International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Issue 11, November 2015
www.ijates.com
ijates
ISSN 2348 - 7550

PARAMETRIC COMPARISON OF DATA MINING TOOLS

Neha Chauhan[1], Nisha Gautam[2]

[1]Student of Master of Technology, [2]Assistant Professor,

Department of Computer Science and Engineering, AP Goyal Shimla University, (India)

ABSTRACT

Presently, to assemble the large volume of dataset at lesser cost, storage technology and data collection has made it possible for any organisation.In order to obtain the useful and convenient information, it is necessary to utilize the stored data for any further use. This overall leads to data mining concept.Today, Data mining is a new and important area in human life. Data mining plays an important role in various fields like business, education, finance, healthsector etc.The main motive of Data mining is to examine the data from different perspective then label it and encapsulate it in order to acquire useful information by using their various new techniques and tools.Today, the various data mining tools available that researchers needs for evaluating their data. In this paper, we overviewed different tools includes in data mining like, WEKA, Rapid Miner, and KNIME. This paper presents pros and cons of each tools and also compare their parameters.By this comparitive study, it can be made easy for the researchers to make a best selection of the tool.

Keywords:Data Mining, Data Mining Tools, KNIME, RapidMiner, WEKA.

I. INTRODUCTION

Data is a collection of facts or interpretation. Computers take data as an input and in various forms of data which can be processed by computer and these facts are stored in databases [1]. So, Database is a collection of data and are used for storing large data items[2].Data warehouse is a process of storing the huge amount of data in a special repository by integrating available and historical data from different sources[1]. For abstracting the required information from this warehouse, we need a technique. So, Data mining is the process of extracting or mining the knowledgeable information from the large volume of data repositories. The core step of data mining is to mine or discover the novel information in terms of patterns or rules from the large volume of data. The key idea behind the Data mining is to design and work efficiently with the large dataset [3]. Datasets may be acquired from variety of resources including: traditional databases, data warehouses, web documents, multimedia databases or simple local textual files [4]. The most specific characteristic of data mining is that it concerns with a massive and complex datasets in which its volume varies from gigabytes to terabytes. To acquire sequence and trends in data, mainly the Data mining uses multiplex algorithm and mathematical analysis [5]. The industries used various algorithms and techniques of Data mining for better decision making. Data mining help the analyst to identify the informative or meaningful facts, sequence, trends, anomalies etc. Data mining has now to begin to be an essential factor in various fields such as business, education, healthcare, finance, scientific etc [6].

Data mining also referred to as Knowledge Discovery in Database, but both Data mining and KDD are distinct with each other. The term KDD refers to discover the useful knowledge from large amount of databases. On the other hand, Data mining refers to the application of the algorithm for mining patterns from data without involving the steps of KDD process [7].

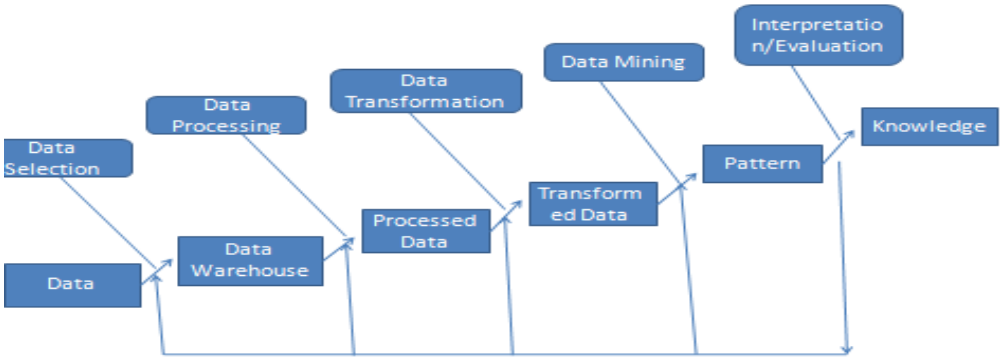

Fig.1 Data Mining and KDD Process

In this fig, KDD process consists of various steps:

1). *Data Selection*: Select the appropriate data from the database on which discovery is to be performed to get the useful information.

2). *Data Pre-processing*: After selecting the data, noisy and irrelevant data is to be removed.

3*). Data Transformation*: Data are transformed into an appropriate form to reduce the effective number of elements and make it ready to perform Data mining.

4). *Data mining*: Data mining is a process where an appropriate task or algorithm has chosen in order to extract data patterns.

5). *Data Interpretation/Evaluation:* Interpreting the mined patterns by removing the redundancy and translate it into discovered knowledge that human understands.

1.1 Data Mining Task

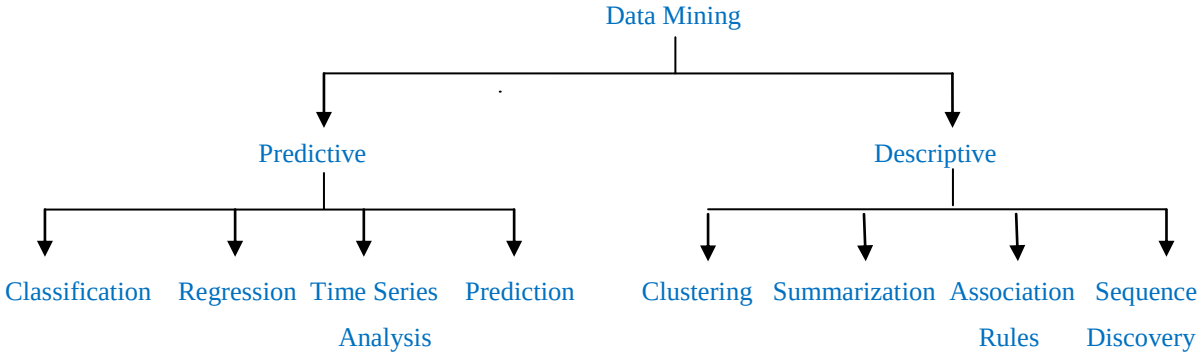

Fig .2 Data mining techniques

Data mining tasks can be classified into two categories: Predictive and Descriptive mining [8].

1.1.1 . Predictive: Predictive Data mining tasks involves prediction of unknown and future values by performing conclussion on the existing data. Predictive data mining tasks include Classification, Regression, Time Series analysis, Prediction.

- Classification: Classification is Data analysis task used to finding a set of models where a model is constructed to predict categorical class labels.
- Regression: Regression is a Data mining technique that are generally used to predict a numeric values i.e., age, income, profit, temp, etc in which target value are known.
- Time Series Analysis: Time Series Analysis is a statistical technique for analyzing time series data in order to extract all meaningful knowledge, statistics and sequence from the shape of the data.
- Prediction: Prediction is a supervised leaning task that is similar to classification used to predict unknown or hidden values.

1.1.2 . Descriptive: Descriptive Data mining tasks involves the specific properties of the data in database and locating the sequence and behavior in data. Descriptive Data mining tasks include Clustering, Summarization, Association Rules, Sequence Discovery.

- *Clustering*: Clustering is the method of making a group of clusters which have similar in their characteristics. The objects which have same characteristics that belongs to one cluster and which are different from other clusters.
- *Summarization*: Summarization is a Data mining concept in which a larger set of data is summarized and gives a smaller set in a meaningful manner that describe the general overview of data.
- *Association Rules*: Association Rules helps to find out the relationship between the data that is unrelated in an information repository or a relationship database.
- *Sequence Discovery*: Sequence Discovery is a concept of Data mining that is used to find out or discover sequences of events that mostly occur together.

II. DATA MINING TOOLS

Today's various data mining tools that are available to handle or manage the large number of datasets and also to improve the quality of data, such tools are RapidMiner, Weka, R, scikit-learn, KNIME, Orange, KEEL, Tanagra etc.These data mining tools makes easy for analyst to get the knowledgeable information.Data mining tools are used to predict future trends, behaviours, allowing business to make proactive , knowledge driven decissions[9]. The various Data mining techniques and algorithms have been implemented on these tools to extract the information and also to check their efficiency and accuracy. In this paper, we are going to discuss and compare only three tools among of these that are; RapidMiner, WEKA, and KNIME which are using the same platform(Java).The description of these tools are as follows:

2.1 Weka

The WEKA(Waikato Environment for Knowledge Analysis) is a Data Mining tool , developed at the University of Waikato , New Zealand that is suitable for machine learning algorithm for data mining tasks and well suited for developing new machine leaning schemes [10]. These algorithms can be applied directly to dataset or can be called from your own java code [9].WEKA is an open source software issued under the GNU(General Public

License) agreement. WEKA provide four application interface : Explorer, Experimenter, Knowledge flow, and Simple Command line [11]. But Explorer is the main interface of WEKA. WEKA is Java based software and can run in different platforms. With the Java based version, the tool is so revolutionary and used in various application including visualization and algorithm for data analysis and predictive modeling [10].It is freely available for download and offers many powerful features.

Features

- WEKA is Java based open source data mining tool.
- It is easy to use for beginners and has the ability of running several learning algorithms and comparing.
- It is platform independent.
- It performs various data mining tasks including: Data preprocessing, Classification rules, regression, Clustering, association rules, visualization, feature selection and improving the knowledge discovery.
- WEKA has 49 Data preprocessing tools, 76 Classification/regression algorithm, 8 Clustering algorithm, 3 algorithm for finding association rules, 15 attribute/subset evaluator plus 10 search algorithm for feature selection [12].
- There are various built in features.
- There is no programming and coding language required.

Advantages

- Easy to manipulate the data.
- Provide access to SQL databases.
- It provides two options for the user to interact through Explorer and Command line [13].
- Specially used for data mining.
- It provides various machine learning algorithms for data mining tasks.
- It supports various standard Data mining tasks that include: Data preprocessing, Clustering and Classification, Regression, Visualization and Feature selection [14].

Disadvantages

- Memory is limited and has lesser performance [5].
- Data visualization and data survey is limited.
- Not better suitable option for the large datasets as they roughly handled.
- Lacking in the representation to the result of processing.
- Limited ability to partition dataset to training and test set [15].
- It doesn't accept data in every format (data format constraints).
- Not good in interfacing with other software [5].

2.2 Rapid Miner

RapidMiner, previously YALE (Yet Another Learning Environment) was developed at the Technical University of Dortmund in 2001 by Ralf Klinkenberg, Ingo Mierswa and Simon Fischer.After, this software name was changed in 2007 from YALE to RapidMiner and is developed by the company RapidMiner, Germany. RapidMiner is an open source java based system for data mining and provides an integrated environment for machine learning, data mining, text mining ,predictive analysis and business analytics and is

mainly used for business and industrial application[9].RapidMiner is the most powerful, easy to use and intutive Graphical User Interface for the design of analytic process, that contain several "operators".The operator functions as a single task in their process in which the input is produced by the existing output of the operator[5].

Features

- It is platform independent.
- It has compatibility with various databases like oracle, MySQL, Excel, SPSS, Microsoft SQL server etc.
- It provides Drag and Drop interface to design the analytics process.
- It supports and accepts new data drivers.
- It provides more than 500 operators for all machine learning procedures, and also combines learning schemes and attributes evaluators of the WEKA learning environment [16].
- It allow user to work with different sizes and types of data sources.

Advantages

- It has enormous flexibility.
- It provides the integration of maximum algorithm of such tools.
- Easy to debug the errors.

Disadvantages

- Limited partitioning abilities for dataset to training and testing sets.

2.3 KNIME

Konstanz Information Miner is an open source general data mining tool that is based on the Eclipse platform, developed and supported by KNIME.com.AG. In 2004, the KNIME initially developed by the team of software engineer at the University of Konstanz, Germany and in 2006, the initial version of KNIME was released [4]. KNIME is very powerful tool for analytical task, extracting data and knowledge from the web communities.The KNIME base version already incorporates hundereds of 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[17].In KNIME, representation of data sources and sinks, mining algorithm,tranformations, visualizations,etc defined by set of nodes called "workflow" and each node has its specific input and output ports that depends on the functionality of the node [18]. For both simple and complex data types, KNIME allows revolutionary analysis to discover trends and predict future results. KNIME uses for teaching as well as research which allows to integrate the new algorithm and tools in a simpler manner.

Features

- Available to everyone i.e., allow users to use the well- defined node API to add proprietary extensions.
- Intuitive user interface.
- It is easy to use and handle different functions.
- KNIME modules cover a wide variety of functionalities like, I/O, data manipulation, views, hilting etc to better understand your data.
- It provides the users to create data flows or pipeline visually, users can selectively execute some or all analysis steps, study the results, prototypes, and collaborative interpretations [13].

- For cross validation and independent validation, it provides functionality to save parameters.

Advantages

- The major benefit of this is easy to use plug-in [18].
- It based on the node work which includes more than 100 nodes to examine the data [19].
- It provides data flow process by dragging and dropping new nodes.

Disadvantages

- Less suitable option for large complex workflows.
- Partitioning ability is limited for dataset [15].

III. COMPARISON OF DATA MINING TOOLS

This comparative study of Data Mining Tools based on their parameters is listed below. The main motive of this comparison is to make the best selection of tool with respect to their areas.

Table1. General Parameters of these tools

Parameters	**WEKA**	**RapidMiner**	**KNIME**
Developer	University of Waikato, New Zealand.	RapidMiner, Germany.	Swiss company Knime.com AG, Switzerland.
Programming language	Java	Java	Java
Released date	1993	2006	2004
License	GNU General Public License	AGPL Proprietary	GNU General Public License
Availability	Open Source	Open Source	Open Source
Current Version	3.7.13	6.5	2.10
Areas	Machine learning, Dta visualization,time series and analysis,text mining,fraud detection etc	Financial forecasting,Targeted marketing,Medical diagnosis,credit card fraud detection,text mining,weather forecasting etc	Customer intelligence,Finance,Manufacturing, Chemical liberary enumeration, Retail, Cross industry etc.
Portability	Cross Platform	Cross Platform	Linux,Windows, OS X
Usability	Easy	Easy	Easy
Compatability with database	MySQL, Postgre SQL, MSQL Server, Oracle,	Oracle, IBM DB2, Microsoft SQL Server,	MySQL, SQLite, Postgre SQL, Hive connector etc.

	ODBC, Sqlite 3.x, HSQLDB etc.	Mysql, Excel, Access, SPSS etc.	
Platform Supporting	Platform independent	Platform independent	Platform independent
Flexibility	Easy to use but not enough flexible	Flexible	Same as WEKA
Visualization	Limited visualization	Better visualization	Better visualization
GUI	Has good but not better as much as other mentioned tools	It has better GUI	It has better GUI

IV. CONCLUSION

This paper has given the brief introduction of Data mining and three different Data mining tools along with its features- WEKA, Rapid Miner, and KNIME. This paper contains the chart comparison between these three tools by using their parameters. Each tool has its own advantages and disadvantages. By employing this parametric study, this concluded that Rapid Miner has much better than other mentioned tools. This comparative study will make things easier to the learner in the selection of data mining tools according to their areas. In future, we will find out the solution to overcome from the limitations of such tools to make them best in every aspect.

REFERENCES

[1]. A. Radhi , A. Essa, Bach,Christian, Data Mining and Warehousing, ASEE 2014 Zone I Conference, April 3-5, 2014.

[2]. S. Joyce M , Nirmalrani V, Privacy in Horizontally Distributed Databases on Association rules, International Conference on Circuit, Power and Computing Technologies [ICCPCT].

[3]. A.V. Saurkar, V. Bhujade, P. Bhagat, A. Khaparde , A Review Paper on various Data Mining Techniques, International Journal of advanced Research in Computer Science and Software Engineering.

[4]. A. Jovic, K. Brikic, N. Bogunovic , An Overview of free software tools for general Data mining.

[5]. S. Sarumathi, N. Shanthi, S. Vidhya, M. Sharmila , A Review: Comparative Study of Diverse Collection of Data Mining Tools, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 8(6), 2014.

[6]. Snehal A. Deshmukh , Data Mining Tools: Review, INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY, 3(9).

[7]. T. Silwattananusarn, K. Tuamsuk, Data mining and its application for Knowledge management: A literature review from 2007 to 2012, International Journal of Data Mining & Knowledge Management Process (IJDKP), 2(5), September 2012.

[8]. N. Jain, V. Srivastava, Data mining techniques: A Survey Paper, International Journal of Research in Engineering and Technology, 2(11),Nov-2013.

[9]. K. Rangra, K.L. Bansal , Comparative Study of Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering, 4(6), June 2014.

[10]. P.S. Patel, S.G. Desai , Comparative Study on data Mining tools, International Journal of Advanced Trends in Computer Science and Engineering, 4(2), April 2015.

[11]. S. Srivastava, WEKA: A Tool for Data Preprocessing, Classification, Ensemble, Clustering and Association Rule mining, International Journal of Computer Applications, 88(10), February 2014.

[12]. S.K. David, Amr T.M. Saeb, K.A. Rubeaan, Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent System, 4(13), 2013.

[13]. K. Saravanapriya, A Study on Free Open Source Data Mining Tools, International Journal of Engineering and Computer Science, 3(12), December 2014.

[14]. S. Singhal, M. Jena, A study on WEKA tool for Data Preprocessing, Classification and Clustering, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2(6), May 2013.

[15]. H. Solanki, Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory, International Journal of Computer Applications, 75(16), August 2013.

[16]. M. Vijayakamal, M. Narendhar, A Novel Approach for WEKA & Study On Data Mining Tools, International Journal of Engineering and Innovative Technology (IJEIT), 2(2), August 2012.

[17]. L. Kataria, Implementation of Knime-Data Mining Tool, International Journal of Advanced Research in Computer Science and Software Engineering, 3(11), November 2013.

[18]. S. Gunnemann, H. Kremer, R. Musiol, R.Haag, T. Seidl, A Subspace Clustering Extension For the KNIME Data Mining Framework, 2012 IEEE 12th International Conference on Data Mining Workshops.

[19]. P. Subathra, R. Deepika, K. Yamini, P. Arunprasad, S.k Vasudevan, A Study of Open Source Data Mining Tools and its Applications, 10(10), 2015.