ISSN No. 0976-5697

Volume 5, No. 6, July-August 2014

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Data Mining Open Source Tools - Review

V. Saravanan
Associate Professor
Department of Computer Applications
Hindusthan College of Arts and Science
Coimbatore, Tamilnadu, India

C. Pushpalatha
Research Scholar
Department of Computer Science
Hindusthan College of Arts and Science
Coimbatore, Tamilnadu, India

C. Ranjithkumar
Assistant Professor
Department of Computer Science
Sri Krishna College of Arts and Science
Coimbatore, Tamilnadu, India

Abstract: Data Mining has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information and also data mining is defined to the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. And also researcher have to applying data mining methods and algorithms in many applications. whose development is driven by strong research interests as well as urgent practical, social, and economical needs. While the last few years knowledge discovery tools have been used mainly in research environments, sophisticated software products are now rapidly emerging. In this paper we have to give survey of most used and popular data mining tools of **Clementine, Rapid miner, R, SAS enterprise miner** and its features. These kind of data mining tools used for prediction and analyzing data mining process and using applications are education, learning environments, statistics and etc.

Keywords: Data mining, clementine, Rapid miner, R, SAS Enterprise miner.

I. INTRODUCTION

Each and every day the human beings are using the vast data and these data are in the different fields .It may be in the form of documents, graphical formats, video , records. As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data. As and when the customer will required the data should be retrieved from the database and make the better decision. This technique is actually we called as a data mining or Knowledge Hub or simply KDD(Knowledge Discovery Process).

The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of "Data mining" is due to the perception of "we are data rich but information poor". Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions.

Data mining is used in a variety of fields and applications. The military use data mining to learn what roles various factors play in the accuracy of bombs. Intelligence agencies might use it to determine which of a huge quantity of intercepted communications are of interest. Security specialists might use these methods to determine whether a packet of network data constitutes a threat. Medical researchers might use them to predict the likelihood of a cancer relapse[1][2].

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process[1].

The sequences of steps identified in extracting knowledge from data are shown in Figure.

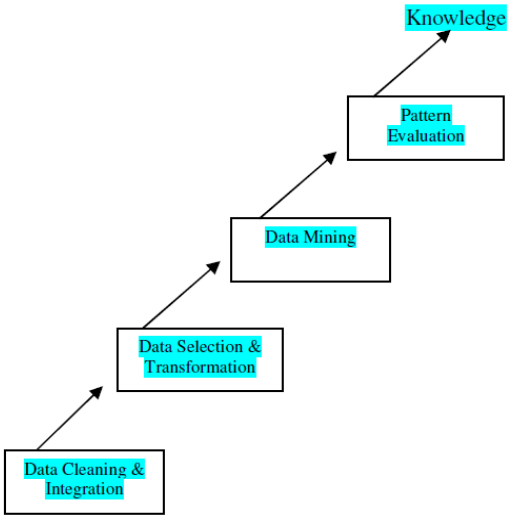

Figure 1: Extracting knowledge from data

II. DATA MINING TECHNIQUES

A. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification .In learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier[3][4].

B. Clustering:

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification[3][4].

C. Prediction:

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) maybe necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models[3][4].

D. Association rule:

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value[3][4].

E. Decision Trees:

Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID)[3][4].

III. DEVELOPMENT OF DATA MINING TOOLS

Some of the most significant advances have been in supporting the management of large data sets, making it possible to store, organize, and sift through data in ways that make it substantially easier to analyze. Google developed Map Reduce to address the substantial challenges of managing data at the scale of the internet, including distributing data and data-related applications across networks of computers.

Map Reduce led to the development of Apache Hadoop, now commonly used for data management. In addition to tools for managing data, an increasing number of tools have emerged that support analyzing it. In recent years, the sophistication and ease of use of tools for analyzing data make it possible for an increasing range of researchers to apply data mining methodology without needing extensive experience in computer programming. Many of these tools are adapted from the business intelligence field, as reflected in the prominence of SAS and IBM tools in education, tools that were first used in the corporate sector for predictive analytics and improving organizational decision making by analyzing large quantities of data and presenting it in a visual or interactive format.

In the early 2000s, many analytics tools were technically complex and required users to have advanced programming and statistical knowledge. Now, even previously complex tools such as SAS, Rapid Miner, and SPSS, weka, mat lap are easier to use and allow individuals to conduct analytics with relatively less technical knowledge. Common desktop software, such as Microsoft Excel, has also incorporated significantly improved visualization and analytics features in recent years. Other tools such as Tableau Software are designed to support the use of analytics tools without advanced technical knowledge.

IV. CATEGORIES OF DATA MINING TOOLS

Most of the data mining tools can be classified into three categories: Traditional data mining tools, dash boards and text-mining tools. Description of each is as follows

A. Traditional Data Mining Tools :

Traditional mining programs help the companies to establish data patterns and trends by using various complex algorithms and techniques. Some of these tools are installed on the desktop computers to monitor the data and emphasize trends and others capture information residing outside a data base. Majority of these programs are supported by windows and UNIX versions. However, some software specializes in one operating system only. In addition to that some may work in only one database type. But, Most of the software will be able to handle any data using online analytical processing or a similar technology[5][6].

B. Dashboards:

Dashboards reflect data changed and update on screen. Dashboards is normally installed in computers to monitor information in a database and it reflects data changes and updates the data in the form of a chart or table on the screen.

It enables the user to see how the business is performing. Historical data can be referenced and checks against the current status in order to see the changes in the business. By this way, dashboards is very easy to use and helps the manager a lot with great appeal to have an overview of the company's performance[5][6].

C. Text-Mining Tools:

The third type of data mining tools is called as a text-mining tool because of its ability to mine data from different kind of text starting from Microsoft Word, Acrobat PDF documents to simple text files. This provides facility of scanning the content and converts the selected into a format that is compatible with the tools database without opening different applications[5][6].

V. DATA MINING TOOLS

A. Clementine:

Clementine is a mature data mining toolkit which aims to allow domain experts (normal users) to do their own data mining. Clementine has a visual programming or data flow interface, which simplifies the data mining process. Clementine applications include customer segmentation/profiling for marketing companies, fraud detection, credit scoring, load forecasting for utility companies, and profit prediction for retailers. SPSS Clementine was one of the very first general purpose data mining tools, and one of the most popular data mining packages. Data Mining Approaches are Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Text Mining, Outlier Discovery, Data Visualization, Discovery Visualization, Sequence Analysis, Web Analytics, Social Network Analysis

Clementine Interface the numerous features of Clementine's data mining workbench are integrated by a visual programming interface. You can use this interface to draw diagrams of data operations relevant to your business. Each operation is represented by an icon or **node**, and the nodes are linked together in a **stream** representing the flow of data through each operation. *Stream canvas* the stream canvas is the largest area of the Clementine window, and it is where you build and manipulate data streams. You can work with multiple streams at a time in Clementine, either in the same stream canvas or by opening a new stream. Streams are stored in the managers during a session.

Palettes the palettes are located across the bottom of the Clementine window. Each palette contains a related group of nodes that are available to add to the data stream. For example, the Sources palette contains nodes that you can use to read data into your model, and the Graphs palette contains nodes that you can use to explore your data visually. The Favorites palette contains a default list of nodes frequently used by data miners. As you become more familiar with Clementine, you can customize the contents for your own use.

Managers at the upper right of the Clementine window are three types of managers. Each tab Streams, Outputs, and Models—is used to view and manage the corresponding types of objects. You can use the Streams tab to open, rename, save, and delete the streams created in a session. Clementine output, such as graphs and tables, is stored in the Outputs tab. You can save output objects directly from this manager. The Models tab is the most powerful of the manager tabs and contains the results of machine learning and modeling conducted in Clementine. These models can be browsed directly from the Models tab or added to the stream in the canvas.

Projects the Projects window is located at the lower right of the Clementine window and offers a useful way of organizing your data mining efforts in Clementine.

Report window located below the palettes, the Report window provides feedback on the progress of various operations, such as when data are being read into the data stream. *Status window* also located below the palettes, the Status window provides information on what the application is currently doing, as well as indications when user feedback is required[6][7].

B. Rapid Miner:

Rapid Miner (formerly YALE) is the world-wide leading open-source data mining solution due to the combination of its leading-edge technologies and its functional range. Applications of Rapid Miner cover a wide range of real-world data mining tasks.Use Rapid Miner and explore data. Simplify the construction of experiments and the evaluation of different approaches. To find the best combination of preprocessing and learning steps. The modular operator concept of Rapid Miner (formerly YALE) allows the design of complex nested operator chains for a huge number of learning problems in a very fast and efficient way (rapid prototyping). Rapid miner supports database management systems likeOracle, Sql server, Postgre Sql and My Sql and it also support many file formats like raff, excel, csv. Rapid miner having two different type of interfaces, that are Edit mode, Result mode[6][8][9][10].

a. Applications:

Rapid Miner has been applied for machine learning and knowledge discovery tasks in a number of domains including feature generation and selection , concept drift handling , and transduction. In addition to the above-mentioned, current application domains of Rapid Miner also include the pre-processing of and learning from time series, meta learning, clustering, and text processing and classification. There exist several plugins to provide operators for these special learning tasks. Among these, there are some unusual plugins like GAStruct which can be used to optimize the design layout for chemical plants. *Data Mining Approaches* are Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Text Mining, Outlier Discovery, Data Visualization[6][8][10][11].

b. Features:

a) Easy-to-use visual environment for predictive analytics. No programming required.
b) Open and extensible, Advanced analytics at every scale
c) Runs on all major platforms and operating systems.
d) Graphical user interface has been completely refined and simplified.
e) Easy data management of user process
f) A simple drag and drop allows data integration in few mouse clicks only.
g) Analysis results are interactively visualized.

h) Results can be used in repository and used in another process.

C. *R:*

R is a well supported, open source, command line driven, statistics package. There are hundreds of extra "packages" available free, which provide all sorts of data mining, machine learning and statistical techniques. It has a large number of users, particularly in the areas of bio-informatics and social science. R is a free software programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls and surveys of data miners are showing R's popularity has increased substantially in recent years. R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. S was created by John Chambers while at Bell Labs [6][12].

R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S.R is a GNU project. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. R uses a command line interface; however, several graphical user interfaces are available for use with R. R is widely used in both academia and industry. *Data Mining Approaches* are Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Text Mining, Outlier Discovery, Data Visualization , Discovery Visualization, Sequence Analysis, Web Analytics, Social Network Analysis. provide collections of packages for different tasks [6][12][13].

a) Machine learning & statistical learning
b) Cluster analysis & finite mixture models
c) Time series analysis
d) Multivariate statistics, Analysis of spatial data

a. Statistical features:

a) R provides a wide variety of statistical and graphical techniques
b) R is easily extensible through functions and extensions.
c) R is highly extensible, stronger object-oriented programming facilities.
d) than most statistical computing languages
e) Dynamic and interactive graphics are available through additional packages
f) R has its own LaTeX-like documentation format.
g) Program features
h) R is an interpreted language , supports matrix arithmetic.
i) R's extensible object-system includes objects for: regression models, time-series and geo-spatial coordinates.
j) R supports procedural programming with functions

D. *Sas Enterprise Miner:*

SAS (Statistical Analysis System) is a software suite developed by SAS Institute for advanced analytics, business intelligence, data management, and predictive analytics. It is the largest market-share holder for advanced analytics. SAS was developed at North Carolina State University from 1966 until 1976, when SAS Institute was incorporated. Forward-thinking organizations today are using SAS data mining software to detect fraud, anticipate resource demands, increase acquisitions and curb customer attrition [6][14].

a. Benefits:

a) Support the entire data mining process with a broad set of tools.
b) Build more models faster with an easy-to-use GUI.
c) Enhance accuracy of predictions and easily surface reliable business information.
d) Ease the model deployment and scoring process.

b. Features:

a) Multiple interfaces
b) Scalable processing
c) Data preparation, summarization and exploration Business-based model comparisons, reporting and management
d) Automated scoring process
e) Open, extensible design
f) SAS components include

VI. DATA MINING TOOLS AND TYPES

Different types of similar data mining tools can be found. The typical characteristics of these types are explained in this section.

A. Data mining suites (DMS):

DMS focus largely on data mining and include numerous methods. They support feature tables and time series, while additional tools for text mining are sometime available. The application focus is wide and not restricted to a special application field, such as business applications however, coupling to business solutions, import and export of models, reporting, and a variety of different platforms are nonetheless supported [15].

B. Business intelligence packages (BIs):

BIs have no special focus to data mining, but include basic data mining functionality, especially for statistical methods in business applications. BIs are often restricted to feature tables and time series, large feature tables are supported. They have a highly developed reporting functionality and good support for education, handling, and adaptation to the workflows of the customer [15].

C. Mathematical packages:

MATs have no special focus on data mining, but provide a large and extendable set of algorithms and visualization routines. They support feature tables, time series, and have at least import formats for images. The user interaction often requires programming skills in a scripting language. MATs are attractive to users in algorithm development and applied research because data mining algorithms can be rapidly implemented, mostly in the form of extensions (EXT) and research prototypes (RES) [15].

D. Integration packages :

INTs are extendable bundles of many different open-source algorithms, either as stand-alone software (mostly based on Java; as KNIME, the GUI-version of WEKA, KEEL, and TANAGRA) or as a kind of larger extension package for tools from the MAT type [15].

E. EXT:

EXT are smaller add-ons for other tools such as Excel, Matlab, R, and so forth, with limited but quite useful functionality. Here, only a few data mining algorithms are implemented such as artificial neural networks for Excel (Forecaster XL and XLMiner) or MATLAB [15].

Table I : Summary Of Data Mining Tools

Tool Name	Type	Remarks
SAS Enterprise Miner	DMS	one of the world's leading tools, enterprise oriented
R	MAT	complete statistical suite, script-based, GNU-GPL
RapidMiner	DMS	formerly YALE, more than 1000 algorithms and operators for data mining, text mining, formerly YALE, more than 1000 algorithms and operators for data mining, text mining, web mining, time series analysis and forecasting, audio mining, image mining, predictive analytics, ETL, reporting, integrates Weka and R and Hadoop
SPSS Clementine	DMS	former Clementine, now in cooperation with IBM, Predictive Analytics Software (PASW), SPSS is an IBM company since 2009.

VII. CONCLUSION

Data mining will be considered one of the most important frontiers and one of the most promising interdisciplinary developments in Information technology. In this paper, we have discussed review the knowledge Discovery Process, various Data mining techniques, data mining tools. These tools are mostly used in data mining prediction, analytics process. This review would help the researchers to focus on the various issues of data mining. we seen tools are mostly using in industries & development centre's and use these tools to get various results in academic research.

VIII. REFERENCES

[1]. Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi," the survey of data mining applications and feature scope ", International Journal of Computer Science, Engineering and Information Technology , Vol.2, No.3, June 2012 , pg: 45-58.

[2]. Michael goebel, Le Gruenwald "A survey of data mining and knowledge Discovery software tools", Sigkdd explorations., june 1999. Volume 1, issue 1 – page 22.

[3]. Abeer badr el din ahmed1, ibrahim sayed elaraby, " data mining: a prediction for student's performance using classification method ", world journal of computer application and technology 2(2): 43-47, 2014.

[4]. Brijesh kumar baradwaj, saurabh pal, "Mining Educational Data to Analyze Students‟ Performance", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011, pg:63-69.

[5]. http://www.theiia.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools-and-techniques/.

[6]. Ashwin satyanarayana, "Software tools for teaching undergraduate data mining course". http://www.asee.org/documents/sections/middleatlantic/fall-2013/9-ashwin-software-tools-for-teaching undergraduate-data-mining-course.pdf.

[7]. "Clementine® 8.0 User's Guide"

[8]. "Rapid Miner – beginner's guide", https://edux.fit.cvut.cz/oppa/MI-ADM/cviceni/c4-RM-intro.pdf.

[9]. Dr. Ayça çakmak pehlivanli, "The comparison of data mining tools", Data warehouses and data mining yrd.doç.

[10]. Huang, Huaming ; Wu, Ge "Introduce to Data Mining with RapidMiner", 2008, Syracuse University, EECS.

[11]. "RapidMiner 4.6 User Guide", Operator Reference, rapid-i.com/downloads/tutorial/rapidminer-4.6 tutorial.pdf.

[12]. Luis Torgo, "Data Mining with R: learning by case studies", http://www.goodreads.com/book/show/7125665-data-mining-with-r.

[13]. Yanchang Zhao, "R and Data Mining: Examples and Case Studies", http://cran.r project.org/doc/contrib/Zhao_R_and_data_mining.pdf .

[14]. "Getting Started with SAS Enterprise Miner 6.1".

[15]. J.Mary Dallfin Bruxella, S.Sadhana, S.Geetha, "Categorization of Data Mining Tools Based on Their Types" International Journal of Computer Science and Mobile Computing, Vol. 3, Issue. 3, March 2014, pg.445 – 452.