Volume 4, Issue 6, June 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Comparative Study of Data Mining Tools

Kalpana Rangra
Research scholar
Department of Computer Science
Himachal Pradesh University
Shimla, India

Dr. K. L. Bansal
Professor
Department of Computer Science
Himachal Pradesh University
Shimla, India

Abstract: -Today the rapid development of information technology and adoption of its several applications has created the revolution in business and various fields significantly. The growing interest in business using electronics and technology has brought vital improvement in data mining field also, since it's an important part of data accessibility. Data mining and it's applications can be viewed as one of the emerging and promising technological developments that provide efficient means to access various types of data and information available worldwide. Not only this, these applications also aids in decision making. A better understanding of these applications helps in aking choice among all available application and tools. The paper gives the comprehensive and theoretical analysis of six open source data mining tools. The study describes the technical specification, features, and specialization for each selected tool along with its applications. By employing the study the choice and selection of tools can be made easy.

Keywords: Data, Data Mining, Data Mining Tools, Open Source Tools, Technical Specification.

I. Introduction

There has been a dramatic increase in amount of information and data which is stored in electronic format since last few decades. The size of data base has been in the process of continuous increment and has reached up to terabytes. This explosive rate of data increment is growing day by day and estimations tell that the amount of information in world doubles every 20 months. Thus the most important question concerned with data is its retrieval which finds the most suitable answer in data mining. Data mining is the process of extraction of predictive information from large data masses. It can also be described as a process of analyzing data from different perspectives and summarizing it into useful information.

With a vast history deeply rooted in machine learning, artificial intelligence, database along with statistics data mining was coined very early. Data mining is strongly associated with data science which involves manipulation and classification of data by applying statistical and mathematical concepts. Data mining is an important phase in knowledge discovery and includes application of discovery and analytical methods on data to produce specific models across data. Data are available everywhere. It can be used to predict the future. Usually the statistical approach is used. Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines. Due to the widespread availability of huge, complex, information-rich data sets, the ability to extract useful knowledge hidden in these data and to act on that knowledge has become increasingly important in today's competitive world .Thus data mining is analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to data owner. [1].

Briefly, data mining is an approach to research and analysis. [2] It is exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. [3]

Sometime, data may be in different formats as it comes from different sources, irrelevant attributes and missing data. Therefore, data needs to be prepared before applying any kind of data mining. Data mining is also known under many other names, including knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing.[4]Many researchers and practitioners use data mining as a synonym for knowledge discovery but data mining is also just one step of the knowledge discovery process. All the techniques follow an automated process of knowledge discovery (KDD) i.e., data cleaning, data integration, data selection, data transformation, data mining and knowledge representation [5]

Types of data that can be mined

- **Flat files**: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.
- **Relational Databases**: Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

- **Data Warehouses**: A data warehouse as a store house, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof.
- **Transaction Databases**: A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items. For example, in the case of the video store, the rentals table.
- **Multimedia Databases**: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies.
- **Spatial Databases**: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.
- **World Wide Web**: The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications. Conceptually, the World Wide Web is comprised of three major components: The content of the Web, which encompasses documents available; the structure of the Web, which covers the hyperlinks and the relationships between documents; and the usage of the web, describing how and when the resources are accessed.
- **Time-Series Databases**: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.
-

II. A Brief Over view of data mining tools

Data mining has a wide number of applications ranging from marketing and advertising of goods, services or products, artificial intelligence research, biological sciences, crime investigations to high-level government intelligence. Due to its widespread use 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 advantages and disadvantages. [6]

Within data mining, there is a group of tools that have been developed by a research community and data analysis enthusiasts; they are offered free of charge using one of the existing open-source licenses. An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences

Data mining provides many mining techniques to extract data from databases. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions.

The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult. [7]

The top six open source tools available for data mining are briefed as below.

A . Weka

Waikato Environment for Knowledge Analysis. Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.

1)Technical Specification:

- First released in 1997.
- Latest version available is WEKA 3.6.11.
- Has GNU general public license.
- Platform independent software.
- Supported by Java
- Can be downloaded from www.cs.waikato.ac.

2)General Features

- Weka is a Java based open source tool data mining tool which is a collection of many data mining and machine learning algorithms, including pre-processing on data, classification, clustering, and association rule extraction
- Weka provides three graphical user interfaces i.e. the Explorer for exploratory data analysis to support preprocessing, attribute selection, learning, visualization, the Experimenter that provides experimental environment for testing and evaluating machine learning algorithms, and the Knowledge Flow for new process

model inspired interface for visual design of KDD process. A simple Command-line explorer which is a simple interface for typing commands is also provided by weka .

3)Specialization:

- Weka is best suited for mining association rules .
- Stronger in machine learning techniques.
- Suited for machine Learning.

Advantages

- It is also suitable for developing new machine learning schemes.[8]
- Weka loads data file in formats of ARFF, CSV, C4.5, binary. Though it is open source, Free, Extensible, Can be integrated into other java packages.

Limitation

- It lacks proper and adequate documentations and suffers from “Kitchen Sink Syndrome” where systems are updated constantly.
- Worse connectivity to Excel spreadsheet and non-Java based databases.
- CSV reader not as robust as in Rapid Miner.
- Not as polished.
- Weka is much weaker in classical statistics.
- Does not have the facility to save parameters for scaling to apply to future datasets.
- Does not have automatic facility for Parameter optimization of machine learning/statistical methods

B. KEEL

Knowledge Extraction based on Evolutionary Learning is an application package of machine learning software tools. KEEL is designed for providing solution to data mining problems and assessing evolutionary algorithms. 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 providing scientific and research methods.

1)Technical Overview

- First released in 2004.
- Latest version available is KEEL 2.0.
- Licensed by GNU, general public license.
- Can run on any platform.
- Supported by java language.
- Can be downloaded from www.sci2s.ugr.es/keel.

2)Specialization

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

Advantages

- It includes regression, classification, clustering, and pattern mining and so on.
- It contains a big collection of classical knowledge extraction algorithms, preprocessing techniques (instance selection, feature selection, discretization, imputation methods for missing values etc.), Computational Intelligence based learning algorithms, including evolutionary rule learning algorithms based on different approaches (Pittsburgh, Michigan and IRL), and hybrid models such as genetic fuzzy systems, evolutionary neural networks etc.[9]

Limitation:

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

C. R

Revolution 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. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

1)Technical Specification

- First released in 1997
- Latest version Available is 3.1.0
- Licensed by GNU General Public License
- Cross Platform
- C, Fortran and R
- www.r-project.org

2)General Features

- The R project is a platform for the analysis, graphics and software development activities of data miners and related areas.

- R is a well-supported, open source, command line driven, statistics package. There are hundreds of extra “packages” freely available, which provide all sorts of data mining, machine learning and statistical techniques. .
- It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems

3)Specification:

- It has a large number of users, in particular in the fields of bio-informatics and social science. It is also a free ware replacement for SPSS.
- Suited for Statistical Computing.

Advantages.

- Very extensive statistical library.
- It is a powerful elegant array language in the tradition of APL, Mathematica and MATLAB, but also LISP/Scheme.
- Ability to make a working machine learning program in just 40 lines of code
- Numerical programming is better integrated in R
- R has better graphics.
- R is more transparent since the Orange are wrapped C++ classes.
- Easier to combine with other statistical calculations.
- Import and export of data from spreadsheet is easier in R, spreadsheet are stored in a data frames that the different machine learning algorithms are operating on.
- Programming in R really is very different, you are working on a higher abstraction level, but you do lose control over the details.

Limitation:

- Less specialized towards data mining.
- There is a steep learning curve, unless you are familiar with array languages

D. KNIME

Konstanz Information Miner, is an open source data analytics, reporting and integration platform. It has been used in pharmaceutical research, but is also used in other areas like CRM customer data analysis, business intelligence and financial data analysis. It is based on the Eclipse platform and, through its modular API, and is easily extensible. Custom nodes and types can be implemented in KNIME within hours thus extending KNIME to comprehend and provide first-tier support for highly domain-specific data format.

1)Technical Specification

- Released on 2004.
- Latest version available is KNIME2.9
- Licensed By GNU General Public License
- Compatible with Linux ,OS X, Windows
- Written in java
- www.knime.org

2)General Features

- Knime, pronounced “naim”, is a nicely designed data mining tool that runs inside the IBM’s Eclipse development environment.
- It is a modular data exploration platform that enables the user to visually create data flows (often referred to as pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models.
- The Knime base version already incorporates over 100 processing nodes for data I/O, preprocessing and cleansing, modeling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others.

3)Specification

- Integration of the Chemistry Development Kit with additional nodes for the processing of chemical structures, compounds, etc.
- Specialized for Enterprise reporting, Business Intelligence, data mining.

Advantages

- It integrates all analysis modules of the well-known. Weka data mining environment and additional plugins allow R-scripts to be run, offering access to a vast library of statistical routines. [8]
- It is easy to try out because it requires no installation besides downloading and un archiving.
- The one aspect of KNIME that truly sets it apart from other data mining packages is its ability to interface with programs that allow for the visualization and analysis of molecular data

Limitations:

- Have only limited error measurement methods .
- Has no wrapper methods for descriptor selection.
- Does not have automatic facility for Parameter optimization of machine learning/statistical methods.

E. RAPIDMINER

is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures.

1)Technical specification:

- Released on 2006
- Latest version available is Rapid miner 6.
- Licensed by AGPL Proprietary
- Cross platform i.e. can be installed on any operating system
- Language Independent
- Can be downloaded from www.rapidminer.com.

2)General Features

- Rapid miner is an environment for machine learning and data mining processes.
- It represents a new approach to design even very complicated problems by using a modular operator concept which allows design of complex nested operator chains for huge number of learning problems.
- Rapid miner uses XML to describe the operator trees modeling knowledge discovery process.
- It has flexible operators for data input and output file formats.
- It contains more than 100 learning schemes for regression classification and clustering analysis. [10].
- Rapid miner supports about twenty two file formats. [7]
- Rapid Miner has a lot of functionality, is polished and has good connectivity.
- Rapid Miner includes many learning algorithms from WEKA.
- Solid and complete package.
- It easily reads and writes Excel files and different databases.
- You program by piping components together in a graphic ETL work flows.
- If you set up an illegal work flows Rapid Miner suggest Quick Fixes to make it legal.

3)Specialization

- Rapid Miner provides support for most types of databases, which means that users can import information from a variety of database sources to be examined and analyzed within the application.
- Specialized for Business solutions that include predictive analysis and statistical computing.

Advantages

- Has the full facility for model evaluation using cross validation and independent validation sets.
- Over 1,500methods for data integration, data transformation, analysis and, modelling as well as visualization – no other solution on the market offers more procedures and therefore more possibilities of defining the optimal analysis processes
- .RapidMiner offers numerous procedures, especially in the area of attribute selection and for outlier detection, which no other solution offers.

Limitations:

- Rapid Miner is the data mining software package that is most suited for people who are accustomed to working with database files, such as in academic settings or in business settings. The reason for this is that the software requires the ability to manipulate SQL statements and files.

F. ORANGE

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is implemented in C++ and Python. Its graphical user interface builds upon the cross-platform framework

1)Technical Requirements:

- Developed in 2009.
- Latest version available is Orange 2.7
- Licensed by GNU General Public License
- Compatible with Python, C++,C.
- Can be downloaded from www.orange.biolab.si

2)General Features

- Orange is a component-based data mining and machine learning software suite.
- It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques.
- Data mining in Orange is done through visual programming or Python scripting.

3)Specialization

- Open source data visualization and analysis for novice and experts.

- It contains components for machine learning and add-ons for bioinformatics and text mining. Along with its also packed with features for data analytics.[11].
- Specialized for data visualization along with mining.

Advantages

- It is an open source data mining package build on Python, NumPy, wrapped C, C++ and Qt.
- Works both as a script and with an ETL work flow GUI.
- Shortest script for doing training, cross validation, algorithms comparison and prediction.
- Orange the easiest tool to learn.
- Cross platform GUI.
- Orange is written in python hence is easier for most programmers to learn.
- Has better debugger.
- Scripting data mining categorization problems is simpler in Orange.
- Orange does not give optimum performance for association rules.

Limitations

- Not super polished.
- The install is big since you need to install QT.
- Limited list of machine learning algorithms.
- Machine learning is not handled uniformly between the different libraries.
- Orange is weak in classical statistics; although it can compute basic statistical properties of the data, it provides no widgets for statistical testing.
- Reporting capabilities are limited to exporting visual representations of data models.

III. Comparative Study of tools

The best six of available open source data mining tools were chosen and analytical study was made by taking into account technical specifications and feature.

Table 1 : Technical Overview of best six data mining open source tools

S. N	Tool Name	Release Date	Release date/ Latest version	License	Operating System	Language	Website
1.	RAPID MINER	2006	21November,2013 /6.0	AGPL Proprietary	Cross platform	Language Independent	www.rapidminer.com
2	ORANGE	2009	6 May,2013/2.7	GNU General Public License	Cross Platform	Python C++,C	www.orange.biolab.si
3	KNIME	2004	6December,2013/2.9	GNU General Public License	Linux ,OS X, Windows	Java	www.knime.org
4	WEKA	1993	24 April,2014/3.7.11	GNU General Public License	Cross Platform	Java	www.cs.waikato.ac.nz/~ml/weka
5	KEEL	2004	5 June,2010/2.0	GNU GPL v3	Cross Platform	Java	www.sci2s.ugr.es/keel
6	R	1997	10 April,2014/3.1.0	GNU General Public License	Cross Platform	C, Fortran and R	www.r-project.org

The table shown gives the technical overview of the tools which includes name of tool and description of release date, latest version release date, licence, operating system, language and official website.

Table II : Analytics of feature of best six open source data mining tools

S. N	Tool Name	Type	Features
1.	RAPID MINER	Statistical analysis, data mining, predictive analytics.	• More than 20 new functions for analysis and data handling, including multiple new aggregation functions • File operators to operate directly from Rapid Miner • A macro viewer that shows macros and their values in real time during process execution • Intutive GUI
2	ORANGE	Machine learning, Data mining, Data visualization	• Visual Programming, Visualization, • Interaction And Data Analytics • Large toolbox, Scripting interface • Extendable Documentation
3	KNIME	Enterprise Reporting ,Business Intelligence ,Data mining	• Scalability , Intutive user interface ,High extensibility • well-defined API for plugin extensions • sophisticated data handling, intelligent automatic caching of data, Data visualization • Import/export of workflows, Parallel execution on multi-core systems • Command line version for "headless”,“batch executions”,Hilting,
4	WEKA	Machine Learning.	• Forty nine data preprocessing tools, seventy six classification/regression algorithms, eight clustering algorithms, fifteen attribute/subset evaluators, ten search algorithms for feature selection. • three algorithms for finding association rules • three graphical user interfaces – “The Explorer” (exploratory data analysis) – “The Experimenter” (experimental environment) – “The Knowledge Flow” (new process model inspired . • poor documentation
5	KEEL	Machine Learning	• Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Data Visualization ,Discovery Visualization, a user-friendly graphical interface,evolutionary learning
6	R	Statistical Computing	• Data Exploration, Outlier detection, Clustering ,Text Mining, Time Series Analysis , Social Network Analysis ,Parallel Computing, Graphics, Visualization of geo spatial data, Web Application Big data • Data and error handling,requires array language,poor mining,

The given table describes the basic features and functionality provided by described six tools i.e. Rapid miner, R, Weka, R, Keel and Orange.

Table III. : Advantages and Limitations of tools

SN	TOOL	ADVANTAGES	LIMITATIONS
1	RAPID MINER	Visualization, Statistical,Attribute Selection, Outlier detection,parameter optimization	Requires prominent knowledge of database handling
2	ORANGE	Better debugger, Shortest scripts,poor statistics,suitable for novoice Experts	Big installation, Limited reporting capabilities
3	KNIME	Molecular analysis, Mass spectrometry. Chemistry Development kit	Limited error measurements, no wrapper methods for descriptor selection,poor parameter optimazation
4	WEKA	Ease of use,can be extended in RM	Poor documentation,weak classical statistics,poor parameter optimization,weak csv reader

5	KEEL	Evolutionary algorithms,fuzzy systems	Limited algorithms
6	R	Purely statistical	Less specialized for data mining, requires knowledge of array language

The given table enumerates the advantages and limitation of each tool separately.

IV . Results and discussions

Of the six data mining packages that have been examined, KNIME is the package that would be recommended for people who are novices to such software to those who are highly skilled. The software is simply very robust with built-in features and with additional functionality that can be obtained from third-party libraries. Based on the analysis, Weka would be considered a very close second to KNIME because of its many built-in features that require no programming or coding knowledge. In comparison, Rapid Miner and Orange would be considered appropriate for advanced users, particularly those in the hard sciences, because of the additional programming skills that are needed, and the limited visualization support that is provided. It can be concluded from above tables that though data mining is the basic concept to all tool yet, Rapid miner is the only tool which is independent of language limitation and has statistical and predictive analysis capabilities, So it can be easily used and implemented on any system, moreover it integrates maximum algorithms of other mentioned tools.

V. Conclusion and future scope

Open-source data mining suites of today have come a long way from where they were only a decade ago. They offer nice graphical interfaces, focus on the usability and interactivity, support extensibility through augmentation of the source code or (better) through the use of interfaces for add-on modules. They provide flexibility either through visual programming within graphical user interfaces or prototyping by way of scripting languages. The study presented the specific details along with description of various open source data mining tools enlisting the area of specialization. With the recent endeavors of various developers concerning the use of tools in various fields one can expect a more enhanced environment along with more technical improvements. The work can be a helping hand to provide an insight in future to develop an application with more efficiency and availability i.e. a tool can be designed which instead of supporting a specific area can be extended to more fields. The effort may be increased and the development may be a complex procedure but indeed it can result in an efficient product.

References:

[1] Hand David, Mannila Heikki, Smyth Padhraic.: "Principles of data mining", Prentice hall India, pp.1, 2004.

[2]. Sethi I. K., "Layered Neural Net Design Through Decision Trees, Circuits,and Systems", IEEE International Symposium,1990.

[3]. Meheta M., Aggarwall R., Rissamen I. : "SLIQ:A fast Scalable Classifier for Data Mining", In Proc. International Conference Extending data base Technology(EDBI), Avignon, France, March 1996.

[4]. Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.),. Advances in Knowledge Discovery and Data Mining, AAAI Press, Cambridge, 1996..

[5]. Kittipol Wisaeng . "An Empirical Comparison of Data Mining Techniques in Medical Databases", International Journal of Computer Applications (0975 – 8887), Volume 77– No.7, September 2013.

[6]. S.R.Mulik, S.G.Gulawani :" PERFORMANCE COMPARISON OF DATA MINING TOOLS IN MINING ASSOCIATION RULES", International Journal of Research in IT, Management and Engineering (IJRIME), Volume1Issue3 ISSN: 2249- 1619

[7]. Ralf Mikut and Markus Reischl Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011.

[8]. Witten, I.H., Frank, E.: "Data Mining: Practical machine Learning tools and techniques", 2nd addition,Morgan Kaufmann, San Francisco(2005).

[9]. Alcala-Fdez, J.,L., del Jesus, M.J., Ventura, s., Garrell, J.M, Otero, J., Romero,C., bacardit, j., Rivas, V.M., Fernandez, J.C., Herrera., F., : "KEEL: A software tool to Assess Evolutionary Algorithms to Data mining Problems", Soft computing 13:3,pp 307-318(2009).

[10]. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler,T. "YALE: Rapid Prototyping for Complex Data Mining tasks", in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-06), pp. 935-940, 2006.

[11] http://orange.biolab.si/features/

[12] https://github.com/Dans-labs/recommender-systems/blob/.../datamining.r

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

[14] http://www.knime.org/

[15] http://rapidminer.com/