Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

Detailed Analysis of Data Mining Tools

Rohit Ranjan
Pre-Final year student
Dept. of Computer Science& Engg.
Dayananda Sagar College of Engineering,
Bangalore-78

Swati Agarwal
Pre-Final year student
Dept. of Computer Science& Engg.
Dayananda Sagar College Of Engineering,
Bangalore-78

Dr. S. Venkatesan
Professor
Dept. of Computer Science& Engg.
Dayananda Sagar College Of Engineering,
Bangalore-78

Abstract: **- Today the magnificient growth of technology and adoption of the several application renaissance in the information technology sector and the related fields.Due to this striking advancement ,collecting and warehousing the data in necessity.This overall leads to the concept of data mining,which can be viewed as one of the emerging and promising technology development. Data mining is explanation and analysis of large quantities of data in order to extract implicit, previously unknown and potentially meaningful patterns by using some tools and techniques. This paper presents the comprehensive and theoretical analysis of five open source data mining tools –Rapidminer, R, Knime, Orange, Weka. The study provides the pros and cons Ziped with the technical specifications features and specialization of each tool.By this complete and hypothetical study, the best slelection of the tool can be made easy.**

Keywords: Data Mining, Open Source, Dataset, R, Rapidminer, Knime , Orange, Weka.

INTRODUCTION:

In this information age, with the advent of technology advances and means for mass digital storage, users typically collect and store all varieties of data, counting on the power of technology to help sort through this amalgam of information. These massive collection of data were initially stored on disparate structures, leading to the creation of the structured databases. The efficient database management system (DBMS) have been very essential and crucial assets for management of large corpus of the data. The proliferation of DBMS has also contributed to massive gathering of varieties of information. Confronted with huge collection of data, the need for the hour is to make letter managerial choices. These emergent needs are:

- Automatic summarization of the data.
- Extraction of the essence of information stored.
- Discovery of the pattern in raw data.

Data Mining is most suitable answer for all the above mentioned emergent needs. Data Mining is the computational process of discovering patterns in data stored in large data repositories or data warehouse, involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The core step of data mining is to mine and discover the novel information in terms of pattern and rules from large volumes of data. The key idea behind data mining is to design and work efficiently with the large dataset. There is non exclusive list of variety of information collected in digital form in datasets :-Businuss Transaction,Scientific Data, Medical and Personal Data, Survilance Video and Pictures, Satellite Sensing,Games and Virtual Works (using CDA) , Text Reports and memos (e-mail message), World Wide Web Repositories.

To acquire the sequence and trends in data, Data mining use multiplex algorithm and mathematical analysis(for efficient discussion making).Data mining is frequently treated as synonym of knowledge discovery in databases (KDD) process.The following figure shows data mining as a step in an iterative KDD process.

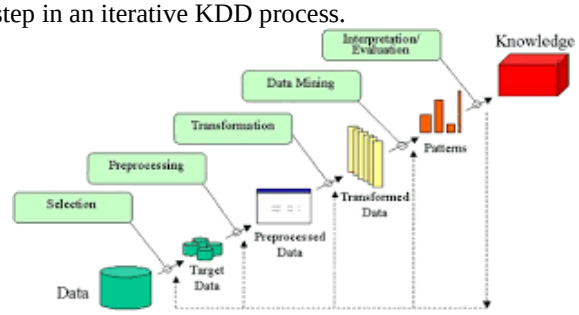


Fig 1:Knowledge Discovery in Database Process

The knowledge discovery in the database process comprises of a few steps leading from raw data collection to some form of new knowledge, The iterative process consists of following steps :

- Data Cleaning :- also known as data cleansing , it is a phase in which noise data and irrelevant data are removed from the collection.
- Data integration:- at this stage multiple data sources , often heterogeneous , may be combined in common source.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

- Data selection:- at this step , the data relevant to analysis is decided on and retrieved from the data collection.
- Data transportation:- also known as data consolidation, it is a phase in which the selected data is transferred into forms appropriate for the mining procedure.
- Data mining:- it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation:- In this step , strictly interesting pattern representing knowledge are identified based on given measure.
- Knowledge Representation:- it is the final phase in which the discovered knowledge is visually represented to the user . This essential step uses visualization technique to help user understand and interpret the data mining results.

The kinds of pattern that can be discovered depend upon the data mining task employed. There are two types of data mining tasks. Descriptive Data Mining- describe the general properties of existing data. Productive Data Mining Tasks - Attempt to do predictions based on inference on available data. The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list :-

- Characterization: - Data characterization is a summarization of general features of objects (exhibited regularly) in a target class, and produces characteristics rules.
- Discrimination :- Comparison of the target class with one or a set of comparative class (often called as contrasting classes) .It produces discriminant rules.
- Association Analysis :- Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.
- Classification :- Classification is data analysis task used for finding a set of models where a model is constructed to predict categorical class tables.
- Prediction :- Prediction is supervised learning task that is similar to classification used to predict hidden unavailable data values or predict a class label for some data.
- Clustering :- Clustering is also called as unsupervised classification, because the class label are known. It is the method of making a group of elements of objects with similar characteristics.
- Outline Analysis :- Outline are data elements that can not be grouped in given class and cluster, also known as exception or surprises. They are often very important to identify.
- Evolution and derivation analysis :- It pertains to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which content to characterizing, comparing or classification of time related data.

Derivation analysis considers difference between measured value and expected values and attempts to find the cause of deviation. Data mining system can be categorized according to various criteria among other classifications are following:-Classification according to the type of source mined : special data , multimedia data , time series etc.Classification according to the data model drawn: relational database, object oriented database, data-warehouse, transactional etc.Classification according the kind of knowledge discovered: based on data mining functionality discovered – classification, clustering etc.Classification according to mining technique used – based on data analysis approach used such as machine learning, neural networks, genetic algorithms etc,Classification based on degree of user interaction query driven systems, interactive exploratory systems etc.

There are mainly three different categories of data mining tools. The categories are as follows: Traditional Data Mining tools-Tools that work with the existing database stored in enterprise services. These tools are used to analyze the large amount of data that is already stored in broad categories in order to reveal sub patterns. Application Based Tools-Application based tools are easy to use and help in administrative work and provide services for company performance. In this historical data is represented as reference and these tools check the current trends in order to see the changes in the business. Web Based Data Mining Tools-This type of tools are called text-mining tools because of its ability to mine various kind of text from any written source. It also helps for scanning and converting data in selected format which is compatible with any tool.

OVERWIEW OF DATA MINING TOOLS

Data can generate revenue. It is a valuable asset of an enterprise. Businesses can use data mining for knowledge discovery and exploration of available data. This can help them predict future trends, understands customer's preferences and purchase habits, and conduct a constructive market analysis .Data mining helps enterprises to make informed business, decisions, enhances business intelligence, thereby reducing the cost overheads.

Due to its widespread use and complexity involved in building data mining applications , a number of innovative and intuitive tools have emerged over decades to fine-tune data mining concepts in a bid to give companies more comprehensive insight into their own data with useful future trends.Some of the popular open source tools available for data mining are briefed below :-

1. R

R is an environment for statistical computing and graphics. This is written in C and FORTRAN, and allows data miners to write scripts just like a programming language. R was founded in 1993 in New Zealand and the first version was released in 1947. It has a general public license and is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

supported by windows, Macros and variants of Linux.R provides a wide variety of statistical (Linear,non linear modelling, classical statistical tests, time series analysis, classification, clustering etc) and graphical techniques and is highly expensible.R is available as free software for data manipulation, calculation and graphical display. R-cloud is the paid-version of R.

R tool provides an effective data handling and storage facility. It consists of a suit of operators for calculations on analysis in particular matrices .It is a large, coherent, integrated collection of intermediate tools for data analysis .Apart from the facilities mentioned above it also provides graphical facilities for data analysis and display either on screen or on hardcopy .R uses a well-developed simple and effective programming language which includes conditional loops , user-defined recursive functions and input and output facilities. It provides data visualization and analysis up to 16TB .

R has outstanding graphical capabilities along with a fully programmable graphics package. It has a very extensive statistical library.R plays well with tools for importing or exporting data, eg:- SAS,SPSS or directly from Microsoft excel .It can produce graphics output in PDF , JPG ,PNG and SVG formats , and output for LATEX and HTML .R has active groups where question can be asked and are quite quickly responded to by the developers of this environment .It has the ability to make machine learning program in just 40 lines of code.R has a steep learning curve -it takes a while to get used to the power R Documentation is sometimes patchy and impenetrable to non statistician.

2. RAPIDMINER

It is data science software platform developed by the company of the same name that provides an integrated environment for machine learning , deep learning , text mining and predictive analytics written in java , it is a readymade , open source no coding required software. Rapid miner is an open source software which works well in the java runtime environment. It shows good performance with the cross-platform operating systems. It's first version was released in 2006.The latest version of Rapid miner was released on 13th February 2017.Rapidminer is Licensed by AGPL website.

Rapidminer provides a visual, code-free environment. It allows the user to work with different types and sizes of data source. Since it runs in java runtime environment, it is platform independent. It also acts as a powerful scripting language engine along with a geographical users. It provides Multi –layered data view for better data analysis. It is typically used for Machine learning, Data mining Predictive analytic.

Rapidminer provides support for most types of databases, which means that users can import info from a variety of database sources .It has the full facility for the model evaluation using cross validation and independent validation sets .Along with the outstanding facility of model evaluation it offers numerous procedure, specially in the area of attribute selection and defining the optimal analysis process .It can be easily integrated with WEKA and R-tools to directly give models from script written in the former two. To utilize the awestruck facilities provided by rapidminer tool one must subscribe to a tutorial. It uses large amount of memory – most often error obtained is "out of memory". It is most suited for people who are accustomed to working with database files.

3. WEKA

Weka is a freely available mining software written in java. it was developed at the university of the Waikato in New Zealand .it is an environment for executing the necessary steps in data mining , including pre-processing of the data and built a productive model. It includes algorithm and tools for clusters analysis, classification , regussion analysis , visualization and feature selection.It is an open source software licensed under GNU general public license. Weka is basically a collection of machine learning algorithms. It works fine on a multiprogram operating system. First version of Weka was released in 1997 which included the work environment with frontend written in Tcl.

Explorer is the main user interface of the weka. It is used for primitive tasks of data mining including data pre-processing, classification, regression, clustering, allocation rules, and utilization. It can also execute data file in multiple formats. One exceptional feature of weka is database connection using JDBC with any RDBMS package.

As WEKA is fully implemented in JAVA programing language, it is independent and portable .WEKA software contains very suitable Graphical user Interface, so that the system is easy to access.There is very large collection of different data mining algorithms. Weka data mining permits companies to find out where their finest selling points are and give them chance to consider benefits of details.Weka is weaker in classical statistic. It doesnot have proper documentation as R. It does not have automatic facility for parameter optimization of machine learning/statistical methods. It does not provide connectivity to Excel spreadsheet and non java based database. It is unable to save parameters for scaling to apply to future datasheet.

4. ORANGE:-

Orange is component-based visual programing software package for data visualization, machine learning , data mining and data analysis. Orange components are called widgets. Visual programing is implemented through an interface in which workflow are created by linking predefined or user-designed widgets.

The core components of orange tool are written in C++ with wrappers in python/cython. Orange makes use of Python open source libraries for scientific computing. It is supported on mac OS, windows and linux platform.In the other words we can say that it works well with the cross-platform operating systems .It is lincensed under GNU General Public License.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

Orange consists of canvas interface onto which the user places widgets and creates a data analysis window. Widgets offer functionalities such as reading the data and visualizing data elements. Data mining in orange is done through visual programming or python scripting. Orange also includes a set of components for data preprocessing, feature scoring, filtering, modeling and exploration techniques.

Orange is the easiest tool to learn with a better debugger which enhances the debugging of the code. Scripting the data mining categorization problems is simpler in Orange as it remembers the choices and suggests the most frequently used combinations. It also has features for different visualizations such as scatterplots ,bar charts, trees , networks and heatmaps. Orange is weak in classical statistics and it does not have automatic parameter optimization of machine learning/statistical methods. It has limited methods for partitioning of dataset into training and testing sets.

5. KNIME

Knime the Konstanz information miner, is an open source data analytics, reporting and integration platform . knime integrates various component for machine learning and data mining through its modular data pipelining concepts .A graphical user interface allow assembly of nodes for data preprocessing(ETL : extraction , transformation , loading), for modeling and data analysis and visualization. Knime is written in java and is supported Linux, macOS and windows.The first version of Knime was released in January 2004.It is licensed by GNU General Public License. Knime has been used in pharmaceutical research but also used in other areas like CRM customer's data analysis, business intelligence and financial data analysis.

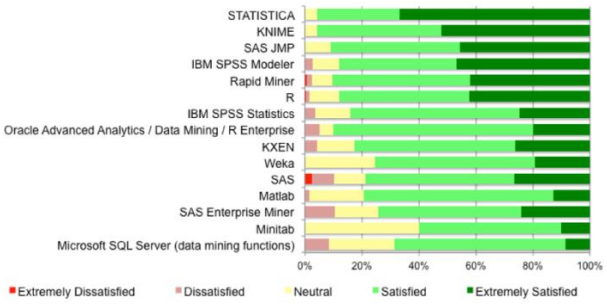
Knime provides scalability through sophisticated data handling. It has an intuitive user interface. In Knime parallel execution is possible on multi-core systems. It integrates all analysis modules of the well known datasets. It is easy to try out because it does not require installation besides downloading and unarchiving.

Knime has only limited number of error measurement methods. Similar to orange even knime does not have wrapper methods for descriptor selection. Along with that it does not contain facility for automatic parameter optimization of machine learning or statistical methods.

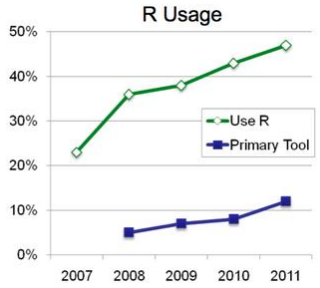
Comparision of open source data mining tools

Comparision Factors	R (Revolution)	Rapid Miner	Weka	Orange	Knime
Initial Release	August 1993	2006	1997	1997	2004
Stable Release	April 21,2017	13 Feb,2017 (7.4)	April 14, 2016 (3.8.1)	6 March, 2017(3.4)	7 April, 2017 (3.3.2)
License	GNU General Public License	AGPL	GNU General Public License	GNU General Public License	GNU General Public License
Operating Systems	Cross Platform	Cross Platform	Windows, OS X,Linux	Cross Platform	Windows, OS X,Linux
Language	C,FORTRAN, R	Language Independent	Java	Python, Cython, C,C++	Java
Partitioning of Dataset into training And testing dataset	Limited methods For partitioning	Limited methods	Limited Methods	Limited Methods	Limited Methods
Descriptor Scaling	Can save parameters to scale future datasets	Can save parameters to scale future datasets	Cannot save parameters to scale future datasets	No scaling methods	No scaling methods
Descriptor Selection	No wrapper methods	Has wrapper methods	Has wrapper methods	No wrapper methods	No wrapper methods
Parameter optimization Of machine learning or Statistical methods	Does not have automatic parameter optimization	Has automatic parameter optimization	Does not have automatic parameter optimization	Does not have automatic parameter optimization	Does not have automatic parameter optimization
Model Validation using Cross validation or Independent validation set	Limited error measurement Methods	Plenty of error measurement methods	Has error measurement methods but needs to rebuild the model each time	Has error measurement methods but needs to rebuild the model each time	Has error measurement methods but needs to rebuild the model each time
Website	www.r-project.org	rapidminer.com	www.cs.waikato.ac.nz/~ml/weka	Orange.biolab.si	www.knime.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017



Graph 1: Satisfactory level of users for different data mining tools



Graph 2: Compared to any of the primitive tools, R is an efficient tool for beginners.

CONCLUSION:-

Over a decade, the steep advancement in the field of technology has revolutionized the holistic view of data mining tools. Today, they offer nice graphical interfaces, focus on usability and interactivity ,support extensibility through augmentation of the source code. The study presented the detailed analysis of five open source data mining tools enlisting the technical specifications garnished with the features offered by each of them. The main aim of this relative study is to emphasize on the rocketing demand for skilled and advanced data miners.Based on the analysis Knime is the package that would be recommended for people who are novices to such software to those who are highly skilled.Weka would be considered very close to KNIME because of its many built-in features that require no programming or coding knowledge. Rapid Miner and Orange would be considered appropriate for advanced users, particularly those in the hard sciences, because of the additional programming skills that are needed, and the limited visualization support that is provided. The focus of the study is to enhance the learning of data mining tools for the beginners with an additional vital perspective of providing them an insight for future developments in this era of technology.

REFERENCES:

[1] A. Komathi1, T.Ramya 2, M. Shanmugapriya3, V. Sarmila4 A Novel Comparative Study on Data Mining Tools,IJIRCCE-16(vol 4,issue 11).

[2] Neha Chauhan1, Nisha Gautam PARAMETRIC COMPARISON OF DATA MINING TOOLS.2nd international conference on recent innovations in science ,Engineering and management(ICRISEM-15 ISBN-978-81-931039-9-9.

[3] Kalpana Rangra ,Dr. K. L. Bansal Comparative Study of Data Mining Tools,International Journal in Advanced Research in Computer Science And Software Engineering-IJARCCE-14(vol 4,Issue 6,June,2014)

[4] Snehal A. Deshmukh , Data Mining Tools: Review, INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY, 3(9).

[5] P.S. Patel, S.G. Desai , Comparative Study on data Mining tools, International Journal of Advanced Trends in Computer Science and Engineering, 4(2), April 2015.

[6] S. Srivastava, WEKA: A Tool for Data Preprocessing, Classification, Ensemble, Clustering and Association Rule mining, International Journal of Computer Applications, 88(10), February 2014.

[7] S. R. Mulik, S. G. Gulawani :“ PERFORMANCE COMPARISON OF DATA MINING TOOLS IN MINING ASSOCIATION RULES”, International Journal of Research in IT, Management and Engineering (IJRIME),

[8] T. Silwattananusarn, K. Tuamsuk, Data mining and its application for Knowledge management: A literature review from 2007 to 2012, International Journal of Data Mining & Knowledge Management Process (IJDKP), 2(5), September 2012.

[9] Ralf Mikut and Markus Reischl Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011.