Data Mining Tools: A Comparative and Analytical Study

Dr. Sandeep Aggarwal[1]
[1] DAV College, Department of Computer Science & Applications, Abohar, India
Email: sandeepaggarwal10@gmail.com

Abstract — **Today the fast development of information technology and its variation among applications has transformed the business and other fields considerably. All the Enterprise tends to access a huge amount of data and information on daily basis. With such amount of data, every enterprise will require powerful techniques for mining of most valuable data pattern and effective understanding of this data that is beyond the human's ability for making decision in a better way. This inevitability also requires of a suitable data mining tool to help in decision making and query processing. The paper provides the complete analysis of various open source data mining tools regarding the features, functionalities and specialization for each tool along with its applications. By employing this study, the assortment of tools can be made effective.**

Index Terms — **Data, Data Mining, Data Mining Tools, KNIME, Open Source Tools, Orange, Technical Specification, Weka.**

I. Introduction

Data mining (DM) is at the core for knowledge discovery from datasets. It combines all the analysis measures that are needed to divulge new and significant information to an interested user. It includes data preparation and data modeling. Datasets can be obtained from a variety of sources like traditional relational databases, data stores, web pages, local data files, cookies and many more. It is imperative to organize data in the most competent way in order to mine maximum possible information [1]. After preparation, a number of models can be realized, depending on the research goal. In order to properly understand the models, various evaluation and procedures are followed. The visualization of the results is also desired. There are a number of tools that were developed by an enthusiast research community of Data Mining; that are provided at no cost using open-source licenses. Various Free and openly available tools for DM are being developed from last two decades.

The goal of these tools is to smooth the progress of the rather complex data analysis process and provide all interested scholars a free option in place of commercial data analysis platforms. Data mining tools envisage imminent trends and behaviors that allow business houses to make more practical, knowledge driven decisions. The growth and solicitation of data mining algorithms requires us to use one of software tools. Here we focus on the analysis of a number of such tools that have grown more proficient and useful over the years, some of them even akin or better in various aspects as compared to their commercial complements. Here in this paper the main characteristics of Weka [2]. KNIME [3], RapidMiner [4], and Orange [5] are outlined and compared.

All of them provide implementations of general Data Mining tasks. The tools are compared on the basis of common features like language, license etc. core DM tasks like supported input, data preparation, data analysis, output etc. and support for some recent more advanced Data Mining topics like big data, data streams, text mining and learning. As the number of available tools is increasing day by day, the choice of most appropriate tool is becoming difficult. The most popular four open source tools that can be used for data mining are briefed as below.

II. Weka

WEKA is Waikato Environment for Knowledge Analysis. It is a collection of open source of various data mining as well as machine learning algorithms that include preprocessing, classification and clustering to visualization and feature selection [10]. These algorithms can either be used directly on data set or we can call it from our own code. The Weka workbench contains a set of abundant tools for visualization and algorithms for analysis of data and predictive modeling. It supports GUI for easy access. It supports .arff (attribute relation file format) file format.

Weka is the tool that is mostly used due to its gigantic functionality and various supported features. This java based tool provides user with both CLI and GUI for carrying out and managing tasks to be performed. Weka provides three graphical user interfaces, one is the Explorer for fact-finding data analysis to support preprocessing, feature selection, learning, conception: second is the Experimenter that runs experimental environment for analysis and assessing machine learning algorithms, and the third is Knowledge Flow used for new process models that provide inspired interface for graphical design of KDD process. Weka also provides a simple Command-line explorer interface for typing commands [8].

A. Technical Specification:

- It was developed by Department of Computer Science, University of Waikato, New Zealand.
- It was first released in 1997.
- The Latest version available is WEKA 3.6.11.
- It is having GNU general public license.
- It is supported by Java.
- It is having GUI as well as command line interface.
- It is available at www.cs.waikato.ac.nz/m1

B. Main features:

- It has 49 data processing tools, 76 classification/regression algorithms, 8 clustering

algorithms and 3 algorithms for finding association rule and many more.
- Weka has provisions for reading of files from numerous different data bases.
- Using Weka we can import the data over internet, from various web pages or from some remotely located SQL database server just by providing the URL of resource.
- Weka is also appropriate for mining association rules.
- Weka is Stronger in machine learning techniques.
- Weka is Suited for machine Learning.

C. ADVANTAGES

- It Possible to apply WEKA tools to Big Data.
- It is also appropriate for evolving new machine learning patterns.[9]
- It provides very flexible environment for random combination of search and estimation methods.
- Weka is free, extensible and relatively easier to use.
- Weka loads data file in various different formats like CSV, C4.5, ARFF and binary. It can also be assimilated into other java packages.
- Weka software contains a GUI making it very easy to access.
- It is having a very large collection of diverse data mining algorithms.
- Models can be built using a graphical user interface or a command line input.

D. LIMITATIONS

- It lacks proper and adequate documentations.
- Its CSV reader is also not as good as it is in Rapid Miner.
- It is not as polished.
- Weka does not support much to the visualization of concluded data.
- Weka is much weaker in classical statistics.
- It does not have the Descriptor scaling facility to pertain to future datasets.
- It does not provide any automatic facility for Parameter optimization of various methods.
- It Has the Descriptor Selection facility but not the part of knowledge flow.
- It has limited facility to partition of data sets to training and testing sets.

III. KNIME

Konstanz Information Miner commonly referred as KNIME is an open source integration platform for data analytics and reporting. This tool allows easy accessibility to new nodes that are to be added into the workflow. It incorporates various components for machine learning and supports data mining with its modular data pipelining approach. It provides us the GUI which allows assembly of nodes for data preprocessing for modeling and data analysis and visualization [12]. Using this we can modify features of a particular node and execution of partial data flow. Various features in KNIME 2.0 provide support for loops, database connection manipulations which further enrich KNIME's capabilities to make it a powerful data exploration and analysis environment with a strong integration backbone that allows for easy access to a number of other data processing and analysis packages [11].

Knime, pronounced as “naim”, is an striking DM tool that is run inside the IBM’s Eclipse development environment and is easily extensible. It is like a modular data investigation platform that facilitates the user to visually create data flows, and selectively accomplish various analysis steps, and later look into the results using interactive views on data as well as models [12]. User defined nodes and types can be generated in KNIME in the span of a few hours thus outspreading KNIME to realize and provide basic support for highly domain-centric data. It has not only been used in medicinal research, but can also be used in various fields like customer data breakdown, business aptitude and financial data analysis.

A. TECHNICAL SPECIFICATION

- It was implemented in java.
- The Latest version available is KNIME2.9.1.
- It is Licensed By open source GNU General Public License 3.
- It is GUI.
- Knime runs on Windows, Mac OS X and Linux.

B. MAIN FEATURES

- It provides an open API system that allows for new nodes to be added to the application in a way that makes integration not only quite easy, but also allows for an efficient way of adding information and functionality to the application[12].
- It allows us to attach Chemistry Development Kit along with additional nodes that support the processing of various chemical structures, compounds, etc.
- It is specialized in analyzing data about Enterprise reporting, Business Intelligence and data mining.
- Knime is a widely used open source, data mining, and reporting workbench used by numerous organizations.

C. ADVANTAGES

- It integrates various analysis modules Weka data mining environment and additional plug-ins that allow various user scripts to be run and providing access to a gigantic library of statistical procedures [9].
- KNIME also provides a unique database port system that allows users to create database connections with almost any JDBC compliant database.
- It is based on the widely used Eclipse IDE platform, making it as much a development platform as a data mining platform.
- Most of the standard data mining methods are included in Knime.
- Knime provides the visualization of data, results, and processes that is intended to be simple for users.
- It has the descriptor scaling facility.
- KNIME provides support for a variety of statistical analysis where statistical functions from basic to more advanced linear models, data clustering and data trees can be performed [12].

- It requires no installation making it easy to try out.
- It has the ability to interface with programs that allow visualization and analysis of molecular data as per requirement.

D. LIMITATIONS

- Have the limited facility to partition of data sets to training and testing sets.
- Have only limited error measurement methods.
- It does not provide facility for Parameter optimization of various methods.
- Have no wrapper methods for descriptor selection.

IV. RAPINMINER

It is possibly the most commonly used open source data mining platform with over 3 million downloads. It is an integrated environment for machine learning, data and text mining, predictive and business analytics. It provides various data mining and machine learning procedures like data loading, Mining, transformation, load (ETL), data preprocessing and conception, evaluation, and deployment [9]. It is majorly used for Industrial applications, research, training, rapid prototyping and application development for various business houses[8]. Rapid Miner implements a client/server model where the server is presented as Software as a Service or it can be on cloud infrastructures.
The graphical user interface and visualization tools are tremendous, with substantial intelligence built into the workflow construction process. This provides very fast error recognition and advocated quick fixes. Its metadata transformation capability is unique among various tools allowing results to be scrutinized at design time. Rapid Miner supports majority of databases, so that users can import information from a diversity of database sources [8]. Various Business solutions that require predictive analysis and statistical computing prefer using it. It integrates numerous operators and the WEKA machine learning library.

A. TECHNICAL SPECIFICATION

- It was released on 2006 by RapidMiner, Germany.
- The Latest version available is Rapid miner 6.
- It is licensed by AGPL Proprietary.
- It is GUI.
- RapidMiner is written in the Java programming language.
- It is Cross platform that can be installed on any operating system and is Language Independent.
- It can be downloaded from www.rapidminer.com.

B. MAIN FEATURES

- It provides a number of learning schemes for regression classification and clustering analysis.
- It is a new approach to design rather complex problems by using a modular operator concept.
- It uses XML to depict the operator trees for representation of knowledge discovery process.
- It provides various flexible operators for data input and output file formats.
- Rapid miner supports a number of file formats. [10]
- Rapid Miner provides a number of learning algorithms from WEKA that you program by piping various components in a graphic ETL work flows.

C. ADVANTAGES

- It provides model evaluation using cross validation and independent validation sets.
- Most data sources are supported including Excel, Access, Oracle, IBM DB2, Microsoft SQL Server, text files and others.
- Over 1,500 methods for data integration, data transformation, analysis and, modeling as well as visualization, no other solution on the market offers more procedures and thus defining the optimal analysis procedures.
- Rapid Miner suggests Quick Fixes to make an illegal work flows as legal.
- RapidMiner offers numerous procedures for attribute selection and outlier detection.
- In RapidMiner, Users hardly have to write any code.
- It provides learning scripts, prototypes and algorithms from WEKA and R scripts.

D. LIMITATIONS

- Rapid Miner is the data mining software package suitable for persons having familiarity in working with database files of academic settings or in business settings because the software requires the ability to handle SQL statements and files.
- It uses large amounts of memory and so it often obtains the errors.
- It is not very user friendly, so the use of the tutorial is almost necessary.

V. ORANGE

Orange is a component-based machine learning suite for explorative data analysis, Python bindings and libraries for scripting by supporting visual programming front-end. It is implemented using C++ Python. Its GUI was built on the cross-platform framework [9]. A large number (over 100) of widgets are supported. This includes data transformation, ordering, regression, connotation, visualization and unsupervised learning methods. There are also some specialized add-ons covering bioinformatics, text mining and other specialist requirements.
It is an open source conception and analysis tool having simple interface. Most of the analysis can be done through its visual programming interface and most visual tools are supported including various scatterplots, bar charts, and heatmaps etc. It is having additional components for machine learning and add-ons for bioinformatics and text mining. Also it is packed with features for data analytics.

A. TECHNICAL SPECIFICATIONS

- The latest version available is Orange 2.7.
- It is licensed by open source GNU General Public License 3.
- It is an open source data mining package built with Python and other languages.
- It runs on Windows, Mac OS X and Linux
- It is GUI as well as Command Line.
- It is available at www.orange.biolab.si

B. MAIN FEATURES

- It is a data mining and machine learning software suite based on components.
- It provides tools for data preprocessing, data representation, model evaluation, and exploration techniques.
- Data mining in Orange is done using either visual programming or Python scripting.
- It provides Open source data visualization and analysis for beginner as well as experts.
- It has a Cross platform GUI.

C. ADVANTAGES

- It can Work both as a script as well as an ETL work flow GUI.
- It generates shortest script for doing training, algorithms comparison and prediction etc.
- Orange the easiest tool to learn.
- It is specialized for data visualization along with mining.
- Orange is written in python hence is easier for most programmers to learn.
- It provides a better debugger as compared to other tools.

D. LIMITATIONS

- It is not super polished.
- The installation is large and cumbersome since you need to install QT.
- It supports a Limited list of machine learning algorithms.
- There is no uniform Machine learning between the different libraries.
- Its Reporting capabilities are restricted to exporting only visual representations of data models.
- Orange does not provide best possible performance for association rules.

VI. RESULTS AND DISCUSSIONS

Among the four data mining packages that we have been studied, KNIME is among the package that can be suggested for people who are novices to data mining software as compared to those who are very much skilled. The software is simply very robust having many in-built features and additional functionality that can be otherwise obtained from third-party libraries. Based on the study, Weka is considered a very close to KNIME due to various built-in features that require no programming or coding knowledge Whereas Rapid Miner and Orange are considered suitable for sophisticated users, predominantly those in the hard sciences, because of the added programming skills that are desired, and the inadequate visualization support that is provided. It can be concluded that even if data mining is the basic concept to all of these tools yet, Rapid miner is the only tool that is independent of language restraint and has statistical and extrapolative analysis capabilities, So it is easily used and implemented, moreover it integrates majority algorithms of other mentioned tools.

VII. FUTURE SCOPE

We have discussed a number of DM tools in this work. The nutshell conclusion is that there is no such tool that can be individually considered best. Each tool has its own strong points and weaknesses. Open-source data mining suites of at present have developed a lot where they were only a decade ago. They offer good graphical interfaces, which provide usability and interactivity, they support extensibility by using interfaces for add-on modules. They offer flexibility either through visual programming within GUI or by using scripting languages. With the recent comings and goings of various developers concerning the use of tools in various fields one can expect a more enhanced environment along with a number of technical improvements. The work can be a useful insight to develop an application with more efficiency and availability i.e. A tool can be designed which instead of supporting a specific area can be extended to more fields. The effort can be increased to a certain extent and the development may be a multifaceted procedure but indeed it can result in an efficient product.

REFERENCES

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

[2] 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.

[3] 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.

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

[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] Witten, I.H., Frank, E.: "Data Mining: Practical machine Learning tools and techniques", 2nd edition, Morgan Kaufmann, San Francisco (2005).

[7] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[8] Kalpana Rangra and Dr. K. L. Bansal, "Comparative Study of Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014.

[9] Manika Verma and Dr. Devarshi Mehta , "A Comparative study of Techniques in Data Mining", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 4, April 2014.

[10] A. Jović, K. Brkić and N. Bogunović, "An overview of free software tools for general data mining".

[11] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias K¨ otter, Thorsten Meinl, Peter Ohl, Kilian Thiel and Bernd Wiswede, "KNIME – The Konstanz Information Miner", SIGKDD Explorations, Volume 11, Issue 1.

[12] Arpita M. Hirudkar and Mrs. S. S. Sherekar, Comparative Analysis of Data Mining Tools and Techniques for Evaluating Performance of Database System, Vol. 6, No.2, International Journal Of Computer Science And Applications.