Available at: www.researchpublications.org

Comparative Analysis of Data Mining Tools and Techniques for Evaluating Performance of Database System

Arpita M. Hirudkar Mrs. S. S. Sherekar

Sant Gadage Baba Amravati University, Amravati

arpitah@gmail.com ss_sherekar@rediffmail.com

ABSTRACT

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Nowadays, large amount of data and information are available, Data can now be stored in many different kinds of databases and information repositories, being available on the Internet. There is a need for powerful techniques for better interpretation of these data that exceeds the human's ability for comprehension and making decision in a better way. There are data mining, web mining and knowledge discovery tools and software packages such as WEKA Tool and RapidMiner tool. The work deals with analysis of WEKA, RapidMiner and NetTools spider tools KNIME and Orange. There are various tools available for data mining and web mining. Therefore awareness is required about the quantitative investigation of these tools.

This paper focuses on various functional, practical, cognitive as well as analysis aspects that users may be looking for in the tools. Complete study addresses the usefulness and importance of these tools including various aspects. Analysis presents various benefits of these data mining tools along with desired aspects and the features of current tools.

KEYWORDS- Data Mining, KDD, Data Mining Tools.

1. INTRODUCTION

Data Mining [1],it is an Extraction of hidden, predictive information from large databases .It is also called as Knowledge Discovery from Databases (KDD).It perform an Identification and evaluation of hidden patterns in database. It is powerful technology with great potential to help organizations to locate and generate information from their data warehouses. Data mining tools predict future trends and behaviors. It help organization to make proactive knowledge driven decisions, they prepare databases for identifying hidden patterns and also automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends.

To mine such type of data there are number of data mining tools are available. As a result of this, it has become rather difficult for an unknown user to select the best possible data mining tool for his work. This paper presents an overview of data mining with the steps included in mining data and the different data mining methods and it also provides the reader the comparisons study of various freely available data mining tools such as WEKA tool, RapidMiner tool and NetTool Spider for web mining available today with their own strengths and weaknesses.

2. DATA MINING

Data mining refers to extracting or “mining" knowledge from large amounts of data. It is also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns such as association rules. It applies many older computational techniques from statistics, information retrieval, machine learning and pattern recognition[1][2]. Following are the data mining steps:

- Data Cleaning: In the first step, data that contain corrupted or empty records are removed.
- Data Integration: In order to proceed with data mining, data need to be collected and integrated into a single formatted structure. However, different sources of data usually do not provide uniform structures and interpretations of data; therefore integration into a single format needs to take place.
- Data Selection: Not all of the data collected are needed though. Data selection allows for choosing only such data that are relevant to the task to be performed.
- Data Transformation: The data that have passed the cleaning step are still not ready for data mining purposes, for they still need to be transformed into format accepted by the data mining algorithm.
- Data Mining: In this step, various algorithms may be applied on the data in order to discover potential knowledge hidden within the data.
- Pattern Evaluation: The importance of results provided by data mining needs to be evaluated, for

not all of the findings may be of interest to the inquiry. Redundant patterns are therefore removed.

- Knowledge Presentation: Results that appear to be the most important undergo transformation and visualization in order to be presented in the most understandable form [1][3].

2.1 Data Mining Methods

- Classification: Supervised Learning. The classes are known
- Clustering: Unsupervised Learning. The classes are unknown
- Association Rule Mining: Identifying the hidden, previously unknown relation between the entities.
- Temporal mining: Use with temporal data, modeling temporal events, time series, pattern detection, sequences and temporal association rules are some tasks.
- Time Series Analysis: Describe the trend, nature and behavior of time series data. Predict the future trend and behavior of the data.
- Web Mining: 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 some task under this mining[4][5].

3. DATA MINING TOOLS

Here in this section the open source data mining tools are mentioned

3.1 WEKA tool

WEKA is **W**aikato **E**nvironment for **K**nowledge **A**nalysis, data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand[6].

It is a collection of open source of many data mining and machine learning algorithms, including pre-processing on data, Classification and regression, clustering, association rule extraction, feature selection. It supports .arff (attribute relation file format) file format

3.2 RapidMiner tool

RapidMiner [7] provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load, a.k.a. ETL), data preprocessing and visualization, modelling, evaluation, and deployment. RapidMiner is written in the Java programming language. It uses learning schemes and attributes evaluators from the Weka machine learning environment and statistical modelling schemes from R-Project.

3.3 KNIME

KNIME, the Konstanz Information Miner, is an open source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows assembly of nodes for data preprocessing (ETL: Extraction, Transformation, Loading), for modeling and data analysis and visualization.

3.4 Orange

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

3.5 NetTools Spider - The Web Mining Spider

A web spider[4] is a software program that searches the Internet for information. The basic 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 behavior 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.

4. PERFORMANCE ANALYSIS OF THE TOOLS

4.1 WEKA Tool

WEKA is a fully functional data mining software package that provides a high level of functionality for users. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and attribute selection.

4.1.1 WEKA-API

In fact, it has been noted that the API functionality of Weka provides users with the ability to achieve increased functionality because of the many freely available programming codes that are available online. Even more, the software contains the ability to perform over 100 types of data mining methods, including Bayesian methods, rule-based methods, and statistical analysis. The inclusion of so many different types of data mining methods in Weka makes the software useful for a variety of data mining methods and in a variety of industries. Users in different industries are not likely to face any inability to use a desired data mining method with Weka [9].

4.1.2 WEKA-Database System Support

Another strength of Weka is that the software natively supports the ability to read files from a variety of database formats [10]. For users who obtain data from the internet, a specific strength of Weka is the ability to acquire data from both SQL databases and from actual webpages by entering the URL of the webpage containing the information. This makes it possible for users to easily input information into the software that might not actually be in a format that would make it easily read by other data mining packages.

4.1.3 WEKA-Visualization Capabilities

While Weka has strong 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 important to note that the software provides visualization of data, results, and processes, but the support that is provided is somewhat limited. What is meant by this is that the visualization of data, results, and processes is not highly colorful or as detailed as other data mining software packages [11]. However, the visualization that is provided is certainly sufficient for being able to view the data on which the analyses are being performed and the results of the data analyses efforts. In addition, add-ons are available that can increase the visualization functionality of the software [10]. As a part of the visualization support in the software and the add-ons that are available,. Weka is able to interface with the R statistical package in order to not only increase its statistical analysis functions, but to also allow for increased visualization of statistical analyses and results [11].

4.1.4 WEKA-PMML Support

Weka has support for PMML. This allows users to import PMML files that are created in both propriety and open-source data mining and statistical software packages. However, the software does not currently have support for exporting data files in the PMML format for use in other applications. This functionality is planned for future releases of the software [9].

4.1.5 WEKA-Statistical Analysis Capabilities

Weka can perform just about any type of statistical analysis. In addition to performing the most basic descriptive and inferential statistical analyses, the software also allows for cluster analysis to be performed. Also, as has already been noted, Weka has the ability to interact directly with the R statistical package. This makes it possible to increase the statistical functionally of the software, as well as makes it possible for users that are more comfortable or familiar with R to use both applications to perform the full range of data analysis and data mining functions that might be required for a given project [11].

4.2 RapidMiner Tool

As with the other open-source data mining software packages that have been examined thus far, RapidMiner has full API support, which makes it possible to access a wide variety of functionality and support [12].

4.2.1 RapidMiner-API

It has been noted that the API functionality within RapidMiner is quite strong, which allows users to interface with other applications and functions without the need to worry about the specific details that allows the interfacing to occur.

4.2.2 RapidMiner-Database System Support

In addition, RapidMiner provides support for most types of databases, which means that users can import information from a variety of database sources to be examined and analyzed within the application. As with other data mining applications, the basis for the database functions is SQL queries. However, it does seem likely that with the basis for database support being SQL queries, some limitations might exist in terms of how data can be imported into the software and how databases can be adjusted. Fortunately, RapidMiner has been created with additional functionality that allows less advanced users to be able to import databases and make changes to those databases with reduced programming and coding knowledge. Once again, however, it should be noted that even with these increased functions to make importation and management of database files easier, the software does generally require at least a medium level of knowledge about database files and about SQL querries. This means that the ability to successfully transform database files with regards to changing or deleting row and column information would require at least some small programming knowledge.

4.2.3 RapidMiner-Visualization

In terms of visualization support for data and analysis, RapidMiner provides a high level of visualization support. It is possible within the software to create detailed results of data analyses. In addition, the visualization of nodes and other information can be highly colorful and appealing to the eye. In this way, RapidMiner can allow 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 some basic programming and coding skills, it seems unlikely that the full performance of the visualization abilities of the software could be achieved. The reason for this is that the visualization support within the software is connected to the functions that are performed[9].

4.2.4 RapidMiner-PMML Support

The data mining methods that are part of RapidMiner are what would be expected of this type of application. Users are able to perform clustering, regression analysis, gaussian processes, and even some more advanced processes. Achieving all of these data mining functions, however, does require a higher level of knowledge that might be required in other data mining software packages. Interestingly, with the advanced functionality of the software, PMML support is only something that has been recently added, and then through an additional extension that must be added to the basic package [12]. While the functionality is now present, it is interestingly to note that it almost seems as though this was viewed as an afterthought by the developers.

4.2.5 RapidMiner-Statistical Analysis Capabilities

RapidMiner does provide for a variety of statistical tests and analyses to be performed. However, as compared to other data mining packages, it does seem that the statistical functionality, much like most of the software's functionality, can truly only be accessed and fully used by those with more advanced skills. For someone who has fewer programming skills, or who wants to use software that requires the use of less programming, RapidMiner's functionality might seem difficult to access or even inaccessible[9].

4.3 KNIME

The Knostanz Information Miner (KNIME) is equipped with an open API system that allows for new nodes to be added to the application in a way that makes integration not only fairly easy, but also allows for an efficient means of adding information and functionality to the application [13].

4.3.1 KNIME-API

The presence of the open API system does make the system more robust and useful than might otherwise be the case, because users can use it to enhance 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.

4.3.2 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 change SQL code. Instead, users can use the ports to acquire database rows and columns. Then, the software provides filter functionality in which users can filter or delete entire rows and columns in a database through the graphical user interface [13]. 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.

4.3.3 KNIME-PMML Support

In terms of the data mining methods that are available in KNIME, most of the standard methods are included. Users are able to perform clustering, rule induction, regressions, and bayes networks [13]. In addition, nearly all of the data mining algorithms that are included in KNIME support PMML. This means 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 noting that just like the database port functionality, the PMML functionality within 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 [9].

4.3.4 KNIME-Visualization

As with the other functions that have already been discussed, the visualization of data, results, and processes in KNIME is intended to be simple for users. The primary workspace in the application, known as the Workbench, allows users to drag and drop different functions or processes so that they can be connected to other nodes. Additional 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. Furthermore, 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 [14].

4.3.5 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 many respects are much more than just basic within the software, or it can be integrated with other open-source and proprietary software to increase its functionality and performance [9].

4.4 Orange

Orange is very similar to the other data mining software packages that have been examined, in terms of the 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 [17].

4.4.1 Orange-API

More specifically, in order to have API functionality, additional libraries and routines must be downloaded and added to the software [17]. 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.

4.4.2 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 [9].

4.4.3 Orange-Visualization

The visualization support within Orange is somewhat 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 [16]. Users of Orange will have to expect that visualizing schemas and other functions will be somewhat limited.

4.4.4 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 [9].

4.5 NetTools Spider

Following are the various utilities provided by NetTools Spider:

4.5.1 Web Site Downloader

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

4.5.2 Offline Browser

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

4.5.3 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. You can share downloaded web sites with your friends, co-workers, or customers.

4.5.4. Web Site Search Utility

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

4.5.5 Internet Search Utility

NetTools Spider can search the Internet for files containing keywords. You can search thousands of web sites and only download the files that contain the words you're looking for.

4.5.6 Link Checker

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

4.5.7 Web Mining Utility

NetTools Spider's most powerful feature is that it can be used as a real-time web mining tool. With the use of simple scripts, NetTools Spider can easily 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 possible uses for NetTools Spider are almost endless.

4.5.8 Link Extraction

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

5. CONCLUSION

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. Acording to study 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 use. NetTools Spider mining tool is basically use 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] Baker, R., Merceron, A., Pavilk, P.I., Educational Data Mining 1010: 3st International Conference on Educational Data Mining, Proceedings, Pittsburgh, USA, 2010.

[6] 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.

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

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

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

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

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

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

[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. http://kops.ub.unikonstanz. de/bitstream/handle/urn:nbn:de:bsz:352-opus-64456/BCDF06_knime_ics.pdf?sequence=1.

[15] 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.

[15] 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.

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