ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

A Novel Comparative Study on Data Mining Tools

A.Komathi[1,] T.Ramya [2,] M. Shanmugapriya[3], V. Sarmila[4]

Assistant Professor, Dept. of CS & IT, Nadar Saraswathi College of Arts and Science, Theni,
Tamil Nadu, India

Assistant Professor, Dept. of CS & IT, Nadar Saraswathi College of Arts and Science, Theni, Tamil Nadu, India

M.Sc(CS&IT) Student, Dept. of CS & IT, Nadar Saraswathi College of Arts and Science, Theni, Tamil Nadu, India

M.Sc(CS&IT) Student, Dept. of CS & IT, Nadar Saraswathi College of Arts and Science, Theni, Tamil Nadu, India

ABSTRACT: Data mining is the method of estimating data from different views and short it into useful information. Data Mining allows the students to discover various data mining algorithms by using diverse open source data mining tools. These data mining tools are very useful to calculate valuable information and gain knowledge based on the input data sets. This paper gives the complete and hypothetical analysis of a five open source data mining tool. In this paper, we discussed about various available data mining tools and compared their utilities.

KEYWORDS: Data, Data Mining, Data Mining Tools, WEKA, orange, Rapid Miner, R, KNIME

I.INTRODUCTION

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. Data handling and managing of large data sets, there arevarious data mining tools are available which improve data quality from past to present. Data mining tools are provided for pre-processing data, feeding it into a variety of learning schemes, and analyzing the resulting classifiers and theirPerformance, concepts.

II.CATEGORIES 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.

Relational Databases: Briefly, a relational database consists of a set of tables containing values

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 includes the combination of video, images, audio and text media. They can be stored on stretched object-relational or object-oriented databases, or simply on a file system.

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 database in data mining. A Data in the World Wide Web is organized in inter-connected data’s. These datas can be text, audio, video, raw data, and even

DOI: 10.15680/IJIRCCE.2016. 0411020

presentations. Theoretically, the WWW consists of three important components: I. The content of the Web, which includes documents available; II. The structure of the Web, which covers the hyperlinks and the relationships between documents; and III. The usage of the web, describing how and when the resources are accessed.

Time-Series Databases: This database systems have time related data such as market data or recorded activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a stimulating real time analysis.

III.TYPES OF DATA MINING TOOLS

There are mainly three different categories of data miningtools. Traditional data mining tools, Application basedtools/Commercial based software and web-based datamining tools. Description of each is as follows:

1. Traditional data mining tools

Some mining programs are work as traditional way tocollect and analyze data which used by various company fordecision making process of large data sets. Majority of theseare supported by windows and UNIX versions. However,sometimes handling with only one database type.

2. Application based tools

An applications which shows the business oriented interfacefor data performance. In this historical data are representedas a references and check the current trends in order to seethe changes in the business. So, application based tools areeasy to use and helps in administrative work and provideservices for company performance.

3. Web based data mining tools

This kind of tools is called text-mining tool because of itsability to mine various kind of text from any writtenresources. And also help for scanning and converting datain selected format which is compatible with any tools.

IV.ASUMMARY OF DATA MINING TOOLS

1. WEKA

Weka is one of the very popular open source data mining tools developed at the University of Waikato in New Zealand in 1992. Waikato Environment for Knowledge Analysis. Weka is a collection of machine learning algorithms for data mining tasks. 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.

Common Features:

- Weka is the open source tool based on Java which is a group of machine learning algorithms
- Robust in machine learning techniques.
- Weka is greatest suited for mining association rules

Advantages:

Weka loads data file in layouts of ARFF, CSV, and C4.5, binary. Though it is open source, Free, Extensible, Can be integrated into otherjava packages.

2. RAPID MINER

Rapid Miner is also one of the open source tool in data mining developed by Ingo Mierswa and Ralf Klinkenberg. Rapid Miner also known as YALE(Yet another Learning Tool) based on XML.

Common Features:

- Rapid Miner has a collection of functionality, is elegant and has good connectivity.
- Rapid Miner consist of many learning algorithms from WEKA.
- Compact and whole package.
- It simply reads and writes Excel files and diverse databases.

Advantages:

DOI: 10.15680/IJIRCCE.2016. 0411020

Rapid Miner has over 1,500 methods 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 optimum analysis processes.

Rapid Miner deals several procedures, mainly in the part of attribute selection and for outlier detection, which no other clarification offers.

3. ORANGE

Orange is an open source data mining tool and imagining software with energetic community and which helps learner and professionals for their investigation. This tool is companionable with windows, Mac OS c and GNU/Linux operating systems.

Common Features

- Orange tool contains a set of modules for data preprocessing, feature scoring and filtering, modeling, model evaluation, and assessment techniques.
- It's also very useful for analytical process which have userfriendly visual programming or python scripting.
- Specially, these toolshave utilities for Bioinformatics Add-On and Text MiningAdd-On.

Advantages

- It is an open source data mining package build on Python, NumPy, enclosed C, C++ and Qt.
- Orange is written in python therefore is easier for most programmers to learn.
- It has superior debugger.

4. R

R is also an open source statistical analysis tool based on C and FORTRAN programming language developed by Ross Ihaka and Robert Gentleman at the University Of Auckland, New Zealand

R(Revolution) is a free software programming
language and software environment for statistical computing and graphics. It is widely used among statisticians and data miners for developing statistical software and analysis.

The R's Strength is the ease with well-designed publication quality plots can be produced, including mathematical symbols and formulae

Common Features

- R is a well-supported, open source, command line driven, statistics package.
- It consists upto hundreds of extra "packages" freely available, which provide all sorts of data mining, machine learning and statistical techniques.
- R provides less support to data mining and Weka, it algorithms as compared to Rapid Miner does implement a few data mining algorithm

Advantages

- R is more transparent since the Orange are wrapped C++ Classes.
- Programming in R really is very different, we are working on a higher abstraction level, but we do lose control over the details.
- It has the ability to make a working machine learning program in just 40 lines of code.

5. KNIME

KNIME known as Konstanz Information Minor It is a user -friendly and inclusive open-source data integration, processing, analysis, and exploration platform. Since day one, KNIME has been developed using difficult

DOI: 10.15680/IJIRCCE.2016. 0411020

software engineering practices and is currently being used actively by over all over the world, in both industry and academia. KNIME is a modular data exploration platform that enables the user to visually create data flows, selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models

Common Features:

- Knime, pronounced "naim", is a nicely designed data mining tool that runs inside the IBM's Eclipse development environment.
- The Knime base version already incorporates over 150 processing nodes
- KNIME is easy to extend and to add plugins.Additional functionalities can be added

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.

V. RELATIVE LEARNING ON FIVE DATA MINING TOOLS

	WEKA	RAPID MINER	ORANGE	R	KNIME
Release Date	1993	2006	2009	1997	2004
Latest Version/Release Date	3.8.0/April 14,2016	7.2/Aug2,2016	3.3.8/Oct 11,2016	3.3.2/Oct 31,2016	3.2.1/Aug 19,2016
Operating System	Cross Platform	Cross Platform	Cross Platform	Cross Platform	Linux, OS X, Windows
Language	Java	Language Independent	Python C++,C	C, Fortran and R	Java
Memory Usage	Less Memory hence works faster	Requires more memory		More memory	
Speed	Works faster on any machine.	Requires more memory to operate		Works fast on any machine.	

VI. CONCLUSION

The main focus of this relative study is to offer a platform to learn several data mining tools without having to spend a lot of time in examining for resources and tutorials. Additional vital aspect is to improve the learning capability for the students to help them in their Data Mining course project and their own specific research. Open-source data mining collections of today have come a long way from where they were only a period ago. They offer nice graphical interfaces, focus on the usability and interactivity, support extensibility through augmentation of the source code. They provide flexibility either over visual programming within graphical user Interface.

REFERENCES

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

[2.] S. Hameetha Begum,**," Data Mining tools and trends-An overview"**,International Journal of Emerging Research in Management & Technology ,ISSN: 2278-9359.

[3].Kalpana Rangra ,Dr. K. L. Bansa, "**Comparative Study of Data Mining Tools"**,International Journal of Advanced Research in Computer Science and Software Engineering

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

[5]. J. Han and M. Kamber. **"Data Mining: Concepts and Techniques"** Morgan Kaufmann, 2000. Article: Exforsys Inc **" What is Data Mining"** Published on: 27th Jul 2006 Source:http://www.exforsys.com/tutorials/data mining/datamining-overview.html

[7]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

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

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

[10]. http://rapidminer.com/

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

[12]. http://www.cs.waikato.ac/nz/~ml/weka

DOI: 10.15680/IJIRCCE.2016. 0411020