Journal of Cardiovascular Disease Research

ISSN: 0975-3583, 0976-2833 VOL11, ISSUE 11, 2020

A Review On Data Analytics Tools And its Application in Data Mining

[1] Mr. P. Sivakumar, **[2]** Mrs. A. Kalaivani, **[3]** Mrs.C. Meera Bai, **[4]** Mrs. S. Menaka

Department of Computer Applications, Nehru Institute of Information Technology and Management, Coimbatore, Tamilnadu, India.

***Abstract*—** Research works often require one or more analytical tools for problem solving, optimization, prediction, system modelling, data analysis or interpretation etc. In a typical academic research work or any professional research assignment, at some point the researcher looks for a suitable analytical tool to progress the work before arriving at some inferences. **Analytical tools are essential for processing, analyzing, and visualizing data to derive actionable insights.** Data mining plays a vital role in the contemporary society and the corporate world as a whole. This paper reviews several different data mining tools including Environment for Knowledge Analysis(WEKA), Konstanz Information Miner (KNIME), GhostMiner, R Analytical Tool To Learn Easily (Rattle), and RapidMiner. More often than not, young researchers face the challenge of choosing a data mining tool to carry out their research. This paper also evaluates the capabilities, attributes, and sources. The strengths and weaknesses of these tools have also been explored. It was established herein, that Waikato Environment for Knowledge Analysis(WEKA), Konstanz Information Miner (KNIME), R Analytical Tool To Learn Easily (Rattle), and RapidMiner are open-source data mining tools and are provided under the GNU GPL licenses while GhostMiner is commercial**.**

Keywords—Classification, Clustering, Data mining; Open source

I. INTRODUCTION

The late 20th century and the 21st century have witnessed a growing importance of information technology in the social and corporate environment. In this regard, different aspects of information technology are needed in decision support, business performance management, query and reporting analytical processing, and predictive analysis. One area of information technology that has witnessed growing importance in decision support and predictive analysis in the current world is data mining. Research and studies indicate that data mining plays a significant role in helping businesses evaluate data and as such make informed decisions regarding different aspects of their processes and operations, [4]. Speaking from this perspective, this paper will evaluate several data mining tools. These include Waikato Environment for Knowledge Analysis(WEKA), Konstanz Information Miner (KNIME), GhostMiner, R Analytical Tool To Learn Easily (Rattle), and Rapid Miner. The fact is that for each research problem or situation, one has to judiciously select an analytical tool that is best suitable for it. But usually if we hold a hammer then all problems seem to be a nail to us. This crucial research tool selection should not be influenced by our own expertise (hammer), wrong suggestions from colleagues or latest trend in tools found in literatures et al.

II. CRITICAL REVIEW OF DATA MINING TOOLS

1) Waikato Environment for Knowledge Analysis (WEKA)

WEKA is one of the most recognized data mining and machine language software. In reference to [3], WEKA was established in 1992 under the funding of the New Zealand government. This software is renowned for its data mining capabilities and is based on JAVA programming platform.One of the main capabilities of WEKA is data pre-processing. With this in mind, WEKA is able to transform raw data into a format that is understandable. According to [3], this data mining tool has a wide array of data pre-processing tools or

rather filters that enable users to perform different functions on data. These filters or tools include Add classification, Add ID, Add values, Attribute reorder, Interquartile range, Kernel filter, Numeric cleaner, Numeric to nominal, Partitioned multi-filter, Propositional to multi-instance and vice versa, Random subset, RELAGGS, Reservoir sample, Subset by expression and Wavelet. Similarly, the software has data classification capabilities. To achieve this, the software uses four main steps to classify data namely data preparation, selecting classify and applying algorithm, generating trees and analyzing the output or results, [11]. In addition to this, WEKA has data clustering capabilities, through unsupervised algorithms. Among the methods that WEKA uses to cluster data include K-means clustering, hierarchical clustering and density based clustering. Similarly, this software supports several data files including ARFF format, CSV, LibSVM's format, and C4.5's format, while WEKA 3.6 supports import of PMML regression, [3]. Moreover, it can be able to read data from databases that supports JDBC driver, files and URLs. With regard to its user interface, WEKA data mining software has four main features namely the Explorer, the Experimenter, Knowledge

ISSN: 0975-3583, 0976-2833 VOL11, ISSUE 11, 2020

flow and Simple CLI. Similarly, WEKA has extensibility capabilities that are implemented using 3 plugins. Importantly, WEKA is an open source, is distributed freely under the GNU General Public License, and if fully portable.

RapidMiner

RapidMiner is a Data Mining Suites (DMS) data mining tool. While most DMS data tools are commercial, RapidMineris an open source software, and is distributed under the GNU Affero General Public License (AGPL). RapidMiner is Java- based and has capability for web mining, data mining and text mining, [10]. RapidMiner has an array of features. In reference [9], some of the feature of this data mining software include multiple user interfaces that are easy to use, easyaccessibility and managing of data, model import and export, an operator for applying models to data sets known as scoring,data partitioning, data replacement, graphs and visualization, attribute generations and Bayesian modelling among others. Similarly, [10] notes that some of the text mining operators include tokenization, extraction, stemming, transformation, utility and other operators such as creating, writing, reading and processing documents. In this case, RapidMiner has text mining and processing capabilities. With reference to [7], RapidMiner has more than 400 operators for data mining in Java. In addition to this, this software has a user friendlygraphical user interface (GUI), drag-and-drop features, andhas options with application wizard that help process data automatically depending on the project objectives. For example, this software can help process data based on churn analysis, direct marketing, and sentiment analysis. Importantly, while RapidMiner is an open source data mining software, it also have a number of versions that are commercial available.

2) Konstanz Information Miner (KNIME)

Konstanz Information Miner (KNIME) is an open source data mining and data analytics, integration and reporting platform, and it is based on Eclipse and written in Java. KNIME is released and provided under GPLv3 license [2]. For users whoneed dedicated support, there is a commercial version of KNIME, which is more appropriate for large businesses and organizations. Nonetheless, most users would still find the freeversion of this software useful in their data mining activities. Importantly, the software has a graphical user interface that helps users to easily and quickly assemble nodes for dataprocessing, that is, extraction, transformation and leading (ETL), in order to model, pre-process and visualize data [2]. KNIME has capabilities that enable loading of data, processing, analysis, transformation and visual exploration module that does not necessarily concentrate on any area of the application. Some of the main capabilities of KNIME include frequent updates, versatility and adaptability. According to [6], KNIME exhibit its capability in adaptability through its ability to work with data prepared in other softwaresuch as Python, R and Java. Whereas this is the case, it is vital to note KNIME's capability to read data from sources is limited to ARFF files, that is, the WEKA files, databases and text-delimited files, which comprise CSV files [1]. KNIME has no capability to read data from Microsoft Excel files,LIBSVM files, and from SVMlight files. In addition, it does not use SPSS files. Speaking from this perspective, individualswho are already using LIBSVM and SVMlight file formatswill experience difficulties in using this software.

3) GhostMiner

GhostMiner is a data mining software that was developed by Fujitsu and is available commercially. According to [8], GhostMiner supports an array of data and data formats including mature machine learning algorithm and databases (including spreadsheets). The software plays a vital role in data selection and preparation, visualization, validating of models and multimodels. There are a number of key features of this software. To begin with, GhostMiner has the capabilitytools, data modelling tools and methods such as cross- validation and Xtest to evaluate the accuracy of data under study. However, this tool is limited when dealing complex data mining process and it has limited support as compared to similar products from companies such as IBM and Oracle. Therefore, the software is recommended for small and midsizeusers rather than large companies.

4) R Analytical Tool To Learn Easily (Rattle)

Rattle is an open source data mining software that is written in R programming language and provides a link into R, and is commercial. According to [13], Rattle utilizes a Gnome graphical user interface, which is implemented through the RGtk2 package. Importantly, this software's graphical userinterface was created using an interactive interface builder, Glade, and as such, the software's interface is independent of the XML description. Rattle supports a number of data formats such as CSV, TXT, ARFF, R datasets and ODBC databases connection via a number of data sources such as

ISSN: 0975-3583, 0976-2833 VOL11, ISSUE 11, 2020

MySQL, Oracle, Teradata, SQLite, SQL Server, IBM DB2, Microsoft Access and Excel and Postgress among others [13]. Similarly, it is essential to note that Rattle has two refined tools, Latticist and GGobi, which support interactive graphical analysis of data. On one hand, the GGobi tool plays a significant role in exploring high-dimensional data using its highly interactive and dynamic graphics such as bar charts, tours, parallel coordinate plots and scatterplots [13]. Importantly, GGobi need to be installed separately on the systems and supports environments such as Windows, OS/X and Linux. On the other hand, the Latticist package uses a graphical interface with lattice graphics to analyze data, and the tools supports data selections annotations, plots, brushing and sub setting among others [13]. In addition to this, theRattle is able to perform a number of functions includingselecting, exploring, graphics, transformation, clustering, associating, modelling and evaluation of data.

III. DISCUSSION

The capabilities and features of these tools are asfollows:

1. Programming language
   - *WEKA* – Java
   - *RapidMiner* – Java
   - *KNIME* – Java
   - *GhostMiner*- Not specified
   - Rattle- R

to process data from a number of file formats including2. Capabilities

spreadsheets, text and ASCII files. Importantly, the software has attributes that enable its users to develop simple user interfaces depending on their data mining needs [12]. Similarly, some of its main capabilities include data visualization and data pre-processing capabilities. The datapre-processing capabilities in GhostMiner enable this softwareto conduct most statistical functions in preliminary statistical data analysis such as calculating the mean and median across the whole data on text, ASCII and database files, [12]. With regard to visualization, GhostMiner has the capability to show how a set(s) of data is related to the other. In addition to these capabilities, this data mining software has a range of statistical

- *WEKA* – Data preprocessing, Classification, Dataclustering etc.
- *RapidMiner* – Pre-processing, Classification, text mining,visualization, application wizard.
- *KNIME*- Data transformation, visualization, frequent dataupdate, versatility etc.
- *GhostMiner* – Pre-processing, visualization, modelvalidation etc.
- *Rattle* – Data transformation, association, classification ,clustering, interactive graphical analysis, etc.

3. Supported file format
   - *WEKA* – ARFF, CSV, C4.5, etc.
   - *RapidMiner* – CVS, DB, Spreadsheet, SPSS etc.
   - *KNIME* – ARFF
   - *GhostMiner* – Spreadsheet, text and ASCII files.
   - *Rattle* – CVS, ARFF, R dataset, ODBC database.
4. Supported data source
   - *WEKA* - URLs, Database, JDBC, etc.
   - *RapidMiner*– URLs, database, SPSS, etc.
   - *KNIME* –None specified
   - *GhostMiner*–None specified
   - *Rattle* – MySQL, Oracle, MS Access, MS Excel, etc.
5. Available sources
   - *WEKA* – Free open source
   - *RapidMiner* – Free and commercial open source
   - *KNIME* – Free open source
   - *GhostMiner* – Commercial
   - *Rattle* – Commercial

ISSN: 0975-3583, 0976-2833 VOL11, ISSUE 11, 2020

- orange

Table1: Summary of strengths and weaknesses of thedata mining tools.

Data mining tools	Strengths	Weaknesses
WEKA	• Robust datamining techniques • Support webservices	• Lack of colourful visualization
Rapid Miner	• User friendly • Applicati onwizard • High level visualization support • Support webservices	• Need for programming knowledge to useto effectively use the software
KNIME	• Adaptability • No need for programming knowledge	• Can only supportARFF file format • No support for web services
GhostMiner	• Wide range ofstatistical and data modeli ngtechniques	• Cannot not handle complex data • Only for small or medium business
Rattle	• Active usergroup	• Not meant fornovice or learners
Orange	• Open source	• Use friendly

Orange tool;

Orange is a perfect machine learning and data mining software suite. It supports the visualization and is a software-based on components written in Python computing language and developed at the bioinformatics laboratory at the faculty of computer and information science, Ljubljana University, Slovenia. As it is a software-based on components, the components of Orange are called "widgets." These widgets range from preprocessing and data visualization to the assessment of algorithms and pred Data comes to orange is formatted quickly to the desired pattern, and moving the widgets can be easily transferred where needed. Orange is quite interesting to users. Orange allows its users to make smarter decisions in a short time by rapidly comparing and analyzing the data.It is a good open-source data visualization as well as evaluation that concerns beginners and professionals. Data mining can be performed via visual programming or Python scripting. Many analyses are feasible through its visual programming interface(drag and drop connected with widgets)and many visual tools tend to be supported such as bar charts, scatterplots, trees, dendrograms, and heat maps. A substantial amount of widgets(more than 100) tend to be supported. The instrument has machine learning components, add-ons for bioinformatics and text mining, and it is packed with features for data analytics. This is also used as a python library.

ISSN: 0975-3583, 0976-2833 VOL11, ISSUE 11, 2020

Orange can read documents in native and other data formats. Orange is dedicated to machine learning techniques for classification or supervised data mining. There are two types of objects used in classification: learner and classifiers. Learners consider class-leveled data and return a classifier. Regression methods are very similar to classification in Orange, and both are designed for supervised data mining and require class-level data. The learning of ensembles combines the predictions of individual models for precision gain. The model can either come from different training data or use different learners on the same sets of data.

Tableau:

Tableau is a Business Intelligence tool for visually analyzing the data. Users can create and distribute an interactive and shareable dashboard, which depict the trends, variations, and density of the data in the form of graphs and charts. Tableau can connect to files, relational and Big Data sources to acquire and process data. The software allows data blending and real-time collaboration, which makes it very unique. It is used by businesses, academic researchers, and many government organizations for visual data analysis. It is also positioned as a leader Business Intelligence and Analytics Platform in Gartner Magic Quadrant.

Charts, graphs, pictures, and maps are all examples of data visualization that most of us are familiar with. While data visualization can help distill complex data, it can also oversimplify information which can create challenges in decision-making.

R programming

R is the leading analytics tool in the industry and widely used for statistics and data modeling. It can easily manipulate your data and present in different ways. It has exceeded SAS in many ways like capacity of data, performance and outcome. R compiles and runs on a wide variety of platforms viz -UNIX, Windows and MacOS. It has 11,556 packages and allows you to browse the packages by categories. R also provides tools to automatically install all packages as per user requirement, which can also be well assembled with Big data.

IV. CONCLUSION

In conclusion, data mining plays a critical role in the modern day decision making process in the business environment. There are a number of data mining software that business can utilize to process and analyze data. These includeRattle, GhostMiner, KNIME, RapidMiner and WEKA. Arguably, while Rattle, KNIME, RapidMiner and WEKA are open source software and provided for free under the GNU GLP licenses, GhostMiner is commercial and can only be obtained at a fee. In addition, while Rattle, KNIME, RapidMiner and WEKA can perform high level data mining functions, GhostMiner is limited to low level data mining functions. A summary of the strengths and weakness of these data mining tools is shown in table 1. Data analytics is the science of analyzing raw data to make conclusions about that information. Data analytics help a business optimize its performance, perform more efficiently, maximize profit, or make more strategically-guided decisions.

REFERENCES

[1] Barcaroli G., Bergamasco S., Jouvenal M., Pieraccini G., Tininini L. Generalised software for statistical cooperation, 2008, [Online]. Available www.researchgate.net/profile/Giulio_Barcaroli/publication/263923216_Generalised_software_for_statistical_cooperation/links/0f31753c52fc97f19d000000.pdf [Accessed 29 March 2015].Akhtar, F. and Hahne, C. Rapid Miner 5 Operator Reference. Rapid-I GmbH, 2012, Retrieved 13:15, February 13, 2015 son, Leif E. "K-nearest neighbor." Scholarpedia 4(2), 1883, 2009.

[2] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. 'The WEKA data mining software: an update'. ACM SIGKDD Explorations Newsletter, 11(1), 10–18,

ISSN: 0975-3583, 0976-2833 VOL11, ISSUE 11, 2020

2009.

[3] Joseph, M. V. 'Significance of data warehousing and data mining in business applications'. International Journal of Soft Computing and Engineering, 3(1), 329–333, 2013.

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

[5] Lin, N. 'Applied business analytics: Integrating business process, big data, and advanced analytics'. New Jersey: Pearson Education Ltd., 2014.

[6] Melin, P., Kacprzyk, J. and Pedrycz, W.'Soft computing for recognition based on biometrics'. Berlin Heidelberg: Springer Science & Business Media, 2010.

[7] Pierre, T. F. Software applications: Concepts, methodologies, tools, and applications: concepts, methodologies, tools, and applications. Hershey, PA: IGI Global, 2009.

[8] Rapid-I GmbH. (n.d.) Fact Sheet: RapidMiner and RapidAnalytics – Business analytics fast and powerful [Online]. Available from: http://www.rapidi.com/downloads/brochures/RapidMiner_Fact_Sheet.pdf [Accessed 29 March 2015].

[9] Shterev, J. 'Demo: Using RapidMiner for Text Mining'. Digital Presentation and Preservation of Cultural and Scientific Heritage, 3, 254–256, 2013.

[10] Singhal, S. and Jena, M. 'A study on WEKA Tool for data pre- processing, classification and clustering'. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2 (6), 250–253, 2013.

[11] Wang, J., Hu, X., Hollister, K. and Zhu, D. 'A comparison and scenario analysis of leading data mining software'. in Selected readings on information technology management: Contemporary issues. ed. by Kelley, G. Hershey, PA: IGI Global, 2008.

[12] Williams, G. J. 'Rattle: A data mining GUI for R.' The R Journal, 1(2), 45–55, 2009.