OPEN SOURCE DATAMINING TOOLS AND THEIR USAGE

K.S. Sindhu[1], Megha J Prakash[2]

ABSTRACT

Data mining is a method to study a large database in order to extract new information. The number of data mining tools helps us to find or extract information in the relevant format. The main objective of data dredging is prediction, which helps to predict the information for future usage. This paper explains the open source software tools which are used to serve various data mining methods.

Keyword : data mining, rapid miner, Weka, Orange, KNIME, Tanagra and XL miner

I. INTRODUCTION

Data mining [1] is the process of extracting constructive information from a very huge catalogue. Data mining is otherwise called as knowledge discovery process, knowledge dredging, and data archeology. The following types of data mining are available,

"Spatial data mining

"Temporal data mining

"Web mining

"Social mining

"Multimedia mining etc.

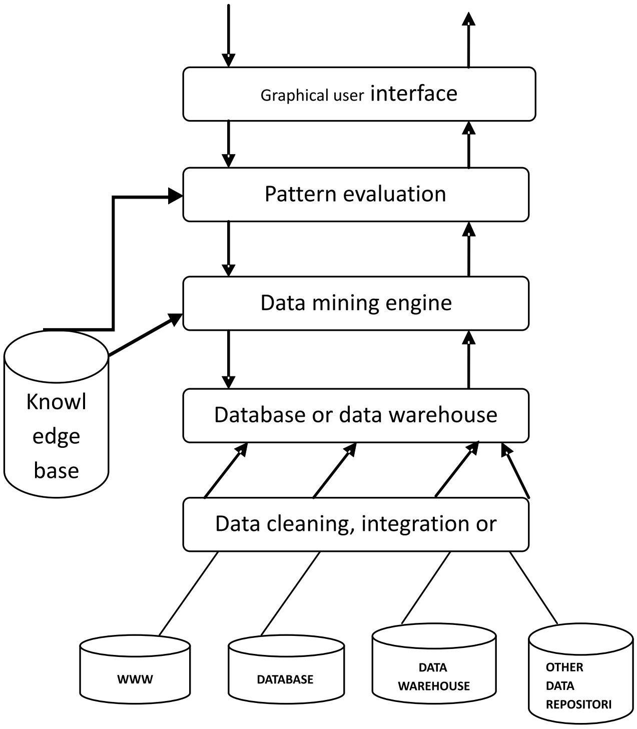

Fig 1.1 data mining architecture

II. DATAWAREHOUSE :

Data warehouse [5] acts as a source for data mining where, all the data are stored in a repository.

The main goal of data mining is classification prediction, clustering, and association. There are several tools and techniques used for mining such data

III. Data mining tools :

1. Rapid miner

2. Weka

3. Orange

[1] Assistant Professor, Department of Computer Science and Engineering, Karpagam Academy of Higher Education Coimbatore

[2] Assistant Professor, Department of Computer Science and Engineering, Karpagam Academy of Higher Education, Coimbatore

4. KNIME

5. Tanagra

6. XL miner

A. RAPID MINER :

Rapid miner[6][10] tool is a platform to perform knowledge dredging activities; it is used for research purposes as well as mining real world task. It is a free open source software coded in java. Rapid miner promotes a new concept called transparent data handling. Rapid miner supports many databases where data can be obtained from multiple sources scrutinized inside the application. This rapid miner tool starts with data warehouse extracts data from data warehouse and transforms the data into a needed format and the loads the data, where each transformation and each revelation is processed as a single process with the same tool .The main advantage of this tool is that it solves the complex analyses very easily.

Guidelines to use rapid miner :

Step 1: allocate the dataset.

Step2: choose the function.

Step 3: implement the function using RM tool.

The relevant dataset can be obtained from UCI repository

To install rapid miner and to work with it, first we have to install rapid miner. Then the new icon is clicked for creating an RM project. Then the dataset which is normally in the ARFF file format or .xsl format is selected.

B. WEKA :

Weka[3] stands for Waikato environment for knowledge analysis. This a machine learning tool coded in java language and an open source software that holds a set of revelation tools and step by step procedure to examine the information, along with GUI for simple access to functions. Weka supports multiple data mining activities such as data preprocessing-clustering, classification, regression, visualization and feature selection .In this tool algorithms can be incorporated in the dataset either directly or from java code.

Installation of weka:

Weka can be downloaded from the official website http://www.cs.waikato.ac.nz/ml/weka.After downloading run the exe file and then select the default setup. Data formats accepted by weka are CSV, ARFF, and database using ODBC, through the file format is in ARFF by default.

Weka explorer consists of 6 tabs. They are preprocess, classify, cluster, association, select attributes and visualize. The first and foremost step in weka tool is data preprocessing. Data preprocessing involves cleaning the dataset such as removing the noisy and inconsistent data and handling missing values. After loading the dataset in the explorer preprocess tab must be clicked and once the data have been preprocessed the dataset can be used for further process such as classification, clustering etc.

Classify: this involves step by step procedures on the processed files.

Cluster : this is done to identify similar or identical data items.

Association : this is done to identify the data items which are associated with one other.

Select attributes : on including and excluding the attributes changes can be identified.

Visualize : output is seen through graph.

C. Orange tool :

Orange [4][8] is a data mining tool as well as a machine learning software; this tool is developed in python language an open source software. Rapid miner uses XML to depict the operative trees modeling information detection procedure. It has supple operator for information input and output file formats. It is an incorporated platform for machine learning, data dredging, text mining, analytical analysis and big business analytics. It is a data visualization tool as well as a collection of data mining tools. The main advantage of this tool is visual programming interface i.e. It is a drag and drop tool for developing the model as well as examining the data and data can be painted. Dummy data (duplicate data) can be created as per our requirement. It has an open source data dredging package developed on Python, NumPy, wrapped C, C++ and Qt. As it is built in python, it will be very much comfortable for the programmers to gain knowledge. Scripting data dredging classification tribulations is simpler in this tool.

This orange tool has four tabs :

1. Data

2. Visualize

3. Model

4. Evaluate

Data :

This consists of 26 functionalities : one can obtain data from multiple recourses such as files, SQL and tables. In this feature we can generate dummy data or merge data or select data. The most important thing is that we can identify cluster and preprocess data.

Visualize :

There are 15 types of visualizations available, which are used to visualize the data from multiple dimensions.

Model :

This orange tool holds 10 supervised modeling languages. They are constant CN2rule induction, K-nearest neighbor, tree, random forest, support vector machine, linear regression, logistic regression, and naviebayes.

Evaluate :

To evaluate the data test score node is clicked

D.Knime tool :

KNIME [2] tool stands for Konstanz information miner which is an open source software written in java language. The main advantage of Knime tool is graphical user interface structure. Information handling, data dredging, and transformations can be carried out efficiently by grouping all the work flow into a single work structure. It is based on the Eclipse environment and, in the course of its modular API, is effortlessly extensible.

SETTING UP KNIME :

Step1 : Install KNIME and setup in Personal Computer

Step2: Fix the display place and locate the functioning register for KNIME to store its records.

Step3: Then create a work flow by clicking file -> new and then name the workflow and click on finish button.

Step 4: To import the data just drag and drop the file reader node in the workplace and click on the file reader to open the data.

The main advantage of KNIME is that it groups all examining modules and eminent. Weka data dredging platforms and further add-ons permit R-scripts to run, providing right to use to a huge files of algebraic routine. The limitations of KNIMEis that it contains only some degree of error dimension method.

E.Tanagra tool

TANAGRA[9] is an open source data exploring software for educational and exploratory purposes. It proposes numerous data dredging methods from examining data investigation, algebraic learning, machine learning and databases area. Tanagra is a descendant of sipina, which implements a variety of supervised learning through step by step procedures. Every node is an algebraic or machine learning practice, and association between two nodes indicates the information exchange. But not like the greater part of the tools based on the workflow standard.

F. XL miner :

XLMiner[7] is a tool which offers a range of methods to examine records. It has widespread exposure of algebraic and machine-learning techniques for categorization, forecast, information examination, and reduction. The XLMiner is classified into five tabs: Data, Data Analysis, Time Series, Data Mining, and Applying Your Model.

Data - This tab includes the Get Data icon. Click Get Data to recover a trial data from a database. Choose File Folder to bring in a set of text documents situated within a file folder. Choose Big Data to use Apache Spark to fetch big data into Excel.

Data Analysis - This tab includes four icons: Explore Transform, Cluster, and Text. Click Explore to use the Chart Wizard, Feature Selection, or view existing charts. Tick Transform to transform information set with lost data, carry out binning, or to transform unconditional information. Tick Cluster to execute cluster analysis by means of k-Means or Hierarchical method. Tick Text to execute an examination on a group of text documents by means of the new Text Miner characteristic.

Time Series - This tab contains three icons: Partition, ARIMA, and Smoothing. This function is used when examining a time series.

DataMining - This tab consists of four Partition, Classify, Predict, and Associate. This function is used to perform data mining actions.

Applying Your Model - This tab contains two icons: Score and Help. Tick Score to score data in a worksheet by means of the classification or prediction algorithms. Tick Help to transform the product, ensure your authorized position, associated with XLMiner.

IV. COMPARISION OF DATA MINING TOOLS:

S.NO	PARAMETER	RAPIDMINER	WEKA	ORANGE	KNIME
1	Developer	RapidMiner, Germany	University of Waikato, New Zealand.	University of Ljubljana	Swiss company Knime.com AG, Switzerland.
2	Programming language	java	java	python	java
3	Authorization	AGPL Proprietary	GNU General Public License	GNU General Public License	GNU General Public License
4	Accessibility	Open source	Open source	Open source	Open source
5	Portability	Cross Platform	Cross Platform	Linux,Windows, OS X	Linux,Windows, OS X
6	Usability	Easy to use	Easy to use	Easy to use	Easy to use
7	Platform sustaining	Platform independen	Platform independent	Platform independent	Platform independent
8	Latest version	9.0	3.8.3	3.13.0	3.6.1

V. Conclusion :

This paper provides a brief introduction to the open source data mining tools, where an introductory knowledge is available to the learners to know about various tools. This paper has presented both the advantages and disadvantages of various tools as well as a comparison chart between various parameters.

VI. References :

1. Ramamohan, Y., Vasantharao, K., Chakravarti, C. K., & Ratnam, A. S. K. (2012). A study of data mining tools in knowledge discovery process. International Journal of Soft Computing and Engineering (IJSCE) ISSN, 2(3), 2231-2307.

2. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., & Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter, 11(1), 26-31.

3. Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. International Journal of Innovative technology and exploring engineering (IJItee), 2(6), 250-253.

4. Kukasvadiya, M.S., & Divecha, N.H. (2017). Analysis of Data Using Data Mining tool Orange. Int. J. Eng. Develop. Res, 5(2), 1836-1840.

5. Mining, W. I. D. (2006). Data Mining: Concepts and Techniques. Morgan Kaufinann.

6. Patel, P. S., & Desai, S. G. (2015). A comparative study on data mining tools. International Journal of Advanced Trends in Computer Science and Engineering, 4(2).

7. Khattak, A. M., Khan, A. M., Rasheed, T., Lee, S., & Lee, Y. K. (2009). Comparative analysis of xlminer and weka for association rule mining and clustering. In Database Theory and Application (pp. 82-89). Springer, Berlin, Heidelberg.

8. http://orange.biolab.si/

9. http://eric.univ-lyon2.fr/~ricco/tanagra/

10. https://rapidminer.com/