Data Mining Tools but provide no experimental evaluation at all.

6 CONCLUSIONS AND FUTURE WORK

In this study we evaluated three open source Data Mining tools, RapidMiner, Orange and KNIME, for five different Data Mining algorithms for Classification. We can conclude that among the studied platforms, RapidMiner is the best one for Small and Medium Enterprises.

Out of the three tools analyzed, RapidMiner is the one that gives the best ratio between execution times and accuracy of results. It is the one that is fastest in executing the majority of the algorithms studied - it has the fastest execution time for four out of five algorithms tested. It is not the best for the majority of the other performance metrics analyzed, but the difference between its values on those metrics and those of the other two platforms is small. It should be noted that Orange is not a platform to be discarded right away. As we discussed in Section 4, Orange is the platform that delivers the biggest number of best results (of 47 performance metrics analyzed, Orange collects the best results for 29 of them, while RapidMiner is the best for 16 of them and KNIME is the best for only 2), although it is slower in executing the algorithms and, in our opinion, not so intuitive to use. Further study is necessary to draw a definitive conclusion on which one of these three platforms is actually the best one.

For future work, we plan to test the platforms analyzed in this paper in a real business environment. We are also interested in testing more features and their usability for other application areas beyond business. We also aim to test additional open source platforms available for Data Mining.

REFERENCES

Abbott, D. Elder, J. (1998) A Comparison of Leading Data Mining Tools. Fourth International Conference on Knowledge Discovery & Data Mining, New York.

Alexander, D. Data Mining. [Online] Available from http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/ [Accessed: 2nd December 2015].

Almeida, P., Bernardino, J. (2015) Big Data Open Source Platforms. IEEE International Congress on Big Data, pp. 268-275.

Almeida, P., Bernardino, J. (2016) A Survey on Open Source Data Mining Tools for SMEs. New Advances in Information Systems and Technologies, Volume 444 of the series Advances in Intelligent Systems and Computing, pp. 253-262.

Borges, C. L., Marques, M. V., Bernardino, J.. (2013). Comparison of data mining techniques and tools for data classification. C3S2E '13 Proceedings of the International C* Conference on Computer Science and Software Engineering, pp. 113-116.

Demšar, J., Curk, T. & Erjavec, A. (2013) Orange: Data Mining Toolbox in Python; Journal of Machine Learning Research, 14. p. 2349-2353.

Dogan, N. & Tankrikulu, Z. (2012) A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. Information Technology and Management, 14 (2). p. 105-124.

Fawcett, T. (2006). An introduction to ROC analysis. Journal Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition, 27(8). p. 861-874.

Fayyad, M. U., Piatetsky-Shapiro, G. and Smyth, P. (1996) Advances in knowledge discovery and data mining. p. 1-34. American Association for Artificial Intelligence, Menlo Park, CA.

Fernández, A., Río, S., López, V., Bawakid, A., Jesus, M. J., Benítez, J. M., & Herrera, F. (2014) Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks. WIREs Data Mining Knowledge Discovery, 4. p. 380-409.

Goebel, M. & Gruenwald, L. (1999) A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations Newsletter, Vol. 1, No., 1, pp. 20-33.

Grzymala-Busse, J, W. & Marepally, S, R. (2010) Sensitivity and Specificity for Mining Data with Increased Incompleteness. Artificial Intelligence and Soft Computing. Volume 6113 of the series Lecture Notes in Computer Science. p. 355-362.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., Dougherty, E, R. (2010). Small-sample precision of ROC-related estimates. Bioinformatics, Vol. 26, No., 6, pp. 822-830.

Hand, D, J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve, Vol. 77, No., 1, pp. 103-123.

Hasim, N. & Haris, A. N. (2015) A study of open-source data mining tools for forecasting. IMCOM '15 Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. Article nº79.

Jović, A., Brkic, K. and Bogunovic, N. (2014) An overview of free software tools for general data mining. 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). p. 1112 – 1117.

KNIME. [Online] Avaliable from http://www.knime.org [Accessed: 2nd December 2015].

Lichman, M. (2013). UCI Machine Learning Repository [Online] Available from http://archive.ics.uci.edu/ml [Accessed: 2nd December 2015] Irvine, CA: University of California, School of Information and Computer Science.

Medri, D. (2013) Big Data & Business: An on-going revolution. [Online] Available from http://www.statisticsviews.com/details/feature/5393251/Big-Data--Business-An-on-going-revolution.html [Accessed: 30th November 2015]

O'Brien, J. A. and Marakas, G. M. (2011) Management Information Systems, 10th Edition, McGraw-Hill, New York, USA.

Petre, R. (2013). Data Mining Solutions for the Business Environment. Database Systems Journal, 4 (4), p. 21-29.

Powers, D. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001. School of Informatics and Engineering, Adelaide, Australia.

Rajagopal, S. (2011). Customer Data Clustering Using Data Mining Technique. International Journal of Database Management Systems (IJDMS), 3 (4), p. 1-12.

RapidMiner. [Online] Available from http://rapidminer.com [Accessed: 2nd December 2015].

Shen, D., Ruvini, J. & Sarwar B. (2012) Large-scale item categorization for e-commerce. CIKM '12 Proceedings of the 21st ACM International Conference on Information and Knowledge Management. p. 595-604.

Wahbeh, A., Al-Radaieh, Q., Al-Kabi, M., & Al-Shawakfa, E. (2011) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence. p. 18-26.

Witten, H. I., Frank, E. & Hall, A. M. (2011) Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition. Morgan Kaufmann, Massachusetts.