Data Mining Tools- An Analytical Approach

Ravi Teja
VI Sem, Dept of M.C.A
R. V. College of Engineering
Bangalore-560059, India

Dr. Jasmine K S
Associate Professor, Dept of M.C.A
R. V. College of Engineering
Bangalore-560059, India

***Abstract*—The world runs on data, and it is money in today's world. Massive data sets contain large, varied and complex data to store, analyze and visualize the process. Data mining with large data sets and repositories has been a major concern in the research community, due to the difficulty of analyzing a large quantity of data. There are different types of tools exist to analyze the data and mine it. With the techniques like clustering, classification, analysis, time series, it is possible to handle large amounts of data. The rapid growth of technology has created a revolution in data mining and developed various tools to work on it. Data mining algorithms work with classification, clustering, regression, decision trees, association. In this context, the paper investigates the suitability of various data mining tools in a practical perspective.**

Keywords* — *Data mining; visualization; analysis; tools

I. INTRODUCTION

Data Mining is the process of discovering, understanding the predictive models from large-scale data which refers to extracting knowledge from large amounts of data. Data mining techniques are rapidly increasing in the application of real-world databases. It is also known as Knowledge Discovery from Databases (KDD). Big data is changing the traditional data analysis platforms.

Unstructured data is either machine generated or human generated. The successful implementation of data mining tools requires careful assessments with the proper functioning of algorithms. With the advanced changes in the current technology, the data should store at multiple locations for each and every second [1].

The data mining tools can predict the future trends and tools are used to mine such type of data. Most of the data is unstructured and hence it takes a process to extract useful information from the data and transform it into understandable. Plenty of tools for data mining tasks are used for artificial intelligence, machine learning to extract the data. Some of the tools are RapidMiner, Weka, Orange, Tanagra, KNIME, R programming language. This paper deals with the major data mining tools which are mentioned below.

- RapidMiner
- R
- Weka

II. CHALLENGES, DRAWBACKS & APPLICATIONS

The existing work deals with the data mining tools, user requirements and their performance issues. Most of the issues are related to user interactions. Mining the data is not so easy and it undergoes the very complicated process as it is a challenge. For example, the process of text mining includes removing numbers, symbols, punctuations, spaces and plotting word frequencies, sparse matrix and applying different types of algorithms [2].

A. Challenges:

- It contains huge amounts of data repositories which is difficult to handle.
- The algorithms must be efficient to extract huge sets data from the database.
- The user requires different types of algorithms and there is a lack of knowledge.
- The system may not work properly if the data contains noise.
- Different types of data types, objects are required and impossible for one system to mine the data.

B. Drawbacks:

- The lack of accuracy, missing attributes, labels will not give the efficient data.
- Data can be misused for unauthorized users and for frauds.
- It may not provide the necessary security to the companies.
- Data mining brings out the patterns, but the significance and validity of those patterns must be made by the user.
- It requires skills and experts to handle the tools.

C. Applications of data mining:

Data mining has a lot of scope and applied in everywhere. Applications of data mining include text mining, predictive analysis, fraud detection, decision making, risk evaluation etc.

III. R

R is the powerful open-source implementation of the language S. R is a very effective statistical tool and well worth the effort to learn. R is polymorphic, which means that the same function can be applied to different types of objects, with results tailored to the different object types [3].

R has the flexibility to work with other programming languages like Java, Python, C/C++. It enables to work with many databases (ODBC, MySql) and statistical packages (SAS). R is a GNU (General Public License) project. R is simple and used for statistical computing. R contains more than 7000 packages which are to be imported. R was created by Ross Ihaka and Robert Gentleman.

A. Characteristics of R:

- R is open source and free.
- Easy to combine with statistical calculations.
- It supports multiple platforms like Windows, Linux.
- It is both object-oriented and functional programming structure [4].
- The graphical capabilities of R are outstanding, providing a fully programmable graphics language that surpasses most other statistical and graphical packages.
- R has more than 4000 packages available from multiple repositories in various specializations.

B. R Packages:

R contains more than 7000 packages available at CRAN (Comprehensive R Archive Network). CRAN contains both FTP and web servers. The packages can either loaded or downloaded.

C. R Visualization:

R supports multiple graphical interfaces. R supports different types of graphical features to analyze the data easily. Using the required packages, the web applications can build. It has multiple colors with different types of charts like pie-chart, histogram [5].

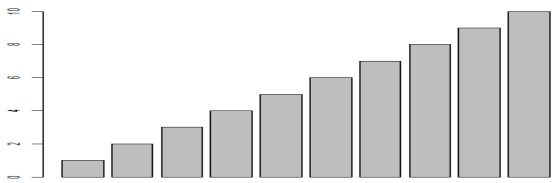

Fig. 1. R visualization with simple data.

D. Advantages:

- The code can be written in C language to manipulate R objects directly to link, call at run time.
- R can be extended through packages.
- R has strong graphical features for data analysis.
- It supports multiple import and export formats.
- Huge amounts of data are stored in database MySql.
- The output can be exported into jpg, pdf formats to understand easily.
- R can import data from CSV files, Excel, SAS and produces the output in pdf, jpg, png formats and also table output.
- It is easy to connect with social media networks like Facebook, twitter.

E. Disadvantages:

- Availability and usage of packages are difficult. They depend on the version
- R has a steep learning curve
- R can quickly consume available memory

F. Algorithms implemented:

The various types of algorithms can be implemented in R to work with different functions. Data mining, Machine learning algorithms can be implemented. Techniques include linear regression, optimization algorithms, classification.

- K-means algorithm
- Apriori algorithm
- Naïve Bayes
- CART
- Random Forests

Examples of usage:

- Extraction of social network data
- Building word cloud

F. Working with R Studio

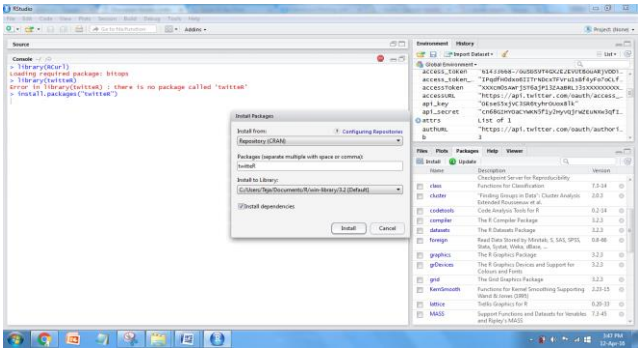

Fig. 2. R Studio screen with basic commands to install packages.

IV. RAPIDMINER

Rapidminer provides an integrated environment for machine learning, data mining, and text mining, predictive analytics. It is the most powerful tool, easy to use and intuitive graphical interface for the design of the analytic process. The code is written in JAVA.

Rapidminer covers a magnificent range of real of real-world data mining tasks and its applications. Due to the unification of its functional range and leading-edge technologies Rapidminer has become the world-wide leading open-source data mining solution to mine the data. Formerly known as YALE (Yet Another Learning Environment) [6].

Initially, it was released in 2006 under the license of AGPL. It was founded by Ingo Mierswa and Ralf Klinkenberg.

A. Characteristics of RapidMiner:

- Easy to use.
- Easily integrate our own specialized algorithms into RapidMiner by leveraging open extension APIs.
- List of data sources includes Excel, Access, Oracle, IBM, Microsoft SQL, and MySql.
- Allows working with large data sources by breaking the limitations of traditional data analysis tools.
- Runs on all major platforms and operating system.
- Save time by identifying possible errors, and get suggested quick fixes.

- It includes all the tools need to make data work from data preparation to model building and validation.
- Maintain and process with repositories.

B. RapidMiner Visualization:

RapidMiner has strong visualizations for predictive analytics. It allows turning data into customizable, exportable different types of automated charts like histogram, pie chart, bar chart. Statistical views process the data. It supports a maximum visual impact with zooming, panning.

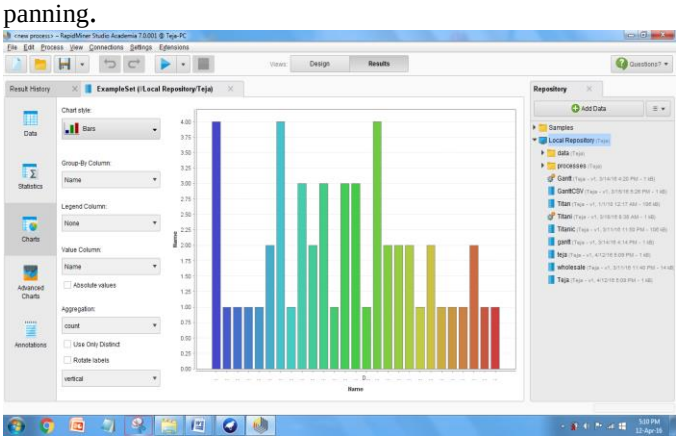

Fig. 3. RapidMiner visualization

C. Advantages:

- RapidMiner supports multiple interfaces.
- It allows handling different models and tasks easily.
- Easy drag & drop process.
- Easy to predict with less effort and optional code.
- Reuse with R and Python code.
- Easily connect with MySql database.
- The data can be stored in the cloud.
- Various extensions are provided which adds new features and increases the productivity [7].

D. Disadvantages:

- It is difficult to work with two files at same time.

E. Algorithms implemented:

The algorithms can be implemented for clustering, regression, neural networks, time series, and classification.

- Principal Component Analysis.
- K-Means.
- Bayesian classification.
- Optimize by generation.
- Generalized Hebbian algorithm.
- Distributed Analytical algorithm.
- Frequent Pattern Growth Algorithm.

Examples of usage:

- Fraud transaction detection in credit cards
- Estimating insurance premium
- Sentiment analysis on text data

F. Working with RapidMiner:

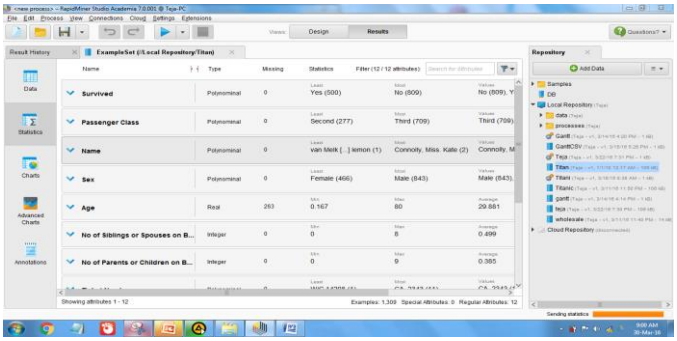

Fig. 4. Statistics of an excel file with the class names, values in RapidMiner.

V. WEKA

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It s written in Java and runs on almost any platform. It supports data mining tasks, data preprocessing, clustering, classification, regression, visualization.

WEKA stands for Waikato Environment for Knowledge Analysis. There are java and non java versions of Weka tool. It was founded in New Zealand [8].

A. Characteristics of Weka:

- Easy to access because of its graphical user interface.
- Large collection of different data mining algorithms.
- It can assist an organization to evaluate and analyze their information in more effective terms.
- Allows individuals to look into their information from a variety of distinct factors as is it incredibly user- friendly [9].
- Freely available under the GNU general Public License.

B. Weka Visualization:

Weka can visualize cluster, classification, association. It has decision-making trees to understand the data.

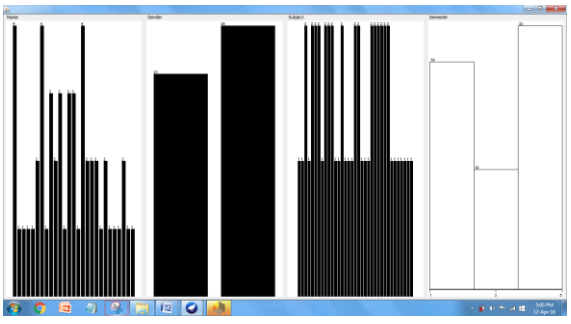

Fig. 5. Visualization and analysis of the data in WEKA.

C. Advantages:

- Weka is possible to handle big data and can connect with databases like JDBC, MySql.
- Ease of use due to its graphical features.
- The code is written in Java language.
- It contains large amounts of predefined data mining algorithms.
- The user interface consists of cluster, classifier, association.
- It can read the files from multiple database formats.
- It can import .arff, .csv file only.
- Easy to handle multiple data repositories [10].

D. Disadvantages:

- It cannot support multiple formats.
- Difficult to store the parameters and datasets.
- Does not have parameter optimization facility.

E. Algorithms implemented:

By default Weka tool has several algorithms to be used. Classification and clustering algorithms can be implemented in Weka.

- FP-growth algorithm.
- Tertius algorithm.
- Clonal Selection Algorithm.
- X means algorithm.

Examples of usage:

- Grid computing with Weka.
- Text classification methods.
- Census data mining and analysis using Weka.

F. Working with WEKA:

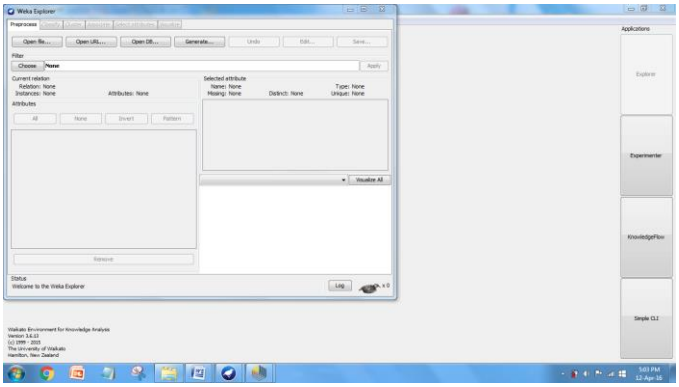

Fig. 6. Importing the files from Explorer in WEKA

VI. ANALYSIS OF SAMPLE EXPERIMENT CONDUCTED WITH R, RAPIDMINER & WEKA

The following table shows the data collected on social behavior.

TABLE I. TABLE SHOWS SOCIAL BEHAVIOR

Keyword	Tweet	Date	Score	Result
Motor Insurance	1000	01-05-16	14	Positive
Third Party	1000	12-05-16	-23	Negative
License	1000	13-05-16	-30	Negative
Vehicle	1000	13-05-16	19	Positive
Premium	1000	13-05-16	48	Positive
Driver Policy	986	04-04-16	-10	Negative
Underwriting	1000	16-05-16	38	Positive
Insurance lapse	38	02-02-16	-1	Negative
Claims	1000	05-05-16	-74	Negative
Gap Insurance	654	01-05-16	8	Positive
Bajaj	500	15-05-16	16	Positive
Tort insurance	31	02-01-16	-7	Negative
Personal Accident Cover	10	22-02-16	1	Positive
No claim bonus	489	02-05-16	6	Positive

The above table contains five columns with the data collected regarding social behavior to find the result depending on the score for the keywords. The data is collected in 2016 from January to May. The analysis is plotted by graphs and charts for all the three tools R, RapidMiner and Weka to visualize in a better way. The graphs are plotted between keywords and scores to obtain the results accurately. The score ranges from -74 to 48.

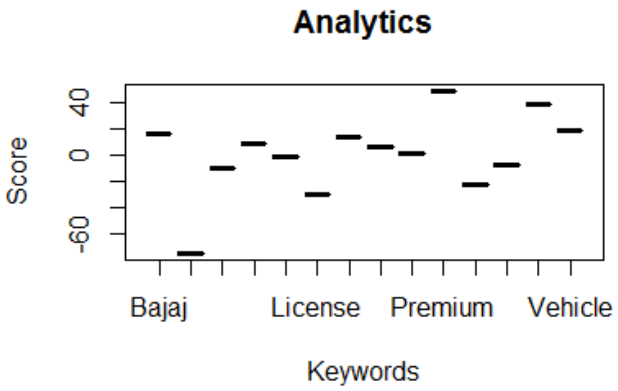

Fig. 7. Plotting between the keywords and scores in R

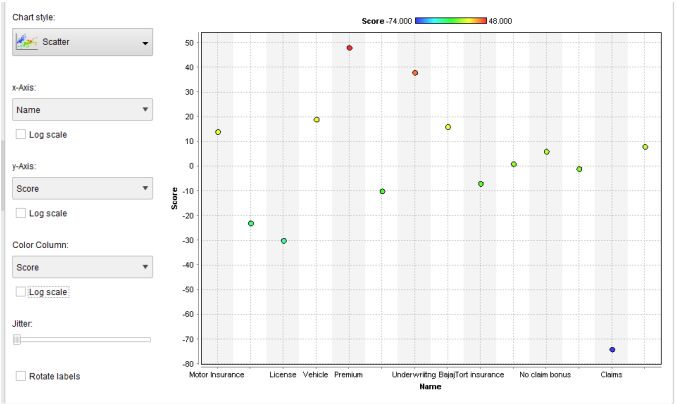

Fig. 8. Plotting a scattered graph between keywords and Score in Rapidminer

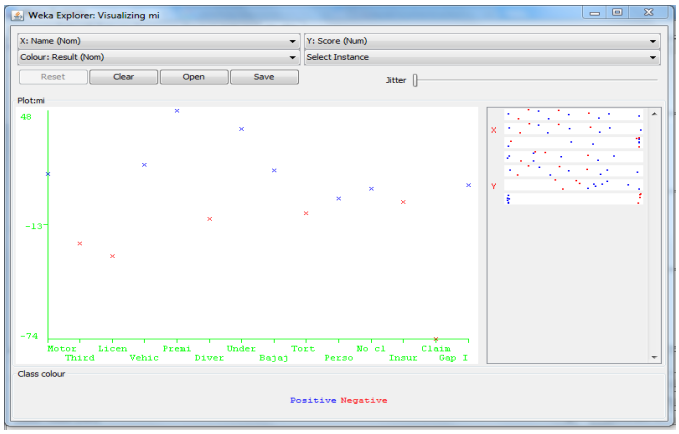

Fig. 9. Plotting a graph between keywords and score in weka.

The above three figures, fig7,8&9 show the graphs plotted from the same data.

If the score is a positive number, then it has positive social behavior which means people have a good opinion on a particular word. If the score is a negative number, then it has negative social behavior which means people have bad opinion on a particular word.

For example, consider the keyword Motor Insurance from table I and its score is 14 which is positive. So the result has positive social behavior for Motor Insurance. Similarly, consider the keyword Third party and its score is -23 which is negative. So the result has negative social behavior.

In the Fig 7, the graph is drawn using R tool which is basic in nature and displays lines for the points with approximate results. The analytics can do better in R tool.

In the Fig 8, the graph is plotted using RapidMiner with the range of scores and the graph is scattered as points between x-axis and y-axis with multiple colors.

In the Fig 9, the graph is plotted using Weka with exact values between -74 to 48 as scores for the respective keywords and it is a scattered graph.

So, for a better visualization RapidMiner is best and can plot graphs for selected columns only. For analytics purpose R is best to handle multiple files and can interact with databases. For implementing any algorithm, Weka can do the best.

VII. COMPARISON STUDY ON TOOLS

TABLE II. TOOL COMPARISON TABLE

Characteristics	R	RapidMiner	Weka
Purpose	Statistical & clustering analysis	Data Mining, Classification	Data mining, association
Data import	.xlsx, csv, txt.	.csv, .xlsx, binary files	.csv, .arff
Specialization	It has a large number of users, in the fields of bio-informatics and social science.	Specialized for solutions that include predictive analysis and statistical computing.	Weka is useful for machine learning techniques and association rules.
Advantages	Purely statistical	Visualization, Parameter optimization.	Easy to use.
Programming Language	C	Java	Java

CONCLUSION

Throughout this paper, the three data mining tools, namely R, Weka and Rapid Miner are discussed. Each tool has its own importance and its usage purely depends on the user and requirements of the algorithm. The research work is going on in data mining to improve big data analytics and thereby improving the quality of tools. There also exists few more tools like Orange, Tanagra, KNIME etc and research and discussion on these tools can be done as a future enhancement.

REFERENCES

[1] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”, 2nd ed., San Fransisco, 2006, pp.6-17.
[2] Qiang Yang, “Challenging problems in data mining research”, International Journal of Imformation & Decision Making, Vol. 5, N0.4 (2006).
[3] Richard Cotton, “Learning R”, O’Rielly, United States of America, 2013, pp.3-7.
[4] Advantages and Disadvantages of R – http://analyticstrainings.com/
[5] Comprehensive guide to Data visualization in R- http://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/
[6] Kalpana Rngra, Dr. K.L.Bansal, “International journal of advanced research in computer science and software engineering”, 2014, pp.3-8.
[7] Vijay Kotu, Bala Deshpande, “Predictive Analytics and Data Mining”, 2015, pp.20-23.
[8] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka/,
[9] Ian H. Witten, Eibe Frank, Mark A. Hall, “Data Mining practical machine learning tools and techniques”, 3rd ed, 2011, pp.403-415.
[10] George M. Marakas, Modern Data Warehousing, Mining, and Visualization, Pearson Education, New Delhi, 2005, pp 125-170.