Tools in Data Mining A Comparative Analysis

Neha Walia[1], Arvind Kalia[2]
[1]*Research Scholar, Department of Computer Science*
Himachal Pradesh University, Shimla, Himachal Pradesh, India
[2]*Professor, Department of Computer Science*
Himachal Pradesh University, Shimla, Himachal Pradesh, India

Abstract- In this era of huge available data in every sphere, mining the knowledge from this data is one of the biggest challenges. Data mining has become a catchphrase, which made researchers to think on the existing data mining techniques and also to develop new analytical techniques to deal with this enormous data. With the increase in the development of data mining techniques, numbers of data mining tools are available in this era to implement these techniques. The paper focuses on the evaluation of seven data mining tools namely WEKA, Rapidminer,Orange, KNIME, Apache Mahout, Oracle Data Mining and Rattle along with their important features, advantages and disadvantages followed by the comparative study of these tools. The tools deliberated for the exploration are easy to expand, works on cross platform, modular and many are freely available. The study can increasescope of theselection of these tools easier for the researchers and developers.

Keywords: Data Mining, WEKA, Rapidminer, Orange, KNIME, Apache Mahout, Oracle Data Mining, Rattle.

I. INTRODUCTION

Accumulation of huge amount of unprocessed data in electronic format resulted because of the swift transformations in technology. Finding patterns, trends and anomalies in these data and summarizing them with simple quantitative models is one of the biggest challenges of the information age, turning data into information and further information into knowledge. Existing conventional data analysis tools and techniques will not solve this problem because of huge size of raw data available. The process of exploring data to discover hidden associations and predict useful trends has anelongatedpast. Data Mining was evolved from evolutionary steps, starting from data collection, data access, data warehousing, decision support and finally to Data Mining. The term Data Mining also known as "Knowledge Discovery in Databases" wasn't coined till 1990s. But its base origin comprises of three interwoven scientific disciplines namely Statistics, Artificial Intelligence and Machine Learning.

Data Mining is a technology that extracts hidden and useful information by processing massive amount of data and blends traditional data analysis methods with sophisticated algorithms. Data Mining is widely used in various applications such as credit scoring, targeted marketing, social network analysis, finance, fraud detection, aerospace, education, telecommunication, intrusion detection etc., to name a few. It is the process of exploration and analysis, by automatic or semiautomatic means, of large quantity of data [1].

Data Mining is also defined as course of action for the data set analysis in various ways to find hidden relationships which are unknown and potentially valuable information from data. This is required in order to predict future trends and behaviors, to make proactive decisions and to answer business questions that consume too much time to answer [2]. Different data mining techniques have been studied to process and analyze several types of data patterns, where the most popular data mining tasks are Classification, Summarization, Association Rule Mining and Clustering [3].

The rest of the paper is organized as follows. Data Mining and Knowledge Discovery are elucidated in section II. Data Mining Tools are enlightened in section III. Comparisons of Data Mining Tools are explained in section IV. Concluding interpretations are given in section V.

II. DATA MINING AND KNOWLEDGE DISCOVERY

The procedure of Data Mining is applied on huge amount of data, which has been used for decision making on different aspects by converting into information. The decisionmaking process is extremely important in errands by obtaining knowledge from information. The fast rising area of research and application is Data Mining and Knowledge Discovery in Databases(KDD) that works on different techniques and algorithmsbased onClassical Statistics, Artificial Intelligence and Machine Learning. The iterative process of KDD is followed by various phases including:

- Data Preprocessing – It includes the cleansing of noisy and irrelevant data,integrating data from heterogeneous sources for the retrieval of useful patterns.
- Data Transformation – Also known as Data Consolidation, relevant selected data are transformed into data appropriate for data mining process in this stage.
- Data Mining − In this step, data are mined using one or more intelligent and smart techniques for the extraction of useful data patterns.

- Pattern Evaluation − Useful patterns are identified based on some important parameters in this stage.
- Knowledge Presentation − It is the last stage in KDD in which the discovered knowledge are represented visually to the user. This phase uses visualization techniques to help the users simply understand and interpret the data mining outcomes [4].

The iterative process has been used in KDD to enhance and refine the results of the given information. The refined information provides knowledge for discovering the interesting patterns and helps in decision making for a particular area.

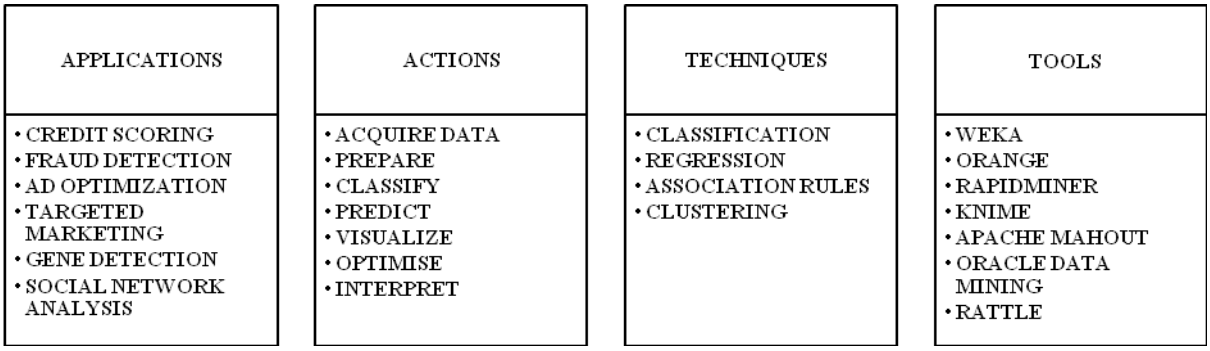

Figure 1. Overview of Data Mining

The data mining concept is totally based on specific actions which are used to process the data into information in various applications on the basic of different algorithms and techniques by using different mining tools and is depicted in Figure 1.

III. DATA MINING TOOLS

3.1. Weka

The most popular tool used for Data Mining is WEKA: Waikato Environment for KnowledgeAnalysis developed at University of Waikato, New Zealand in the year 1993.This tool is a group of machine learning algorithms that is used to solve all important tasks of data mining such as preprocessing, classification, clustering, visualization, association, feature selection etc. [5].

Features

- Free open source tool.
- Written in Java language.
- Platform independent.
- Licensed under GNU General Public License.

Advantages

- Data files can be loaded in various formats like .ARFF, .CSV, .C4.5 and binary.
- Provides interactive GUI namely Explorer,Experimenter, Knowledge Flow, Workbench and simple CLI.
- Used to develop new machine learning methods[6].
- Complete choice for data preparation, feature selection and data mining algorithms are incorporated.
- Easier to use and is portable.

Disadvantages

- Do not implement the newest techniques.
- The documentation for the GUI is quite limited.
- Inferior connectivity with non-java based databases.
- Does not have strong control with traditional statistics.

3.2 Rapidminer

Rapidminer,a data mining tool, earlier popularly known by the name YALE YetAnother Learning Environment. It was developed by Ingo Mierswa and Ralf Klinkenberg in the year 2006. Tool provides an incorporated environment for data preparation, machine learning,deep learning, text miningand predictive analytics. This tool is popular 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.

Features

- Free open source tool.
- Written in Java Language.
- Platform independent tool.
- Licensed under Affero General Purpose License.

Advantages

- Powerful visual programming environment.
- Access, load and analyze any type of data.
- Extract statistics and key information.
- Cleansesthe data expertly for predictive analytics.
- Build and deliver models faster and efficiently.
- Estimates the model performance accurately [7].
- Score models for the Rapidminer platform or for other applications.

Disadvantages

- Expensive license needs to be purchased for commercial use.
- Unintuitive tool.
- Its use cases are limited to the set of processors/modules it contains.

3.3. Orange

Orange is a component-based, open source tool developed at University of Ljubljanain theyear 1996. It is a visual programming software tool used for exploratory analysis of data, providesinteractivevisualization and binds with Python library for scripting. It has set of components which are used for data preprocessing, exploratory data analysis and model evaluation.

Features

- Free open source tool.
- Written in Python, Cython, C++ and C.
- Platform independent tool.
- Licensed under GNU General Public License.
- Advantages
- Tool provides machine learning and data visualization for beginners and experts. Interactive data analysis workflows are available with a large toolbox.
- Performssimple data analysis with data visualization.
- Interactive data exploration for rapid qualitative analysis with clean visualizations.
- Graphical user interface allow users to focus on exploratory data analysis.
- Numbers of add-ons are available within Orange tool for data and text mining from external data sources and has the capability toperform natural language processing.
- Data analysis feature are available in this tool.
- Simple and easier to learn.

Disadvantages

- Tool does not offer widgets for classical statistics.
- Machine learning is not handled homogeneously between the diverse libraries.
- Orange does not provide optimum performance for association rules.
- May not be compatible with Java platforms.

3.4 Knime

Konstanz Information Miner is a free open-source data mining tool used for data analytics, reporting and integration, developed by KNIME AG in the year 2006. This tool providesmulti-language software development environment which consists ofIntegrated Development Environment (IDE) and an extensible plug-in system. Various components for machine learning and data mining are integrated in KNIME through its modular data pipelining concept.

A graphical user interface and use of JDBC allows assembly of nodes that combines different data sources and preprocessing, modeling, data analysis and visualization are done without or with very little programming.

Features
- Free open source tool.
- Written in Java language.
- Compatible with Windows, Linux and OS x.
- Packageconsisting of Eclipse software licensed under the Eclipse Public License (EPL) and separate KNIME plug-ins licensed under the General Public License.

Advantages
- Easier to use and user friendly tool.
- Compatiblewithall platforms.
- Specialized for enterprise reporting, business analytics and data mining.
- Best for molecular analysis.
- Well known analysis modules are available offering vast library of statistical routines.
- WEKA machine-learning algorithms can be integrated in KNIME.

Disadvantages
- Machine learning parameter optimization does not haveautomatic feature.
- Error measurement methods are inadequate.
- Descriptor selection does not have wrapper methods.

3.5 Apache Mahout

Apache Mahout is a project undertaken by Apache Software Foundation in the year 2009. The goal of the project is to develop free scalable or distributed machine learning algorithms on Hadoop platform. The project mainly focused on collaborative filtering, clustering and classification. It uses Java libraries for linear algebra and statistics math operations and primitive Java collections. Recommendation mining, clustering mining, frequent itemset mining and classification mining are supported by this tool. Map/Reduce paradigm is used on the top of Apache Hadoop to implement the algorithms [8].

Features
- Free open source tool.
- Written in JAVA and Scala.
- Cross Platform.
- Licensed under Apache release 2.0.

Advantages
- Useful for distributed environment.
- Has ready to use framework.
- Applications can analyse data faster and effectively.
- Greater community support.

Disadvantages
- Limited availability of algorithms.
- Doesnot confine commitment to Hadoop based executions.
- Poorvisualization and less supportof scientific libraries[9].

3.6 Oracle Data Mining

Oracle Data Mining popularly known as ODM, is a module of Oracle Database Enterprise Edition developed by OracleCorporation in the year 2002. It is an outstandingtool as itcontains several data mining and dataanalysis algorithmsfor classification, prediction,regression, association, feature selection, anomaly detection, feature extractionand specialized analytics. Data mining models for the design, management and operational functions are available in the database environment. Application and tool developers can embed predictive and descriptive mining capabilities using PL/SQL or Java APIs[6].

Features
- Not a free source tool.

- Distributed under Oracle.
- Proprietary license under Apache release 2.0.

Advantages

- Simple model as it uses Oracle SQL on the database.
- High performance and reliability.
- Eliminates data extraction and movement.
- Provides a platform for analytics-driven database applications [10].
- Provides increased security by leveraging database security options.
- Delivers lower total cost of ownership as compared to traditional data mining vendors.

Disadvantages

- Limited language support.
- Uses PL/SQL or Java API only.

3.7 Rattle

R Analytical Tool to Learn Easily is a popular open source GUI-based data mining tool using Gnome graphical interface from Togaware. Rattle supports unsupervised and supervised data mining models.Data can be summarised visually.

Features

- Free open source tool.
- Written in R Language.
- Runs on GNU/Linux, Macintosh OS/X and MS/Windows.
- Licensed under GNU General Public License.

Advantages

- Provides considerable data mining functionality.
- Allows thedataset to be partitioned into training, validation and testing.
- Option for scoring an external data file.
- Provides more sophisticated processing in R language.
- Provides an easy environment to learn R language.

Disadvantages

- GUI of Rattle is not analytical.
- Restrictedcapability to producevarious graph types.
- BIGLM packages and parallel programming are not supported.

IV. COMPARISON OF DATA MINING TOOLS

Seven popular data mining tools namely WEKA, Rapidminer, Orange, KNIME, Apache Mahout, Oracle Data Mining and Rattle along with their important features, advantages and disadvantages are considered for this study. Comparisons of these tools are prepared in the tabulated format.

Table 1: Comparison of Data Mining Tools

Tools→ Features ↓	WEKA [11]	RAPID MINER [7]	ORANGE[12]	KNIME[13]	APACHE MAHOUT [8,9]	ORACLE DATA MINING[14]	RATTLE[15]
Release Year	1993	2006	1996	2004	2009	2002	-
Latest Version,Released Year	3.8.3, September 4, 2018	9.2, January 20, 2019	3.17.0, October 26, 2018	4.0, June 27,2019	0.14.0, March 6, 2019	11gR2, September, 2009	5.1.0, September 5 2017
License	GNU General Public License	APGL Affero General Purpose License	GPLv3	GNU General Public License	Apache License 2.0	Proprietary	GNU General Public License

Operating System	Windows, OS X, Linux	Cross Platform	Cross Platform	Windows, OS X, Linux	Cross Platform	LINUX, Oracle Solaris	GNU/Linux, Macintosh OS/X, MS/Windows
Language	Java	Language independent	Python, Cython, C++, C	Java	Java, Scala	PL/SQL, Java APIs SQLlanguage	R
Developer/ Author	University of Waikato, New Zealand	Ingo Mierswa,Ralf Klinkenberg	University of Ljubljana	KNIME AG	Apache Software Foundation	Oracle Corporation	Graham Williams
Website	www.cs.waikato.ac.nz/~ml/weka	www.rapidminer.com	www.orange.biolab.si	www.knime.com	www.mahout.apache.org	www.oracle.com	www.rattle.togaware.com
Type	Machine Learning	Data Science, Machine Learning, Predictive Analytics	Machine Learning, Data Mining, Data Visualization, Data Analysis	Mining/ Deep Learning / Data Analysis / Text Mining	Machine Learning	Data Mining and Data Analysis	Data Mining / Statistical Analysis
Availability	Open source	Open source	Open source	Open Source	Open source	License	Open source

Table 1 compares the various characteristics of seven data mining tools. Nine characteristics are included such as operating system on which it works, released year, latest released version and so on.

V. CONCLUSION

The paper gives a concisepreamble to the concept of data mining and the steps used to improve the results that provide the knowledge for discovering the remarkable patterns. Various data mining tools are also discussed in this paper. The details of seven tools are discussed along with their important features and the applications where these tools can be used. Comparativestudy is also done on these tools. In future, the above work will be extended on new data mining tools.

VI. REFERENCES

[1] Xingquan Zhu and Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978-1-59904-252, Hershey, New York, 2007.

[2] M.Dunham, "Data Mining Introductory and Advanced Topics". ISBN 978-0-13-088892-1, Prentice Hall,2003.

[3] Han Jiawei, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques" Third Edition,ISBN 978-0-12-381479-1, Morgan KaufmannPublishers, 2011.

[4] J.Grabmeier and A. Rudolph, "Technique of Clustering Algorithms in Data Mining", Data Mining and Knowledge Discovery,2002.

[5] A Sharma and B. Kaur, "A Research Review on Comparative Analysis of Data Mining Tools , Techniques And Parameters",International Journal of Advanced Research in Computer Science, Vol. 8, No. 7, pp. 523–529, 2017.

[6] I.H. Witten and E.Frank, "Data Mining: Practical machine Learning tools and techniques", Secondedition,Morgan Kaufmann, San Francisco, 2005.

[7] Rapidminer[online] Available at https://rapidminer.com/products/studio/feature-list.

[8] Apache mahout [online] available at http://www.hortonworks.com/hadoop/mahout.

[9] Apache mahout [online] available at www.intellectx.com/techstack/mahout.

[10] Oracle Data Mining [online] at https://www.oracle.com/technetwork/database/options /advanced-analytics /odm/twp-data-mining-11gr2-160025.pdf.

[11] WEKA [online] available at https://www.cs.waikato.ac.nz/ml/weka.

[12] Orange [online] available at https://orange.biolab.si.

[13] Knime [online] available at https://www.knime.com.

[14] Oracle data mining [online]available at https://www.oracle.com/technetwork/database/options/advanced-analytics/odm/overview/index.html.

[15] Rattle [online] available at https://rattle.togaware.com.