World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:6, 2014

A Review: Comparative Study of Diverse Collection of Data Mining Tools

S. Sarumathi, N. Shanthi, S. Vidhya, M. Sharmila

***Abstract*—**There have been a lot of efforts and researches undertaken in developing efficient tools for performing several tasks in data mining. Due to the massive amount of information embedded in huge data warehouses maintained in several domains, the extraction of meaningful pattern is no longer feasible. This issue turns to be more obligatory for developing several tools in data mining. Furthermore the major aspire of data mining software is to build a resourceful predictive or descriptive model for handling large amount of information more efficiently and user friendly. Data mining mainly contracts with excessive collection of data that inflicts huge rigorous computational constraints. These out coming challenges lead to the emergence of powerful data mining technologies. In this survey a diverse collection of data mining tools are exemplified and also contrasted with the salient features and performance behavior of each tool.

***Keywords*—**Business Analytics, Data Mining, Data Analysis, Machine Learning, Text Mining, Predictive Analytics, Visualization.

I. Introduction

THE domain of data mining and discovery of knowledge in various research fields such as Pattern Recognition, Information Retrieval, Medicine, Image Processing, Spatial Data Extraction, Business and Education has been tremendously increased over the certain span of time. Data Mining highly endeavors to originate, analyze, extract and implement fundamental induction process that facilitates the mining of meaningful information and useful patterns from the huge dumped unstructured data. This Data mining paradigm mainly uses complex algorithms and mathematical analysis to derive exact patterns and trends that subsists in data. The main aspire of data mining technique is to build an effective predictive and descriptive model of an enormous amount of data. Several real world data mining problems involves numerous conflicting measures of performance or intention in which it is need to be optimized simultaneously. The most distinct features of data mining is that it deals with huge and complex datasets in which its volume varies from gigabytes to even terabytes. This requires the data mining operations and algorithms to be robust, stable and scalable along with the ability to cooperate with different research domains. Hence the various data mining tasks plays a crucial role in each and every aspect of information extraction and this in turn leads to the emergence of several data mining tools. From a pragmatic perspective, the graphical interface used in the tools tends to be more efficient, user friendly and easier to operate in which they are highly preferred by researchers.

Mrs.S.Sarumathi, Associate Professor, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9443321692; e-mail: rishi_saru20@rediffmail.com).

Dr.N.Shanthi, Professor and Dean, is with the Department of Computer Science Engineering, Nandha Engineering College, Tamil Nadu, India (e-mail: shanthimoorthi@yahoo.com).

Ms.S.Vidhya, PG Scholar, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9443960666; e-mail: vidhyapsubramani@gmail.com).

Ms.M.Sharmila, Assistant Professor, is with the Department of Information Technology, M.Kumarasamy College of Engineering, Tamil Nadu, India (phone: 9443581688; e-mail: sharmi28.it@gmail.com).

II. Different Data Mining Tools

A. Rapid Miner

Rapid miner [1] is a software platform and open source system for data mining. It has being used in integrated environment which provides facilities for data mining, text mining, machine learning, business analytics and predictive analytics. For integrating the products of data analysis and data mining engine, the rapid miner software is used as a stand-alone application. It provides competitive edges in its applications to their users when applied in more than 40 countries. It has being useful for several fields such as business, education and industrial applications also. They may also include visualization, optimization, validation which supports overall steps in data mining. It provides the efficient GUI for its designing and execution of analytical workflows. Such workflows in rapid miner are said to be “process”, that it contains several “operators”. The operator functions as a single task in their process in which the input is produced by the existing output of the operator. In rapid miner from the command line the individual functions can be called.

1) Merits
a) Integration and expandability to Web services
b) Stronger in analytical ETL, data and text mining
c) Reusable and can be implanted into business processes
d) Programming the analysis without actually writing the source code by Gray-Box programming
e) Scalable with more data
2) Limitations
a) Possibly smaller user base compared to R tool
b) Rapid miners and R tool integration improvement is still ongoing process
c) Complexity of the Graphical User Interface
d) Statistical methods are lesser than R tool

B. Orange

Orange is an open source [2] and also component based software. It is also machine learning software used for an explorative data analysis and visualization, python bindings for scripting. For data preprocessing it provides a set of

components that has some exploration techniques and features like modeling, scoring and filtering. It can be implemented in C++ and python. It provides an efficient graphical user interface for building the cross platform QT framework. In case of GPL, it is distributed free. Orange has been supported by several versions of Linux, Apple's Mac OS etc. It can be add-ons for bioinformatics and text mining. This software [3] has packed with features for data analytics. It has the widgets that provide the graphical user's interface for orange's data mining and machine learning methods.

1) Merits
a) Best for beginners in the areas of research and analysis
b) User friendly work model
c) Interactive visualization function to fix the parameters straight from the graph
2) Limitations
a) No widgets offered for classical statistics
b) Compatibility concerns may arise, since it is written in NON-JAVA platform
c) Not suited for intensified classical Statistics research

C. PSPP

PSPP is a free software [4] application that is sampled data for statistical analyzing. It provides a graphical user interface and also conventional command-line interface. PSPP is considered as the system for data management and statistical analysis. It gives the SPSS proprietary program for free replacement. This enables to reads syntax and data file, analysis the data and writes the result for standard output. Both the languages accepted by PSPP and SPSS statistical products are similar. Whereas the developments of PSPP are ongoing, the current version of PSPP is incomplete due to its statistical procedure support. From the SPSS files, PSPP can collect the data and use it for functions like generating tabulated reports, descriptive statistics and to conduct complex statistical analyses. In PSPP, it is possible to make selection of simple menus and dialog box for performing complex analysis without typing any command syntax.

1) Merits
a) Supports over a billion of cases and variables
b) Choice of user interface and output formats
c) Easy import of data from external sources
d) Ability to open and analyze many data sets simultaneously
e) Completely indexed user manual with free license and no expiration period
f) Inter Operable with many free software
2) Limitations
a) Less customizable output
b) Unavailability of many advanced statistical analysis methods, such as MANOVA
c) Improvement of table and chart outputs produced by PSPP is still an ongoing process

D. KNIME

KNIME is a perfect fit tool [5] for well-designed which is an effective open source data mining software. It is an open source data analytics, useful for reporting and integrating platform. Using the concept of modular data pipelining it is enable to integrate with various components for machine learning and data mining. It provides the graphical user interface [6] for data preprocessing and for modeling, data analysis and visualization by assembly is of nodes. It is easy for intuitive data flow user interface and also gives powerful data mining elements.

1) Merits
a) Compatible with all systems, as it is written in JAVA platform
b) Very user friendly
c) Best for Molecular analysis
2) Limitations
a) Dense fuzzy extensions
b) Complexity of Lab Nodes

E. WEKA

WEKA is the freely available open source [7] application under the GNU. It supports many tasks in data mining such as data preprocessing, classification, clustering and some other process also. The purpose of this application is to utilize the given computer application that allow to perform the machine learning capabilities and useful information is derived to form patterns and trends. It provides an efficient user friendly graphical interface which allows its operation and setup quickly. It is written in C but later the WEKA application has been rewritten into java which can able to survive in almost every computing platform. It provides a tool for novice users for identifying the hidden information in the database and file systems with simple use of options and visual interface. In WEKA [8], the user data is available in the form of flat or relation, in which each data object is specified by fixed number of, attributes which has specific type or numeric values.

1) Merits
a) Good for beginners to do simple analysis with given set of data
b) Capable of pulling information from various database formats
c) Ability to get details from SQL database as well as from actual webpage just by typing the URL in IE.
2) Limitations
a) Deficiency in the prospects of interfacing with other software
b) Lesser Performance
c) Limited Memory

F. RATTLE

RATTLE (R Analytical Tool to Learn Easily) is a tool [9] in data mining that gives an uncomplicated and logical interface. It is built on top of the open source and free statistical language R with the help of Gnome graphical interface. This interface takes the user through the basic step of data mining. It turns out to be user friendly software by illustrating the R code that is used to achieve this. It uses the R Statistical Software through a graphical user interface. The software

contains a Log Code tab, which may replicates the R code for any activity by GUI, which can be copied and posted. The software permits the dataset to be partitioned into training, validation and testing. The dataset can also be viewed and edited by the user. This software [10] also has the option for scoring an external data file. Rattle is compatible with GNU/Linux, Macintosh OS X and MS/Windows. It presents statistical and visual summaries of data, which transforms their data into forms that is readily modeled, support for building both unsupervised and supervised models from the data, it represents the performance of models graphically and scored new datasets. The Rattle software is used in Australia and other countries for business, government, research, statistical analysis, model generation and for teaching data mining. Whilst the tool itself may be sufficient for all of a user's needs, it also provides a more sophisticated processing and modeling in R.

1) Merits
a) Innate design with sequential tabs that makes it easily understandable
b) Good for building a complete analytical platforms like clustering, decision trees
c) Auto generated log with time stamps
2) Limitations
a) It is more of data mining GUI reasonably than analytical GUI
b) Limited ability to generate different types of graphs
c) No support to BIGLM packages and parallel programming

G.jHepWork

jHepWork [11] is an interactive framework for scientific computation, data analysis and visualization which is useful for scientists, engineers and students. It runs on any operating system where the Java virtual machine can be installed, as the code is written in JAVA. The program is fully multiplatform and integrated with the Jython (Python) scripting language. The symbolic calculation can be done using MATLAB/Octave high-level interpreted language. The libraries include both numerical and analytical calculations, linear algebra operations and equation solving algorithms. The program is developed for interactive scientific plots in 2D and 3D and contains numerical scientific libraries implemented in Java for mathematical functions, random number, curve fitting and for other data mining algorithms also. jHepWork [12] is an attempt to create a data-analysis environment using open-source packages with a coherent user interface and tools competitive to commercial programs. The aim is to incorporate open-source mathematical and numerical software packages with GUI-type of user interfaces into a coherent program in which the main user interface is based on short-named Java/Python classes. It is used to build an analysis environment using Java scripting concept. jHepWork can be used everywhere for analysis of large numerical volumes of data, data mining and statistical data analysis. jHepWork is considered among five best free and open source data-mining software. It was renamed to SCaVis project from 2013.

1) Merits
a) Can display 1D and 2D functions and histograms
b) Math operations with high level data containers is possible
c) Neural network and linear regression analysis
2) Limitations
a) Not completely free for commercial usage
b) High Memory requirements

H. Apache Mahout

Apache Mahout [13] is a project of the Apache Software Foundation. The aim of this project is to produce free implementations of distributed or otherwise scalable machine learning algorithms on the Hadoop Platform. The project mainly focuses on collaborative filtering, clustering and classification. Hadoop platforms are useful for many implementations. Mahout provides Java libraries for common math operations (more for linear algebra and statistics) and primitive Java collections. Though the number of implemented algorithm has grown quickly, mahout supports four use cases: Recommendation mining which takes user' behavior and tries to find items users might like. Clustering which takes the text documents and groups them into groups of topically related documents. Classification which learns from existing categorized documents. Classification learns from existing categorized documents and assigns unlabeled documents to the (optimistically) correct category. Frequent item set mining that takes a set of item groups and identifies, in which each items normal appear together. These algorithms can be implement on top a of Apache Hadoop using the map/reduce paradigm. But it does not restrict Hadoop based implementations. The Contributions that run on a single node or on a non-Hadoop cluster are also welcomed.

1. Merits
a) Compatible since it is written in JAVA
b) Provides Java libraries for common math operations
c) Uses Apache Hadoop platform
2. Limitations
a) It is a still work in progress
b) Unavailability of various algorithms

I. Alpha Miner

Alpha Miner [14] is an open source data mining platform which is designed by the E-Business Technology Institute (ETI). It offers the best cost and performance ratio for data mining applications. Workflow style case construction facilitates simple drag-and-drop operations for general business managers in construction of data mining case. Plug-able component architecture is very much useful in adding new BI capabilities in data import and also for export, modeling algorithms, model assessment and deployment, data transformation, thus affording extensibility. Data mining capabilities from Xelopes and WEKA have been incorporated in the first release. Versatile data mining functions [15] is built with powerful analytics in order to demeanor industry specific analysis such as customer profiling and clustering, product association analysis and finally for classification and

prediction.

1) Merits
a) Easy construction of a data mining case by simple drag-and-drop operations
b) Extensibility in addition new BI capabilities in data import and export, data transformations, modeling algorithms
c) Versatile data mining functions
2) Limitations
a) Efficiency is less compared to BI solutions

J. KEEL

KEEL [16] is an open source (GPLv3) java software tool to impose evolutionary algorithms for Data Mining problems. KEEL is designed for solving data mining problems and assessing evolutionary algorithms. The application includes regression, classification, clustering, and pattern mining and so on. It permits the user to perform a complete analysis of any learning model in comparison to existing ones; it includes a statistical test module for making comparison. It contains a vast collection of classical knowledge of extracting algorithms and preprocessing techniques. The Computational Intelligence based on learning algorithms, which includes evolutionary rule learning algorithm based on different approaches (Pittsburgh, Michigan and IRL …) and hybrid models such as genetic fuzzy systems, evolutionary neural networks, etc. The two main goals [17] beyond the development of KEEL are research and education. The implemented programs are useful in wide research and educational goals such as evolutionary fuzzy rule learning, genetic artificial neural networks and Learning Classifier Systems etc. It has a collection of libraries for preprocessing and post-processing techniques for data manipulating, soft-computing methods in knowledge of extracting and learning, and then for providing scientific and research methods etc.

1) Merits
a) Less programming work to the user
b) Spreads the range of users applying evolutionary learning algorithms
c) Independent of Operating System
2) Limitations
a) Do not support cross-validation
b) Less comprehensible models

K. Monarch

Monarch is a desktop report mining tool [18] used to extract data from human readable report files, such as text, PDF, XPS and HTML. The program was developed by Math Strategies and the software is published by Data watch Corporation 1991. Over 500,000 copies of Monarch have been licensed and the software is in use in over 40,000 organizations with the latest release version as 11.Monarch can import data from OLE DB/ODBC data sources, spreadsheets and desktop databases. It also permits users to re-use information from existing computer reports, Such as text, PDF and HTML files. Users define models that explain the layout of data available in the report file, and the tabular format can be done by the software parses of data. The parsed data can be again enhanced with links to external data sources, filters and sorts, finally for calculated fields and summaries. The data that can be exported to a variety of and primarily spreadsheets also included.

1) Merits
a) Allows users to re-use data from existing computer reports
b) Can import and export data to various sources
c) OLAP Manipulations
2) Limitations
a) No available for a wide range of customers

L. TANAGRA

TANAGRA [19] is open source data mining software which is used for purposes like academic and research. It proposes various data mining methods from exploratory analysis of data, statistical and machine learning and for databases area also. The main purpose of Tanagra project is to give researches and students easy-to-use data mining software and allowing analyzing either real or synthetic data. It is also used to propose architecture to researchers and allowing them to easily add their own data mining methods, to compare their performances. This software has free access to source code and so it can be considered as a pedagogical for learning programming techniques.

1) Merits
a) Performs multiple linear regression
b) Predicts the instances with ID3
c) Good to use in supervised discretization and hybrid clustering
2) Limitations
a) K means method uses a large number of cores
b) Time consuming computation of distance

M. SIPINA

SIPINA implements supervised learning [20] and are available for free academic and research purposes. It implements various supervised learning paradigms and especially intended to do decision trees induction. It has some features which have Data Access, Feature Transformation and Selection, Error Evaluation, Classification and Learning Algorithms. SIPINA is mainly a Classification Tree Software. But, other supervised methods are also available such as k-NN, Multilayer perceptions, Naïve Bayes, etc. We can perform the performances comparison and model selection.

1) Merits
a) Implements the decision graphs algorithm
b) Able to perform performances comparison and model selection
c) Delivers much functionality for the data exploration methods
2) Limitations
a) Inability to handle large datasets
b) Sometimes there is error rate degradation

III. Summarization of Data Mining Tools

The following Table I illustrate the comparative view of different data mining tools based on their compatibility characteristics and its application domains. The main aim of this comparison is not to scrutinize which is the best Data Mining tool but to exemplify the usage paradigm and the awareness of tools in several fields.

TABLE I
Comparison of Data Mining Softwares

Name of the Data Mining Tool	Platform supported	Mode of Software	Applications
Rapid Miner	Cross Platform	Open Source	Statistical Analysis, Data Mining, Predictive Analytics
Orange	Cross Platform	Open Source	Machine Learning, Data Mining, Data Visualization
PSPP	GNU	Open Source	Statistical Analysis
Knime	Windows, Linux, Macintosh	Open Source	Enterprise Reporting, Business Intelligence, Data Mining
Weka	Cross Platform	Open Source	Machine Learning
Rattle	Ubuntu, GNU/Linux, Microsoft Windows, Macintosh	Open Source	Statistical Analysis, Model Generation
jHepWork	Cross Platform	Open Source	Data Analysis Visualization, Scientific Computation
Apache Mahout	Cross Platform	Open Source	Machine Learning, Data Mining
Alpha Miner	Windows and Linux	Open Source	Data Mining
Keel	Cross Platform	Open Source	Machine Learning, Regression, Classification, Clustering, Pattern Mining
Monarch	Windows, Unix	Commercial	Report Mining, Data Analysis, Business Intelligence
Tangara	Windows	Open Source	Machine Learning, Data Mining, Multivariate Analysis, Data Analysis
Sipina	Windows	Open Source	Machine Learning, Data Mining

IV. Conclusion

Several Data Mining tools were elucidated along with their usage of various tasks in this paper. Each and every sub tasks of data mining tends to be highly essential underpinning process for efficient information extraction. This requisite paves the way for the development of many data mining tools. These tools has the extensive technical paradigm, outstanding graphical interface and inbuilt multipart algorithms in which it is very useful for handling substantial amount of data more precisely and legibly. Thus the main role of this survey is to enrich knowledge about the data mining tools and its appliance in several industries which will be very useful for the readers and also it meets the needs of data mining researchers to innovate more advanced tools in future.

References

[1] Rapid Miner [Online]. Available at: http://www.rapid-i.com/downloads/tutorial/rapidminer-4.6-tutorial.pdf
[2] Orange [Online]. Available at: http://www.slideshare.net/jt_4285/data_mining_tool_orange
[3] Orange [Online]. Available at: ftp://kibernetica.fov.uni-mb.si/PES/orange.pdf
[4] PSPP [Online]. Available at: http://data-mining-tutorials.blogspot.in/2012/03/PSPP-alternative-to-spss.html
[5] Knime [Online]. Available at: http://www.dataminingresearch.com/index.ph/2010/07/knime_open_source_data_mining_software/.
[6] Knime [Online]. Available at: http//unipi.it/lib/exe/fetch.php/dm/knime_slides.pdf
[7] Weka [Online]. Available at: http://www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf
[8] Weka [Online]. Available at: http://www.cs.ccsu.edu/nmarkov/weka.tutorial.pdf
[9] Rattle [Online]. Available at: http://cs.anu.edu.au/rattle.pdf
[10] Rattle [Online]. Available at: http://rattle.togaware.com/
[11] jHepWork [Online]. Available at: http://www.top4download.com/jhepwork/noslwfpa.html
[12] jHepWork [Online]. Available at: http://download.cnet.com/jHepWork/3000_2070 4_75833656.html
[13] Apache Mahout [Online]. Available at: http://hortonworks.com/hadoop/mahout
[14] Alpha Miner [Online]. Available at: http://alphaminer.software.informer.com/2.0
[15] Alpha Miner [Online]. Available at: http://www.eti.hku.hk/alphaminer/
[16] Keel [Online]. Available at: http://www.keel.es
[17] Keel [Online]. Available at: http://www.salleurl.edu/GRSI/docs/keel_softcomputing.pdf
[18] Monarch [Online]. Available at: http://www.ion.icaew.com/ClientFiles/72181ac8-e560-4bc4-8c20-6be215bef4bc/Monarch%20BI%20Brochure.pdf
[19] Tanagra [Online]. Available at: http://eric.univ_lyon2.fr/~ricco/tanagra/en.tanagra.html
[20] Sipina [Online]. Available at: http://eric.univ_lyon2.fr/~ricco/sipina

Mrs. S.Sarumathi received B.E. degree in Electronics and Communication Engineering from Madras University, Madras, Tamil Nadu India in 1994 and the M.E. degree in Computer Science and Engineering from K.S.Rangasamy College of Technology, Namakkal Tamil Nadu, India in 2007. She is doing her Ph.D. programme under the area Data Mining in Anna University, Chennai. She has a teaching experience of about 16 years. At present she is working as Associate professor in Information Technology department at K.S.Rangasamy College of technology. She has published 5 reputed International Journals and two National journals. And also she has presented papers in three International conferences and four national Conferences. She has received many cash awards for producing cent percent results in university examination. She is a life member of ISTE.

Dr.N.Shanthi received the B.E. degree in Computer Science and Engineering from Bharathiyar University, Coimbatore, Tamil Nadu, India in 1994 and the M.E. degree in Computer Science and Engineering from Government College of Technology, Coimbatore, Tamil Nadu, and India in 2001. She has completed the Ph.D. degree in Periyar University, Salem in offline handwritten

Tamil Character recognition. She worked as a HOD in department of Information Technology, at K.S.Rangasamy College of Technology, Tamil Nadu, India since 1994 to 2013, and currently working as a Professor & Dean in the department of Computer Science and Engineering at Nandha Engineering College Erode. She has published 29 papers in the reputed International journals and 9 papers in the National and International conferences. She has published 2 books. She is supervising 14 research scholars under Anna University, Chennai. She acts as the reviewer for 4 international journals. Her current research interest includes Document Analysis, Optical Character Recognition, and Pattern Recognition and Network security. She is a life member of ISTE.

Ms.S.Vidhya holds a B.Tech degree in Information Technology from N.P.R College of Engineering and Technology, affiliated to Anna University of Technology ,Tiruchirappalli, Tamil Nadu, India in 2013. Now she is an M.Tech student of Information Technology department in K.S.Rangasamy College of Technology. She has presented three papers in National level technical symposium. Her Research interests includes Data Mining, Wireless Networks.

Ms. M.Sharmila received B.Tech degree and M.Tech degree in Information Technology from K.S.Rangasamy College of Technology, affiliated to Anna University Chennai, Tamil Nadu, India in 2012 and 2014 respectively. At present she is working as an Assistant Professor in Information Technology department at M.Kumarasamy College of Engineering Karur. She has published 4 international journal and presented three papers in National level technical symposium. She is an active member of ISTE. Her Research interests include Mining Medical data, Opinion Mining and Web mining. Most of her current work involves the development of efficient cluster ensemble algorithms for extracting accurate clusters in large dimensional database.