# Comparative Analysis of various Data Mining Tools

**S.Subhitsha[1] S.Selvakumar[2] V.P.Sumathi[3]**
[1,2]Student [3]Assistant Professor (SRG)
[1,2,3]Department of Computer Science & Engineering
[1,2,3]Kumaraguru College of Technology, Coimbatore, India

*Abstract*— Today, the rapid growth in data has created the need to analyse large volumes of data for which several data mining tools have been developed over decades. Data Mining and its applications play an important role to analyse the data and provide useful insights to users. Several Data Mining tools are developed by research communities and data analysts. These tools provide users an easy to use environment to analyse data. This paper provides an overview of five most popular software tools: Weka, Orange, Rapid Miner, R, KEEL and three web based tools: PREDICT, Breast cancer treatment outcome calculator and Adjuvant online. The tools are compared based on their type, features, pros and cons, and the file formats supported by them. The common issues in the existing tools have been identified and a future work to address the issues has been proposed.

***Key words:*** Data Mining, Data Mining Tools, Weka, Orange, Rapid Miner, R, KEEL

## I. Introduction

There has been a steady increase in the amount of information and data which is stored in electronic format for the last few decades. The size of the data has reached terabytes and is increasing continuously. The recent advancements in the field of Information Technology have made data mining simple and easily affordable. The increased size of the data has made manual analysis very arduous and led to the development of several data mining tools to manipulate data. Such requirements have increased the development of automated tools which assists in transforming data into information and knowledge.

Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. Data mining involves an integration of techniques from multiple disciplines such as database and data warehousing technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis.

Data mining has many application fields such as marketing, business, science and engineering, healthcare analysis, economics, games and bioinformatics. Due to its widespread applications and complexity involved in building data mining applications, a large number of Data mining tools have been developed over decades. Every tool has its own merits and demerits. The tools are developed by a research community and data analysis enthusiasts with open-source licenses. This development style offers a means of incorporating the diverse experiences. The key features of Data mining tools include: User-friendly environment, ability to handle big data, suitable for use by novice users and limited programming knowledge.

This study presents a comparison of the five most popular software tools: Weka, Orange, Rapid Miner, R, KEEL and three web based tools: Predict, Breast cancer treatment outcome calculator and Adjuvant online. The tools are compared based on their features, advantages and drawbacks.

## II. Data Mining Tools

Today various data mining tools are available to handle huge volumes of data. Data mining tools are used to predict future trends, behaviours, allowing business to make proactive and knowledge driven decision. It is also widely used in healthcare to predict the future course of a disease. Various Data mining techniques and algorithms have been implemented on these tools to extract the information and also to check their efficiency and accuracy. The technical features of various tools along with their advantages and disadvantages are discussed in the following section.

## III. Software Tools

### A. Weka

Weka (Waikato Environment for Knowledge Analysis) is a very popular open source java based data mining tool. It can be used to implement various Machine learning and data mining algorithms. Weka supports a variety of data formats like CSV, arff and Binary. Weka supports many model evaluation procedures and metrics but lacks many data survey and visualisation methods. It is more oriented towards classification and regression problems and less towards descriptive statistics and clustering methods, although some improvements were made recently with respect to clustering.

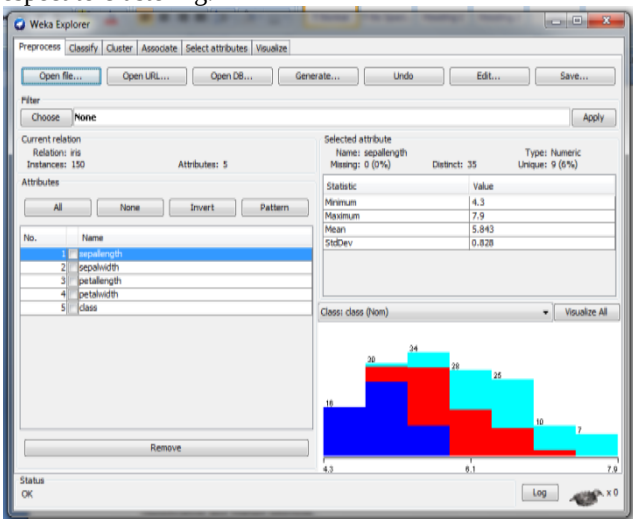


Fig. 1: Weka User Interface

#### 1) Features

- Open source and freely available.
- Platform independent.
- Flexible facilities for scripting experiments.
- Easily usable by people who are not data mining experts.

 

*2) Advantages*

- It supports various data mining tasks like data pre-processing, clustering, classification and feature selection.
- It can be integrated into java packages.
- Weka is best suited for mining association rules.

*3) Disadvantages*

- It does not perform well with big data.
- Memory is limited and hence lesser performance.
- Does not have the facility to save parameters for scaling to future datasets.
- Lack of data visualisation techniques.

*B. Rapid Miner*

Rapid Miner is a XML based data mining tool used to implement various machine learning and data mining processes. It is a popular tool to implement classification and clustering algorithms. It provides drag and drop to design analytics process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures. Everything in Rapid Miner is focused on processes that may contain subprocesses. Processes contain operators in the form of visual components. Operators are implementations of DM algorithms, data sources, and data sinks. The dataflow is constructed by drag-and-drop of operators and by connecting the inputs and outputs of corresponding operators.

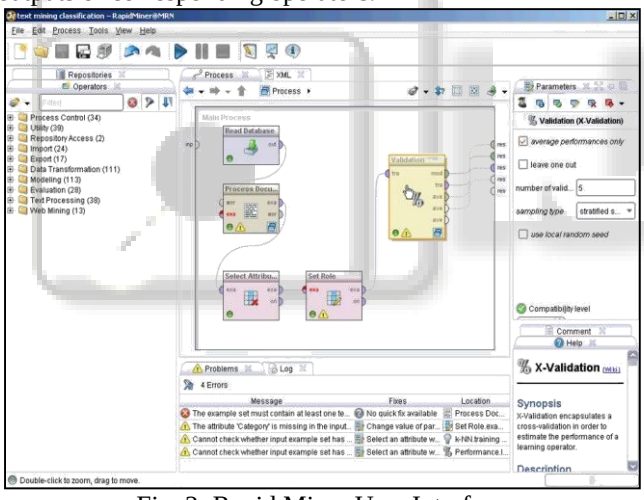

Fig. 2: Rapid Miner User Interface

*1) Features*

- Graphical user interface.
- Analysis processes design.
- Multiple data management methods.
- Data from file, database, web, and cloud services.
- In-memory, in-database and in-Hadoop analytics.
- Application templates.
- GUI or batch processing.

*2) Advantages*

- Rapid Miner offers numerous procedures, especially in the area of attribute selection and for outlier detection, which no other solution offers.
- Easy to debug the errors.
- It easily reads and writes Excel files and different databases.

*3) Disadvantages*

- The software requires ability to manipulate SQL statements and files because it is suited for people accustomed to working with database files.
- Limited partitioning abilities for dataset to training and testing sets.

*C. Orange*

Orange is a Python-based tool for data mining. Functionalities are visually represented by different widgets. A short description of each widget is available within the interface. Programming is performed by placing widgets on the canvas and connecting their inputs and outputs. Data Mining is done through visual programming or python scripting. Orange Canvas offers a structured view of supported functionalities grouped into nine categories: data operations, visualization, classification, regression, evaluation, unsupervised learning, association, visualization using Qt, and prototype implementations.

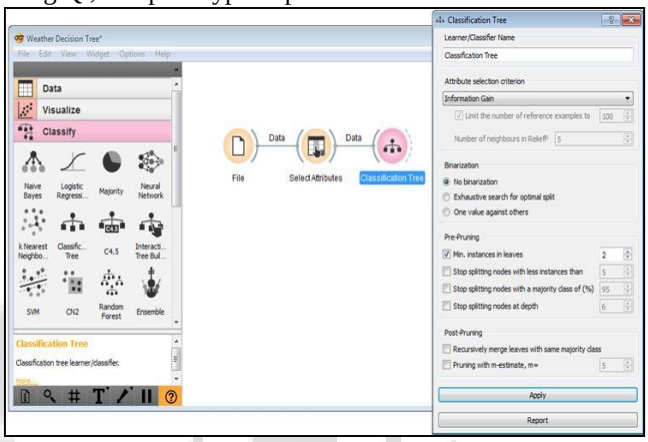

Fig. 3: Orange User Interface

*1) Features*

- Open and free software
- Platform independent software
- Programming support
- Scripting interface

*2) Advantages*

- It works both as a script and GUI.
- The interface provides a pleasant user interface.
- Provides better data visualisation.

*3) Disadvantages*

- The number of widgets is limited.
- Limited list of Machine learning algorithms.
- Visualisation is not appealing.

*D. R*

Revolution is a free software programming language and software environment for statistical computing and graphics. The tool offers only a simple GUI with command-line shell for input. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R provides a wide variety of graphics and statistical techniques such as linear and non-linear modelling, classical statistical tests, time series analysis, classification clustering and is highly extensible.

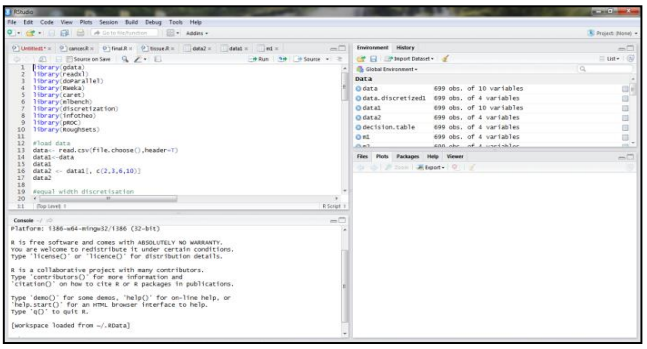

Fig. 4: R GUI

*1) Features*

- Free and open source software
- Supports unique data visualisation
- Allows statisticians to do very intricate and complicated analyses

*2) Advantages*

- Suited for statistical computing.
- The tool offers a simple GUI with command line shell for input.
- Offers very fast implementation of algorithms compared to other tools.

*3) Disadvantages*

- It does not provide a user friendly environment because all commands need to be entered in the command shell.
- R language is difficult to learn thoroughly enough to become productive in data mining.
- Less specialised towards data mining.

*E. KEEL*

KEEL (Knowledge Extraction based on Evolutionary Learning) is an open source Java software tool that can be used for a large number of different knowledge data discovery tasks. KEEL provides a simple GUI based on data flow to design experiments with different datasets and computational intelligence algorithms in order to assess the behaviour of the algorithms. It contains a wide variety of classical knowledge extraction algorithms, pre-processing methods, computational intelligence based learning algorithms, hybrid models and statistical methodologies for contrasting experiments. It allows performing a complete analysis of new computational intelligence proposals in comparison to existing ones.

*1) Features*

- Keel is a software tool to assess evolutionary algorithms for Data Mining problems.
- Machine learning tool.
- Platform independent.

*2) Advantages*

- It contains a library with evolutionary learning algorithms which reduces programming work.
- Due to the use of a strict object-oriented approach for the library and software tool, these can be used on any machine with Java
- It contains a big collection of classical knowledge extraction algorithms.

*3) Disadvantages*

- Efficiency is restricted by the number of algorithms it support as compared to other tools.

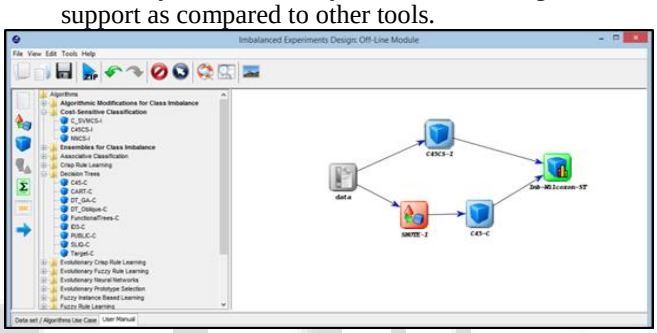

Fig. 5: KEEL User Interface

| Tool | Input Format | Output Format |
|---|---|---|
| WEKA | .csv, .dat, .arff, From database | Bar chart, line chart, scatter plot, histogram |
| Rapid Miner | Text file, .csv, audio, PDF, HTML, images, XML | Bar chart, line chart, pie chart, histogram, box, 3D scatter plots |
| Orange | .csv,.xls,.txt | Box plot, histograms, scatter plot, mosaic display, sieve diagram |
| R | Text file, binary file, .csv, from database, XML, Pdf, HTML, images | Bar chart, line chart, pie chart, histogram, box, scatter, conditioning plot, 3D scatter plots |
| KEEL | ASCII,.dat | Training(.tra), Testing(.tst) |

Table 1: Input and Output Format supported by Tools

| Tool | Type | Features |
|---|---|---|
| WEKA | – Machine Learning | – 49 data pre-processing methods<br>– 76 classification and regression algorithms<br>– 8 clustering algorithms<br>– 10 algorithms for feature selection<br>– 3 GUI(Explorer, Experimenter and Knowledge flow) |
| Rapid Miner | – Statistical analysis<br>– Data mining<br>– Predictive analytics. | – Supports 22 file formats<br>– 20 functions for analysis and data handling<br>– 80 data pre-processing methods<br>– 31 classification and regression algorithms<br>– 9 clustering algorithms |
| Orange | – Machine learning<br>– Data mining<br>– Data visualization | – 8 data pre-processing methods<br>– 21 classification and regression algorithms<br>– 1 clustering algorithm |

 

| | | |
|---|---|---|
| R | – Statistical computing | – R supports various data mining tasks by providing several libraries. |
| KEEL | – Machine learning | – 5 classification algorithms<br>– 6 regression algorithms<br>– 7 data-processing methods |

Table 2: Comparison of Features

## IV. ONLINE TOOLS

### A. *Predict Tool*

PREDICT is a mathematical model, accessed by the internet and has been designed for patients and doctors to help them decide on the ideal course of treatment following breast cancer surgery. Using the Eastern Cancer Registration and Information Centre (ECRIC) dataset, information was collated for 5,694 women who had surgery for invasive breast cancer in East Anglia from 1999 to 2003. The total number of deaths at Years 5 and 8 after diagnosis was estimated by summing the breast-specific and competing mortality. Observed and predicted deaths were compared using a standard Chi-squared test. Model discrimination was evaluated by calculating the area under the receiver-operator-characteristic (ROC) curve (AUC) calculated for breast cancer specific and overall deaths at Year 8 past diagnoses. The ROC curve plots sensitivity against 1-specificity at different predicted risk thresholds.

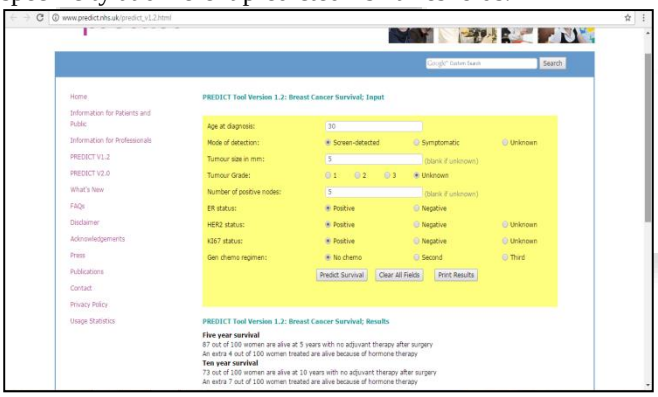


Fig. 6: PREDICT Input Screen

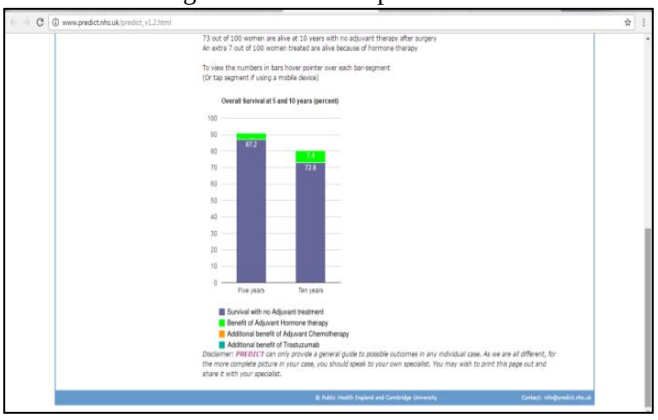


Fig. 7: PREDICT Output Screen

### B. *Breast cancer treatment outcome calculator*

The goal of this tool is to provide medical professionals with web-based calculators for accurately predicting the clinical outcome for individual cancer patients, as well as for accurately estimating the impact of various treatment choices on that outcome as shown in Figure 7.

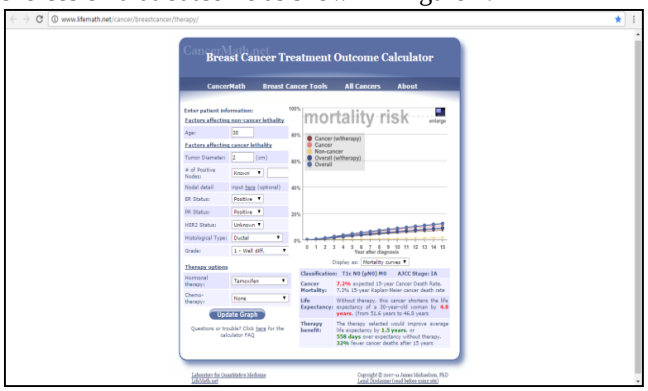


Fig. 8: Outcome Calculator

### C. *Adjuvant Online*

Adjuvant Online was developed using data from the United States Surveillance Epidemiology and End Results (SEER) tumor registry. It has been widely used as a clinical tool that helps patients reach decisions on the option best for them, according to their own preferences and priorities. It is one of the first computer programs designed to help doctors and patients understand the implications of different treatment options tailored to individual risk factor. This was developed for patients with early invasive breast cancer. The output screen of the tool is shown in Figure 8.

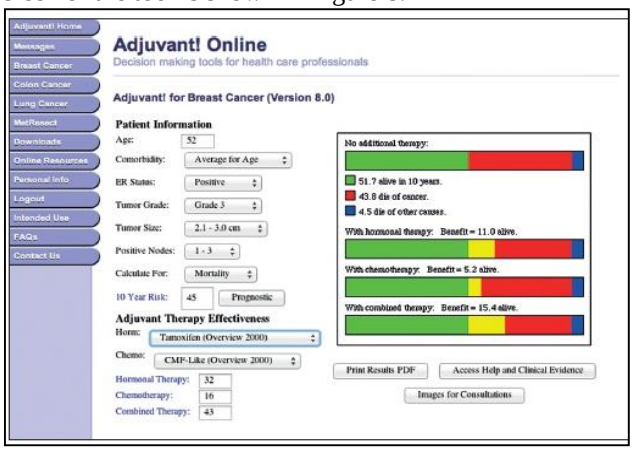


Fig. 9: Adjuvant Online

| Tool | Description | Algorithms Used | Drawbacks |
|---|---|---|---|
| Predict Tool | A mathematical model, accessed by the internet and has been designed for patients and doctors to help them decide on the ideal course of treatment following breast cancer surgery | Chi-squared test | – The chi-square test does not give much information about the strength of the relationship.<br>– The chi-square test is sensitive to sample size.<br>– Small sample size was used and hence may not be reliable. |

 

| | | | |
|---|---|---|---|
| Breast cancer treatment outcome calculator | Web-based calculators for accurately predicting the clinical outcome for individual cancer patients, as well as for accurately estimating the impact of various treatment choices on that outcome. | Pure mathematical models | – Pure mathematical models were used to predict cancer.<br>– This tool is suitable for use only by medical professionals.<br>– Results are difficult to interpret by common people. |
| Adjuvant online | A clinical tool that helps patients reach decisions on the option best for them, according to their own preferences and priorities. | Not mentioned | – It has been found to underestimate risk in some groups.<br>– The statistical model behind this calculator has never been published. |

Table 3: Comparison of Online Tools

## V. SUMMARIZATION OF ISSUES AND FUTURE WORK

The common issues identified with the existing tools are:

- Does not perform well with big data
- Limited memory
- Requires ability to manipulate SQL statements
- Number of algorithms is limited
- Some tools require programming knowledge
- Does not provide user friendly environment
- Online tools are suitable for use only by medical professionals
- Results are difficult to interpret by common people

As a future work, an online web application tool will be developed such that it focuses on scaling with new changes and dataset. The issues to be addressed in the future work are:

- Handles large dataset
- Does not require ability to manipulate SQL statements
- Programming knowledge not required
- User-friendly environment
- Suitable for use by common people

## VI. CONCLUSION

The study presented the specific details along with description of various open source data mining tools enlisting their advantages and limitations. The complete analysis of these data mining and web mining software tools focuses the usefulness and importance of these tools by considering various aspects. Analysis presents various benefits of these data mining tools with respect to functionalities, advantages and disadvantages, and compared them accordingly.

## REFERENCES

[1] Rangra, Kalpana, and K. L. Bansal. "Comparative study of data mining tools." International journal of advanced research in computer science and software engineering 4.6 (2014).
[2] Wahbeh, Abdullah H., et al. "A comparison study between data mining tools over some classification methods." (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence (2011): 18-26.
[3] Karur,Parminder, Qamar. "Comparison of various tools for data mining." (IJERT) International Journal of Engineering Research and Technology (2014): 393-397.
[4] Chauhan, Neha, and Nisha Gautam. "Parametric Comparison of Data Mining Tools".
[5] Solanki, Harshvardhan. "Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory." International Journal of Computer Applications 75.16 (2013).
[6] David, Satish Kumar, Amr TM Saeb, and Khalid Al Rubeaan. "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics." Computer Engineering and Intelligent Systems 4.13 (2013): 28-38.
[7] Jovic, Alan, Karla Brkic, and Nikola Bogunovic. "An overview of free software tools for general data mining." Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on IEEE, 2014.
[8] Patel, Priti S., and S. G. Desai. "A Comparative Study on Data Mining Tools." International Journal 4.2 (2015).
[9] Al-Khoder, Ahmad, and Hazar Harmouch. "Evaluating four of the most popular Open Source and Free Data Mining Tools".
[10] Hirudkar, Arpita M., and S. S. Sherekar. "Comparative analysis of data mining tools and techniques for evaluating performance of database system." Int J Comput Sci Appl 6.2 (2013): 232-237.
[11] Wimmer, Hayden, and Loreen Marie Powell. "A comparison of open source tools for Data Science." Journal of Information Systems Applied Research 9.2 (2016): 4.
[12] Elder, John F., and Dean W. Abbott. "A comparison of leading data mining tools." Fourth International Conference on Knowledge Discovery and Data Mining. 1998.
[13] Chen, Xiaojun, et al. "A survey of open source data mining systems." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2007.
[14] Jain, Divya. "A Comparison of Data Mining Tools using the Implementation of C4.5 Algorithm." International Journal of Science and Research Vol3 8 (2014)

 