International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 41-46
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

AN OVERVIEW OF DATA MINING TOOLS

Parmeet Kaur
Department of Computer Science
Punjab Institute of Technology (PTU)
Kapurthala, India

Parmjeet Kaur
Department of Computer Science
Punjab Institute of Technology (PTU)
Kapurthala, India

Abstract*—data mining term refers to the finding of relevant and helpful data from database. Data mining scours databases for hidden patterns, finding predictive information and trends that experts may miss, as it goes past their desires. When implementation on a high performance parallel processing computer or client server, data mining tools can analyze massive databases to deliver answer to questions such as which clients to use most likely to respond to the next promotional mailing. This paper mainly focuses on the review of several tools that have grown more efficient and helpful over the years, some even comparable or better in certain aspects. The tools are compared in terms of general properties (language, license, development team etc.) as well discuss which type data can be mined with data mining tools. *(Abstract)

*Keywords-**data mining, database, tools, Weka, Rapid Miner, Orange, R, and KNIME.***

I. INTRODUCTION

Data mining [1] (DM), is the extraction of understood and potentially information from huge and is the main step in learning revelation in datasets. It incorporates every one of the examination techniques that are needed with a specific end goal to uncover most recent and pertinent data to an interested customer. Data mining incorporates data arrangement and information model Datasets may be obtained from various sources, including: traditional database, relational databases, data warehouses, transaction database, multimedia databases web documents, or simple local textual files.

It is vital to get ready data in the well-organized way in order to extract as much information as possible. After arrangement, different models can be built, depending on the research objective. In order to properly interpret the models, standard evaluation and statistical measures are pursued. Free and openly available software tools for Data mining have been being developed for as far back as 20 years the primary goal of these tools is to facilitate the rather complicated data analysis process and to offer every interested scientist a free distinct option for business information examination stages. They do so mainly by proposing integrated situations or particular packages on top of standard programming languages, which are frequently open source

II. DIFFERENT TYPES OF DATA THAT CAN BE MINED

A. Flat documents

Flat documents are really the most widely recognized information hotspot for data mining at the exploration level. Flat files are basic information documents as plain content or binary. There is unstructured relationship between the records. In a relations database a documents contain a record for each line. In a record, the different columns are delimited by a comma or tab to isolate the fields. Contrasting a relational database, it does not contain multiple tables. The information in these records can be experimental estimations, exchanges, and time-arrangement reports and so on.

B. Relational Databases

A relational database characterizes as an arrangement of tables containing either estimations of attributes or values of entity attributes from entity relationships. Tables have rows and columns, where rows represent tuples and columns represent attributes. Each table record (or row) holds a unique data defined for a relating column category.

C. Data Warehouses

Information distribution center (DWH) data ware house, otherwise called an endeavor information stockroom is a store of information gathered from numerous information sources or heterogeneous information sources and is planned to be utilized all in all under the same united schema. A data warehouse gives the choice to analyze data from various sources under the same roof.

D. Transaction Databases

A transaction database is an arrangement of records speaking to exchanges, every record with a period stamp, an identifier and an arrangement of things. Supplementary with the transaction files could also be expressive information for the items. For instance, in the case of the audio and video store, the rentals table

E. World Wide Web

Data in the World Wide Web is organized in inter-connected documents. These documents can be raw data, text, audio, video, and different applications. Uniquely, the World Wide Web is contained three noteworthy segments: The contents of the Web, which incorporates records accessible; the structure of the Web, which lives up to expectations with the hyperlinks and the relations in the middle of archives; and the use of the web, describing how and when the data are accessed.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 41-46
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

F. Multimedia Databases

Multimedia media databases characterize as a database which incorporates sound and video, pictures and content media. They can be stored in a file system or extended or object-oriented databases. Data mining from multimedia repositories may require image interpretation, computer vision, computer graphics and natural language preparing procedures.

G. Spatial Databases

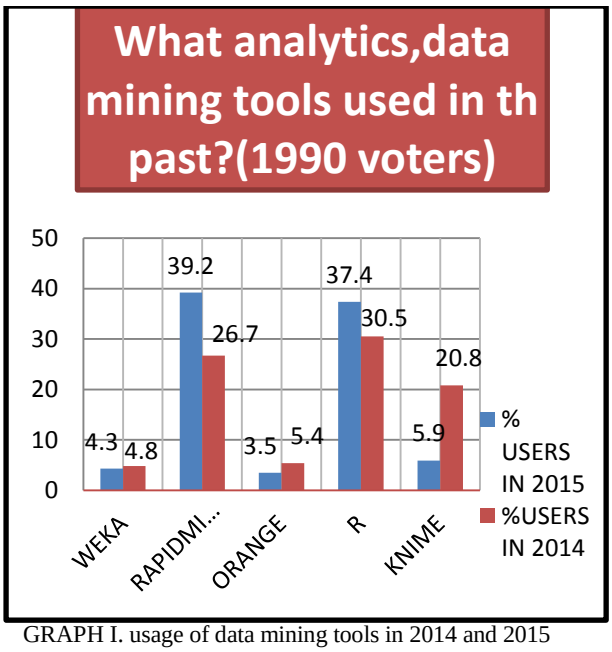
Spatial databases are databases that store geographical information like worldwide or local situating and maps. Such spatial databases exhibit new difficulties to data mining calculations.

H. Time-Series Databases

Time-series databases are a database which contain time related data such stock market data or logged activities. These databases normally have a constant stream of new information coming in, which some of the time causes the crucial for a challenging real time analysis. Data mining in time series databases usually incorporates the study of patterns and relationships among developments of distinctive variables, and in addition the expectation of new patterns and developments of the variables in time.

III. DATA MINING TOOLS

Various types of data mining tools are available in the business, each with their quality and shortcomings. The greater part of the tools has usage for Mac, Windows, Linux and OS X working frameworks. General characteristics like developer, programming language, license, and current version, GUI (graphical user interface), main purpose and Community support of the six DM tools are recorded in Table I [2].



GRAPH I. usage of data mining tools in 2014 and 2015

In a late, 2013, survey distributed on the powerful KDnuggets entryway concerning the utilization of Data mining and huge data tools in a genuine task, it is interesting to examine that in the top 5 tools there is only one business tool: Excel. The force of free tools, mainly RapidMiner, R, KNIME probably stems from the maturity and accessibility of a huge number of machine learning algorithm implementations. In bar graph I [2], a modified version of the survey is demonstrated, with only free tools listed, for comparison of tools. In any case, it is critical and valuable to consolidate their capacities so that intrigued clients can pick the suitable environment for dealing with their issue.

1. WEKA

WEKA[3] standards for Waikato environment for knowledge learning, is a java based computer program that was developed at the University of Waikato, New Zealand with the end goal of recognizing realities from (raw data gathered from cultivation domains.) WEKA offers four data mining alternatives, such as Explorer, Experimenter, and Knowledge flow, command-line interface (CLI). To begin with choice Explorer permits the definition of data source, data arrangement, machine learning algorithms, and visualization. Ensuing alternative Experimenter is primarily used to think about the execution of diverse calculation on the same dataset.. Knowledge flow specifies the data flow using relevantly connected visual components. WEKA is free and open source application that is accessible under GNU general public license agreement for noncommercial purposes. At first written in C the WEKA application has been at long last changed in java.

It is compatible with each computing platform and is easily to understand with a graphical interface that allows for quick setup and operation. WEKA supports a wide range of standards tasks for data mining: data preprocessing, data classification, data clustering, data regression, data visualization and features election. The essential reason of WEKA is software that trained to perform machine learning operation and derive the important information in the form of trends and patterns.

Figure 1 presents the WEKA .

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 41-46
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

Table 1. GENERAL CHARACTERISTICS OF THE DATA MINING TOOL

Characteristic	WEKA	RAPIDMINER	ORANGE	R	KNIME
Developer:	Univ. of Waikato, New Zealand	Rapid Miner, Germany	Univ. of Ljubljana, Slovenia	worldwide development	KNIME.com AG
Programming language:	Java	Java	C++, Python, Qt framework	C, Fortran, R	Java
License:	open source, GNU GPL 3	open source (v.5 or lower); closed source ,free Starter ed. (v.6)	open source, GNU GPL 3	free software, open source GNU GPL 2+	open source, GNU GPL 3
Current version:	3.6.10	6	2.7	3.0	2.10
GUI /command line:	Both	GUI	Both	Both	GUI
Main purpose:	general data mining	general data mining	general data mining	data mining, statistical techniques	Data Mining Data Analysis / Text Mining
Community support (est.):	Large	Large (~200 000 users)	Moderate	very large (~ 2 M users)	Moderate

2. *RAPIDMINER TOOL*

RapidMiner[4], some time ago Rapid-I YALE (Yet Another Learning Environment) is an experienced, java based, general information mining environment that was produced by the organization RapidMiner, Germany for giving Statistical investigation, predictive analytics, data mining and knowledge learning procedures including: data loading and transformation(ETL: Extraction, Transformation, Loading), data preprocessing ,modeling, evaluation and results representation, validation and optimization.

Earlier versions (v.5 or lower) were open source and latest version (v. 6) is proprietary for now, with several license options (Starter, Personal, Professional, and Enterprise).A Starter Edition is available for free to use , a Personal Edition is offered for US$999, a Professional Edition is $2,999 and pricing for the Enterprise Edition is accessible from the engineers.

RapidMiner is essentially focused on procedures that feature may contain sub procedures. Procedures are assembled by administrators as visual parts. Administrators' contain Data Mining algorithms, data sources, and data sinks. The dataflow is design by drag-and-drop of administrators and by connecting the inputs and outputs of corresponding administrators.

RapidMiner can be utilized for feature engineering, tracking and drafting ideas ,data stream mining, distributed data mining, text mining, statistical analysis, predictive analytics, business analytics, data processing, multimedia mining. RapidMiner is found in the: vitality business, hardware industry, car industry, protection, and saving money, information transfers, trade, generation, IT organizations, research division, statistical surveying, pharmaceutical industry and different fields.. The accompanying figure 2 demonstrates the graphical client interface for Rapid Miner

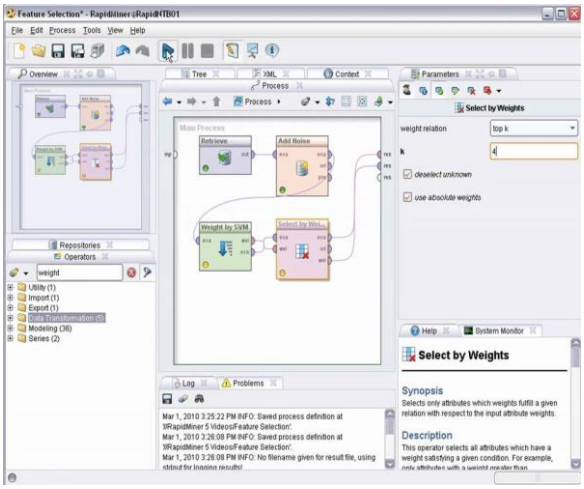


Figure 2: GUI interface of KNIME

3. *ORANGE*

Orange [5] is a powerful free and open source software suite for machine learning and data mining that was developed at the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana, Slovenia. It is a complete set of structured components for data operations, visualization, Classification, regression, evaluation, unsupervised learning, association, visualization using Qt, and prototype implementations.

Orange is c++ based part that can be gotten to straightforwardly through Python script, or through GUI objects called widgets. Functionalities of orange are visually

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 41-46
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

represented by diverse widgets (e.g. read file and train SVM classifier etc.). A short depiction of every widget is accessible within the interface. By putting widgets on the canvas and uniting their inputs and outputs, we perform the programming. The interface is exceptionally cleaned and visually appealing, offering a wonderful user experience.

Orange is a Python-based free tool under GPL and downloads from the download page. It is segment based structure which implies you can utilize existing modules and additionally makes your own ones. You can use your own module in the spot of standard orange components. It supports on Linux, Mac OS X and windows.Figure3 demonstrates the graphical user interface for ORNAGE.

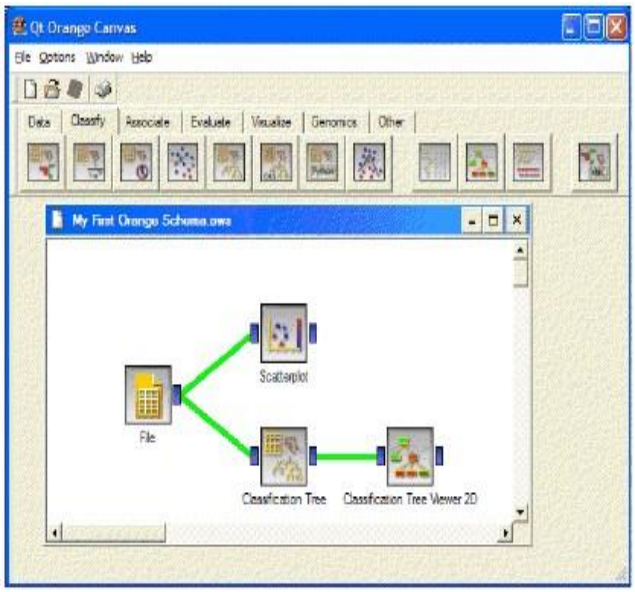


Figure 3: ORANGE GUI

4. ***R***

R[6] formally called revolution for statistical computing and graphics, is a software programming language and software environment that was outlined by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is as of now created by the R Development Core Team.

R is current execution of of S, is a statistical programming language created by John Chambers and (in earlier versions) Rick Becker and Allan Wilks of Bell Laboratories in 1970. The thought behind the dialect, as expressed by John Chambers, is "to transform thoughts into software, immediately and faithfully [7].

The R project is a cross stage for the examination, visuals and software development activities of data mineworkers, Statistician and related areas. R is free for all under GNU GPL (General Public License) and pre-compiled binary versions are provided for various systems. The source code of R interpreted language is written in C++, Fortran, and in R and Highly created clients can write C, C++, Java, Dot NET or Python code to control R objects directly.

R is a fully-supported, open source, command line driven. There are many additional "packages" freely accessible on web, which gives a wide range of data mining, classification, time-series analysis, machine learning and statistical techniques [8], as well as linear and nonlinear modeling, classical statistical experiments, clustering, and others. Figure 4 demonstrates the GUI for R.

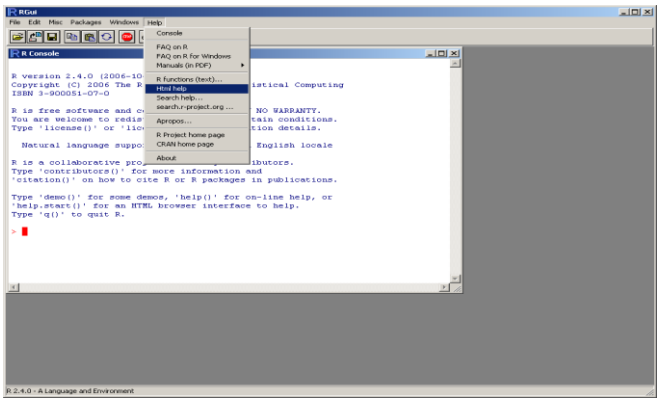


Figure 4: GUI interface of R Language

5. KNIME

Konstanz Information Mineris[9] an open source data analysis, reporting and integration platform that was developed and maintained by the Swiss companyKNIME.com AG. Itis a general purpose Data mining tool based on the Eclipse platform and is open-source, though commercial licenses exist for companies requiring professional technical support. According to the official website, KNIME is used by over 3000 organizations in more than 60 countries.

KNIME has been utilized as a part of medical research, but on other hand is utilized as a part of different areas like CRM (Customer relationship management) customer data analysis, business Intelligence, enterprise Reporting, data analysis, text mining. It is free for all under GNU GPL and pre-compiled binary versions are provided for various systems, Compatible with Linux, OS X, Windows and its latest version is 2.10 KNIME is written in java.

It is a modular data exploration tool that enables the client to visually create data flows, selectively execute some or all analysis steps, and later analyses the outcomes through interactive views on information and models. The KNIME base version already incorporates over 100 processing nodes for data I/O, preprocessing and cleaning, modeling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 41-46
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

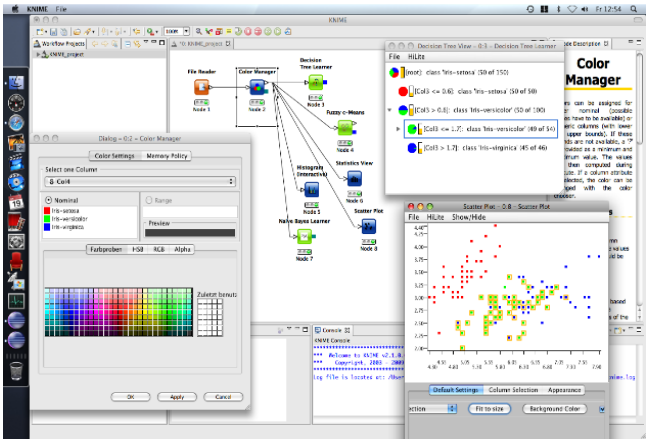


Figure 5 shows the GUI for KNIME.

Table II. : Advantages and Limitations of tools

Tools name	Advantage	Limitation
Weka	1. suitable for developing new machine learning scheme 2. Ease of use 3. Suitable for data mining tool 4. Weka is best suited for mining association rules 5. Weka loads data file in formats of ARFF, CSV, C4.5, binary	**1.** Poor documentation **2.** weak csv reader **3.** weak classical statistics **4.** poor parameter optimization
RAPIDMINER	1. Attribute Selection 2. Outlier detection 3. Solid and complete package 4. Rapid Miner has a lot of functionality, is polished and has good connectivity.	1. Requires prominent knowledge of database handling
ORANGE	1. Better debugger 2. Cross platform GUI 3. Written in python hence is easier for most programmers to learn. 4. Works both as a script and with an ETL work flow GUI.	1. Big installation 2. Limited reporting capabilities 3. Not super polished 4. Orange is weak in classical statistics;
R	1. Better graphics. 2. Purely statistical 3. Ability to make a working machine learning program in just 40 lines of code 4. more transparent	1. Less specialized towards data mining. 2. There is a steep learning curve, unless you are familiar with array languages
KNIME	**1.** Molecular analysis **2.** It is easy to try out because it requires no installation besides downloading and un archiving **3.** Chemistry Development kit **4.** Specialized for Enterprise reporting, data mining ,Commercial Intelligence	**1.** Limited error measurements 2. poor parameter optimization 3. no wrapper methods for descriptor selection

IV. CONCLUSION

Some data mining tools were presented in this paper. Overall conclusion is that there is no particular best tool. Each tool has its strong points and weak points. Nevertheless, RapidMiner, R, Weka, and KNIME have most of the desired characteristics for a fully-functional Data Mining platform and therefore their use can be suggested for most of the Data Mining operation.

V. REFERENCES

[1] Yrd.Doç.Dr. Ayça ÇAKMAK PEHLİVANLI, the comparsion of data mining tools, 16 November 2011

[2]Kalpana Rangra and Dr. K. L. Bansal,Comparative Study of Data Mining Tools,Volume 4, Issue 6, June 2014

[3] A. Jović*, K. Brkić* and N. Bogunović, An overview of free software tools for general data mining, MIPRO 2014, 26-30 May 2014, Opatija, Croatia

[4] 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.

[5] G. Piatetsky, KDnuggets Annual Software Poll: RapidMiner and R vie for first place, 2013, Available at [last accessed 2014-02-22]: http://www.kdnuggets.com/2013/06/kdnuggets-annualsoftware-poll-rapidminer-r-vie-for-first-place.html

[6]http://www.r-project.org/

[7] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, A Study of Data Mining Tools in Knowledge Discovery Process,A Study of Data Mining Tools in Knowledge Discovery Process

[8] Y. Zhao, R and Data Mining: Examples and Case Studies, San Diego: Academic Press, 20128

[9] 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