www.ijarcsse.com

Volume 4, Issue 5, May 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Comparative Study of Data Mining Tools and Techniques for Appraising Enactment of Database System

Abhilash Pandey[*]
M.Tech.(CSE), Invertis University, India

Deepak Kumar Pathak
M.Tech.(CSE), Invertis University, India

Garima Gupta
M.Tech.(CSE), Invertis University, India

Abstract*—Data mining, the abstraction of hidden foretelling information from enormous databases, is a potent technology with great potential to Service Company's emphasis on the most essential information in their data warehouses. Currently, enormous amount of data and information are available, Data can now be stored in various different types of databases and information repositories, being accessible on the Internet. A potent technique is a need for better interpretation of these data that exceeds the human's ability for comprehension and making decision in an enhanced approach. There are data mining, web mining and knowledge discovery tools and software packages such as RapidMiner tool, WEKA tool, NetTools, KNIME tool and Orange. The work deals with analysis of RapidMiner tool, WEKA tool, NetTools, KNIME tool and Orange. There are numerous tools available for data mining and web mining. Therefore awareness is essential about the quantitative investigation of these tools. This paper focuses on numerous practical, functional, rational as well as analysis facets that users can be observing for in the tools. Whole study statements the efficacy and importance of these tools including numerous facets. Analysis acknowledges various benefits of these data mining tools along with wanted aspects and the features of present tools.

***Keywords*— Data Mining, Databases, Data Warehouses, KDD, Data Mining Tools.**

I. INTRODUCTION

The Data Mining [1] is an Extraction of hidden, foretelling information from enormous databases .It is also knows as Knowledge Discovery from Databases (KDD).It performs an Identification and estimation of hidden patterns in database. It is great technology with great potential to aid organizations to discover and produce information from their data warehouses. Data mining tools forecast behaviours and future trends. It support organization to take positive knowledge driven decisions, they organize databases for identifying hidden patterns and also automates the discovery of relevant patterns in a database, using well-defined approaches and algorithms to look into current and historical data that can then be analysed to predict future trends. There are number of data mining tools are available to mine such kind of data. So, it has become rather difficult for an anonymous user to select the finest possible data mining tool for his work. In this paper presents an summary of data mining with the phases involved in mining data and the various kind data mining methods and it also delivers the reader the comparisons study of various easily and freely available data mining tools such as RapidMiner tool, WEKA tool, NetTools, KNIME tool and Orange for web mining accessible currently with their own strengths and weaknesses.

II. DATA MINING

The Data mining states to extracting or “mining" knowledge from enormous amounts of data. It is also knows as Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the method of robotically searching enormous volumes of data for patterns such as association rules. It applies various older computational techniques from statistics, information reclamation, machine learning and pattern recognition [1][2]. There are following data mining steps:

- Data Cleaning: This is the first step of data mining. In this step; data that contain empty or corrupted records are removed.
- Data Integration: In order to continue with data mining, data must be collected and integrated into a single formatted structure. However, different sources of data generally do not deliver uniform structures and interpretations of data; so integration into a single format requirements to take place.
- Data Selection: Not everything of the data collected are needed though. Data selection consents for electing only such data that are significant to the task to be performed.
- Data Transformation: The data that have conceded the cleaning phase are still not prepared for data mining, so there still need to be transformed into format accepted by the data mining algorithm.
- Data Mining: In this step, several algorithms may be applied on the data in order to determine prospective knowledge hidden within the data.

- Pattern Evaluation: The significance of results provided by data mining essentials to be evaluated, for not all of the findings may be of concern to the inquiry. Redundant patterns are consequently removed.
- Knowledge Presentation: Results that seem to be the most essential undergo transformation and visualization in order to be presented in the most understandable form [1][3].

A. Data Mining Methods

- Classification: Supervised Learning. There are classes known
- Clustering: Unsupervised Learning. There are classes unknown
- Association Rule Mining: Detecting the hidden, previously unknown relation between the entities.
- Temporal mining: Use with temporal data, forming temporal events, time series, pattern recognition, sequences and temporal association rules are certain tasks.
- Time Series Analysis: Define the trend, nature and behaviour of time series data. Forecast the future trend and behaviour of the data.
- Web Mining: There mining web data; Web content mining, Web structure mining and Web usage mining.
- Spatial Mining: Use with GIS for mining knowledge from spatial database. Spatial classification and clustering and rule generation are certain task under this mining [4][6].

III. DATA MINING TOOLS

In this section specific open source data mining tools are mentioned

A. RapidMiner Tool

RapidMiner [8] delivers data mining and machine learning procedures with: data loading and transformation, data pre-processing, visualization, modelling, estimation, and deployment. RapidMiner is written in the Java language. It practices learning schemes and attributes evaluators from the Weka machine learning environment and statistical modelling schemes from R-Project.

B. Weka Tool

WEKA is Waikato Environment for Knowledge Analysis, data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand [5]. It is a group of open source of various data mining and machine learning algorithms, with pre-processing on data, Classification and regression, clustering, association rule extraction, feature selection. It supports .arff (attribute relation file format) file format

C. NetTools Spider - The Web Mining Spider

A web spider [4] is a software package that searches the Internet for information. The simple process of a web spider is to download a web page and to search the web page for links to other web pages. It then repeats this process in all of the new pages that it found. By repeating this process a web spider can find all of the pages within a web site and all of the pages on the Internet.

D. KNIME

KNIME stand for the Konstanz Information Miner. It is an open source data analytics, integration platform and reporting. KNIME integrates numerous components for machine learning and data mining over its modular data pipelining model. A graphical user interface lets assembly of nodes for data preprocessing (ETL: Extraction, Transformation, Loading), for modelling, data analysis and visualization.

E. ORANGE

Orange is a component-based data mining and machine learning software package, featuring a visual programming front-end for explorative data analysis and visualization as well as Python bindings and libraries for scripting. It contains a set of components for data preprocessing, feature scoring and modelling, filtering, model assessment, and exploration techniques. It is implemented in C++ and Python.

IV. PERFORMANCE ANALYSIS OF THE TOOLS

A. The RapidMiner Tool

In respect of other open-source data mining software suites that have been examined thus far, RapidMiner has complete API support, which makes it promising to access a wide range of functionality and support [11].

1) RAPIDMINER API

It has been well-known that the API functionality within RapidMiner is quite robust, which permits users to interface with other applications and functions without the requirement to concern about the particular details that permits the interfacing to occur.

2) RapidMiner-Database System Support

In addition, RapidMiner delivers support for most kinds of databases, which means that users can import information from a variety of database sources to be inspected and analysed within the application. As with other data mining applications, the origin for the database functions is SQL queries. However, it does look likely that with the origin for database support being SQL queries, several limitations might exist in terms of how data can be imported into the software and how databases can be attuned. Fortunately, RapidMiner has been formed with additional functionality that permits fewer advanced users to be able to import databases and perform changes to those databases with compact programming and coding knowledge. Once again, however, it should be well-known that even with these increased functions to make importation and management of database files easier, the software does commonly need at least a medium level of knowledge about database files and about SQL queries. So this means that the capability to successfully transform database files with regards to updating or deleting row and column information would need at least few small programming knowledge.

3) RapidMiner-Visualization

In terms of visualization support for data and analysis, RapidMiner delivers a high level of visualization support. It is probable within the software to produce detailed results of data analyses. In further, the visualization of nodes and additional information can be highly colourful and pleasing to the eye. In this way, RapidMiner can permit users with higher level programming and coding skills to have increased output from their efforts in terms of being able to visualize the data and results. However, without at least certain basic programming and coding skills, it looks unlikely that the complete performance of the visualization abilities of the software could be attained. The cause for this is that the visualization support within the software is associated to the functions that are performed [10].

4) RapidMiner-PMML Support

The data mining techniques that are part of RapidMiner are what would be predictable of this kind of application. Users are able to implement clustering, regression analysis, Gaussian processes, and even several more advanced processes. Attaining all of these data mining functions, however, does need a higher level of knowledge that might be required in other data mining software suites. Interestingly, with the innovative functionality of the software, PMML support is only something that has been newly added, and then through an additional extension that must be added to the basic package [11]. While the functionality is now present, it is interestingly to note that it almost looks as though this was regarded as an afterthought by the developers.

B. WEKA TOOL

WEKA is a fully functional data mining software suite that delivers a high level of functionality for users. WEKA supports various different standard data mining tasks such as data preprocessing, clustering, classification, visualization regression, and attributes selection.

1) WEKA-API

In fact, it has been well-known that the API functionality of Weka delivers users with the ability to attain increased functionality because of the various freely available programming codes that are available online. Even additional, the software contains the ability to perform over 100 kinds of data mining methods, with rule-based methods, Bayesian methods and statistical analysis. The presence of several different kinds of data mining methods in Weka makes the software useful for a variety of data mining techniques and in a variety of industries. Users of different industries are unlikely to face any inability to practice a desired data mining method with Weka [10].

2) WEKA-Database System Support

Strength of Weka is that the software natively provisions the ability to read files from a variety of database formats [9]. For users who achieve data from the internet, a specific strength of Weka is the ability to obtain data from both SQL databases and from actual webpages by entering the URL of the webpage having the information. This makes it probable for users to easily input information into the software that might not really be in a format that would make it easily read by other data mining packages.

3) WEKA-Visualization Capabilities

Weka has robust support for the use of APIs, a variety of data mining methods, and supported database systems, one of the weaknesses of the software is its visualization support. It is significant to note that the software delivers data's visualization, results, and processes, but the provision that is delivered is slightly limited. What is expected by this is that the visualization of data, results, and processes is not extremely colourful or as detailed as other data mining software packages [12]. However, the visualization that is delivered is indeed sufficient for being able to view the data on which the analyses are being implemented and the results of the data analyses efforts. In addition, add-ons are existing that can increase the visualization functionality of the software [9]. As a part of the visualization support in the software and the add-ons that are existing. Weka is able to interface with the R statistical package in order to not only increase its statistical analysis functions, but also allow for increased visualization of statistical analyses and results [12].

4) WEKA-PMML Support

Weka has support for PMML. This lets users to import PMML files that are produced in both propriety and open-source data mining and statistical software packages. However, the software does not now have support for exporting data files in the PMML format for use in other applications. This functionality is strategic for future issues of the software [10].

5) WEKA-Statistical Analysis Capabilities

Weka can implement just about any type of statistical analysis. In addition to performing the most elementary descriptive and inferential statistical analyses, the software also permits for cluster analysis to be performed. Also, as has already been well-known, Weka has the ability to interact directly with the R statistical package. This makes it probable to increase the statistical functionally of the software, as well as makes it probable for users that are further comfortable or familiar with R to use both applications to make the full range of data analysis and data mining functions that might be required for a given project [12].

C. NetTools Spider

Following are the several utilities provided by NetTools Spider:

1) Web Site Downloader

At the core of NetTools Spider is a powerful web site downloader package. It is flexible sufficient for the most demanding user, yet modest sufficient that user can download a web site with only 2 mouse clicks and a couple of key-presses. User can exactly start downloading a web site in less than 30 seconds [4].

2) Offline Browser

NetTools Spider includes a built-in web browser that makes viewing downloaded web sites a snap. User can navigate over downloaded web sites much faster than possible if they were viewed online. At a user's look can see an entire web sites' structure and pick the files user want to view.

3) Web Site Localizer

NetTools Spider will convert a web site to a localized copy making it possible to copy the web site to a CD and view it with any web browser. User can share downloaded web sites with user's friends, customers, or co-workers.

4) Web Site Search Utility

NetTools Spider can search downloaded web sites with amazing speed. It can search whole web pages or their titles, Meta tags and links.

5) Internet Search Utility

NetTools Spider can search the Internet for files having keywords. User can search thousands of web sites and only download the files that contain the words user's looking for.

6) Link Checker

NetTools Spider is capable of checking all links in a web site, with links generated in dynamic content and links to external web sites. It will then provide user a detailed list of where broken links are so they can easily be corrected.

7) Web Mining Utility

NetTools Spider's most powerful feature is that it can be used as a real-time web mining tool. With the help of simple scripts, NetTools Spider can simply extract pieces of information from a web site and store that information in a database or text file. These scripts can be written in either VBScript or JavaScript and are simple enough for most web developers to use. With its web mining features, the probable uses for NetTools Spider are almost endless.

8) Link Extraction

NetTools Spider permits user to easily export web page information and links in many different formats including Excel csv files [4].

D. KNIME

The Knostanz Information Miner (KNIME) is well-appointed with an open API system that permits for new nodes to be added to the application in a method that makes integration not only properly easy, but also permits for an efficient means of adding information and functionality to the application [14].

1) KNIME-API

The existence of the open API system does make the system more robust and beneficial than might otherwise be the case, because users can use it to improve the functionality of the software either through their own programming efforts, or through the APIs that are freely available and have been written by others.

2) KNIME-Database System Support

KNIME also has a unique database port system that allows users to establish database connections with nearly any database that is JDBC compliant. While the ability to acquire data from a large number of different types of databases is actually not unique to KNIME or most data mining software packages, what is unique is that the database port functionality allows databases to be manipulated without the need to modification SQL code. Instead, users can use the ports to secure database rows and columns. Then, the software delivers filter functionality in which users can filter or delete entire rows and columns in a database through the graphical user interface [14]. For users who are not familiar with SQL statements, or who simply want to avoid the need to edit SQL statements, KNIME provides for the ability to import and filter databases entirely through the graphical user interface.

3) KNIME-PMML Support

In terms of the data mining methods that are available in KNIME, most of the standard methods are included. Users are capable to perform clustering, rule induction, regressions, and bayes networks [14]. In addition, nearly all of the data mining algorithms that are included in KNIME support PMML. This shows that users can perform most types of data mining methods and then export the models and results to other propriety and open source applications that utilize the PMML format. It is also worth stating that just like the database port functionality, the functionality of PMML inside KNIME makes it possible for users to clean up PMML files without the need for any coding to occur. Instead, users can transform PMML files within the graphical user interface and then use the files, or export them to other software packages [10].

4) KNIME-Visualization

As with the other functions that have already been discussed, the visualization of data, results, and processes in KNIME is proposed to be simple for users. The primary workspace in the application, well-known as the Workbench, lets users to drag and drop different functions or processes so that they can be connected to further nodes. Further functionality can be added to KNIME that makes it possible to increase its visualization abilities. For example, by integrating with R statistical software package or JfreeChart, it is possible to improve the visualization of statistical functions and results. Moreover, it is possible to implement additional functionality for chemistry data so that the software is able to visualize molecular data types and the various properties of molecular structures [13].

5) KNIME-Statistical Analysis Capabilities

KNIME provides support for a large variety of statistical analysis of data. Statistical functions from basic descriptive statistics to more advanced linear models and data clustering and data trees can be performed. In addition, the ability to interface with the R statistical software package means that the statistical functionality of KNIME is greatly increased as even more advanced statistical functions can be performed. The overall result is that KNIME can be used for basic functions, which in various respects are much more than just basic within the software, or it can be included with other open-source and proprietary software to increase its functionality and performance [10].

E. ORANGE

Orange is similar to the other data mining software packages that have been examined, in terms of functions that can be performed. The dramatic difference, however, is that in order to achieve full functionality from Orange, additional add-ons, known as widgets, generally have to be obtained and added to the program. The reason for this is that Orange is actually a library of objects and routines written in C++. The basic program has the essential functionality of the processes and functions that the software is intended to perform. This is not to say that the software cannot perform advanced features. However, in order to perform advanced features, the full range of libraries and routines must be obtained [18].

1) ORANGE-API

More specifically, in order to have API functionality, further libraries and routines must be downloaded and added to the software [18]. While this might not seem to be a major issue for advanced users, it could be an obstacle for novices as they might expect the software to be fully functional when it is downloaded. Even more, because of the fact that other data mining software packages have built-in API support, it might just be assumed that Orange would also have this functionality.

2) ORANGE-Database System And PMML Support

At the same time, it seems as though the support for other database systems may be limited. The reason for this is that PMML support within the software is very limited and only available by adding additional routines and libraries to the basic software package. While the interoperability with other database formats and other database software is automatic for other data mining packages, the interoperability in Orange is manually guided. Users must be able to perform the import functions on their own with little actual assistance from the software. Along with the lack of direct PMML support, Orange also has little built-in support for other database systems. Through the basic program, the only database support is for SQL. Users must be able to understand and work with SQL documents and statements in order to import database files. Any database files in other formats are much more difficult to import into the system, if some of them can be imported at all. Once again, this demonstrates the lower level of ease of use and functionality of Orange as compared to the other data mining software packages that have been examined. For users who want to be able to easily import and work with database files in a variety of formats, Orange will likely show its functional limitations. It should be noted however, that Orange does provide support for most data mining methods. The software has the ability to perform Bayes, decision trees, and other types of data mining methods. Once again, however, the issue is not so much the functionality that is supported, but the ease of use for the user [10].

3) ORANGE-Visualization

The visualization support within Orange is slightly limited. While visualization is certainly available, and users are able to visualize data, processes, and results, the visualization is not as appealing to the eye or easy to work with as other data mining packages [15]. Users of Orange will have to expect that visualizing schemas and other functions will be somewhat limited.

4) ORANGE-Statistical Analysis Capabilities

Orange does provide support for most statistical tests and analyses that users would want to perform. In fact, because of the nature of Orange with regards to be more accessible to advanced users, it is advanced users who may find the software's statistical features and functionality most impressive. The software provides support to run various types of statistical tests and analyses and create charts and graphs for the results. The key to the success of using the software's statistical functionality would rest with the user and his or her ability to understand how to input files and information and to perform the processes required to obtain the results of various statistical functions [10].

V. CONCLUSIONS

Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, the web etc. Various free toolkits are available to understand and extrapolate data and information. This research has conducted a comparison between different data mining toolkits and web mining. The complete analysis of these data mining and web mining software tools focuses the usefulness and importance of these tools by considering various aspects. Analysis presents various benefits of these data mining tools with respect to functionalities, advantages and disadvantages, and compared them accordingly. The analysis took into account support of APIs, various database systems, PMML support, statistical analysis capabilities and visualization specific to the respective software packages. According to learning the functionality built into to Weka and available through add-ons makes the software highly robust for a variety of users. The RapidMiner are for those users who with the skills to write code or to seek out add-ons, the software can perform many high-level functions related to the process of data mining. The description of the functions of KNIME might make it seem to be an application that is intended for those who either do not have coding and programming skills or who want something that is easy to practice. NetTools Spider mining tool is mainly practice for the web mining purpose.

REFERENCES

[1] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[2] Du Mouchel, W., Volinsky, C., Johnson, T., Cortes, C., and Pregibon, D. (1999) Squashing flat files flatter. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM Press.

[3] Jiawei Han, Benjamin W. Wah, Vijay Raghavan, Xindong Wu, Rajeev Rastogi, Fifth IEEE International Conference on "Data Mining", ICDM 2005, Houston, Texas, 2005.

[4] Baker, R., Barnes, T., Beck, J.E., Educational Data Mining 2008: 1_st International Conference on Educational Data Mining, _Proceedings. Montreal, Quebec, Canada, 2008.

[5] Bouckaert, R. R.; Frank, E; Hall, M. A.; Holmes, G; Pfahringer, B; Reutemann, P; Witten, I. H.. WEKA—Experiences with a Java Open- Source Project. Journal of Machine Learning Research, vol 11, pp 2533-2541, 2010.

[6] Baker, R., Merceron, A., Pavilk, P.I., Educational Data Mining 1010: 3st International Conference on Educational Data Mining, Proceedings, Pittsburgh, USA, 2010.

[7] http://www.questronixsoftware.com/

[8] http://rapid-i.com/

[9] Hall, M; Frank, E.; Holmes, G.; Pfahringer, B., The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009, 11, pp 10-18,2009.

[10] Samuel Kovac, S 2012,1,3 http://is.muni.cz/th/255695/fi_b/suitability_analysis_of_data_mining_tools.pdf

[11] RapidMiner http://rapid-i.com/content/view/181/190/lang,en/

[12] Hornik, K.; Buchta, C.; Zeileis, A., Open-Source Machine Learning: R Meets Weka. Research Report Series 50, pp 1-7.

[13] Berthold, M. R.; Cebron, N.; Dill, F.; Di Fatta, G.; Gabriel, T. R.; Georg, F.; Meinl, T.; Ohl, P.; Sieb, C.; Wiswedel, B., KNIME: The Konstanz Information Miner: Version 2.0 and Beyond. SIGKDD Explorations 2009, 11, 26-31, 2009.

[14] Berthold, M. R.; Cebron, N.; Dill, F.; Di Fatta, G.; Gabriel, T. R.; Georg, F.; Meinl, T.; Ohl, P.; Sieb, C.; Wiswedel, B., KNIME: The Konstanz Information Miner. http://kops.ub.unikonstanz.de/ bitstream/handle/urn:nbn:de:bsz:352-opus-64456/ BCDF06_knime_ics.pdf?sequence=1.

[15] Nelson, A.; Menzies, T.; Gay, G., Sharing Experiments Using Open Source Software. Software—Practice and Experience 2002, 9, pp 1-7,2002.

[16] Wahbeh, A. H.; Al-Radaideh, Q. A.; Al-Kabi, M. N.; Al-Shawakfa, E. M., A Comparison Study Between Data Mining Tools Over Some Classification Methods. IJACSA 2011, 2, pp 18-26, 2011.