RESEARCH ARTICLE OPEN ACCESS

Elaborative Analysis on Data Mining Techniques and Tools

Ms. Palvi
Lovely Professional University, Phagwara
India

ABSTRACT

Data mining is the process of finding out latent pattern from huge set of data so that the solely purpose is to receive hidden information. Data mining is a technique of exploring latent patterns from huge databases. This acts as a boon to knowledge discovery which provides a great support to business world and academia. To make this knowledge discovery occurring various data mining tools are developed. These tools provide way to get data and to extract some interesting patterns out of it which are further useful to attain new knowledge.. In this paper author elaborate introduction in section 1, data mining process in section 2, data mining techniques in section 3, different tools that provide aid to data mining in section 4, comparative analysis of different tools in section 5, section 6 describes relevant conclusion.

Keywords :— Data Mining techniques and tools, RapidMiner, Orange, WEKA, KNIME, Tanagra.

I. INTRODUCTION

Data mining is the process of extracting patterns from enormous amount of data by applying various techniques. This is used for knowledge discovery in databases to be used in decision making process. Big organizations use it primarily for finding new ways to increase their profits and to minimize cost. Data mining analyze the data and helps to bring up the hidden factors so that useful patterns and information can be generated. Various data mining tools are there to find accuracy and precision.

II. DATA MINING PROCESS

Data mining is the process of getting relevant data according to one's business values and interests from the large bunch of data using various techniques and algorithms such as Association, Clustering and Classification.

The steps involved in knowledge extraction are as follows:

A. Data Cleaning*:* The information obtained may contain some bugs which is preprocessed here to obtain relevant data.

B. Data Integration*:* Integration of the data is done here. Various steps are deployed to jon the data from different Data sets

C. Data Selection: The data which is required is selected for user application and further steps.

D. Removal of noisy data*:* Since they are large amount of data it captures more space, so using this method we mitigate the space but it achieves the same results.

E. Data Mining*:* A knowledge discovery process to extract the essential data.

F. Pattern Evaluation*:* It is the process in which a pattern is identified.Clustering can also be deployed here.

G. Knowledge Representation*:* This is the final stage in which the knowledge is represented using varied visualization techniques.

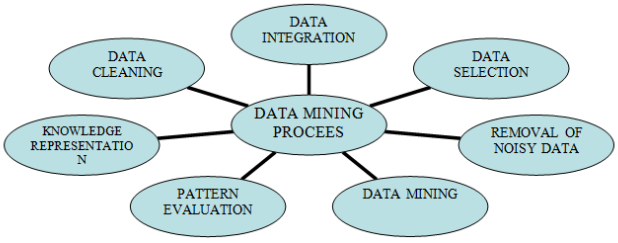

Figure 1 Data Mining Process

III. DATA MINING TECHNIQUES

There are following techniques in data mining that can be used.

1) ***Rough Set Techniques***: By using rough set techniques, the main goal is to obtain learning from the approximation of concepts. It gives us a scope to find out latent patterns within the data provided. It can be used for getting features and patterns and also for generation of rules used in decision tree.
2) ***Fuzzy Logic:*** An environment which consists of imprecision and risk,where we actually dealing with the fuzzified values whose range lies in between 0 and 1.Here membership functions are deployed that allows to map a value in fuzzified form.
3) ***Machine Learning*****:** In the models of machine learning, there will be information of non-

quantifiable.The independence of variables and parameters is divided and preassumed by the methods or models. It is an important part of Artificial intelligence in which the computers achieved the ability to learn without being programmed. These can be supervised learning or unsupervised learning.

4) ***Neural Network*:** Neural Network or an artificial neural network is a biological system that detects various patterns and makes predictions according to prespecified pattern. The greatest breakthroughs in neural network in recent years are in their application to real world problems like customer response prediction, fraud detection etc. Artificial neural network have become very versatile tool in tasks like speech recognition, pattern recognition, decision problem or prediction applications.

IV. DATA MINING TOOLS

There are various tools that are deployed in data mining. Some of them are listed below:-

A. ***WEKA(WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS):***Weka tool is the most commonly used tool in data mining.It has value due to its high functionality and good features. This is java based data mining tool. It has the ability to provide user with both GUI and simple CLI for performing and managing tasks to be performed. It supports all relevant data mining tasks from preprocessing, classification, and clustering to visualization and feature selection[1]. GNU has declared WEKA as free of cost under General Public License, which is a big advantage as compared to Rapid Miner, because users can customize it according to their will[2]. Weka is data mining software that utilizes a bunch of machine learning algorithms. These algorithms can be applied directly to the data or called from the Java code[6]. Weka contains a group of visualization tools and algorithms for analysing of data and predictive modeling, together with graphical user interfaces[8].

Figure 2 Weka Tool[7]

B. ***RAPIDMINER:***-RapidMiner is a data mining software platform which is developed by the company of the same name that provides an integrated environment for preparing data, machine learning, deep learning , text mining, and predictive analytics. It is used for business based applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization[9].RapidMiner utilizes a client/server model with the server offered as cloud infrastructures. RapidMiner is written in the Java programming language. RapidMiner provides a GUI to design and execute analytical workflows.According to Bloor Research, RapidMiner is able to provide template based work which minimizes the need of writing huge source codes.

Figure 3 RAPIDMINER[13]

C. ***KNIME*:** KNIME is an open source data mining tool which was developed by a Swiss company KNIME.com AG, Germany. The tool is named as Konstanz Information Miner (KNIME)as its development started at University of Konstanz, Germany and thus Like Orange, it is also component based tool for machine learning and data mining. Its first version came into the market in 2006.

KNIME runs on Linux, OS X and Windows platforms. KNIME has been used in Customer data analysis, pharmaceutical research, business intelligence and financial data analysis[3]. KNIME provides with a great ability to establish database connections with as many number of databases that supported alongwith JDBC. This tool also provides with a unique functionality in many aspects. [4].

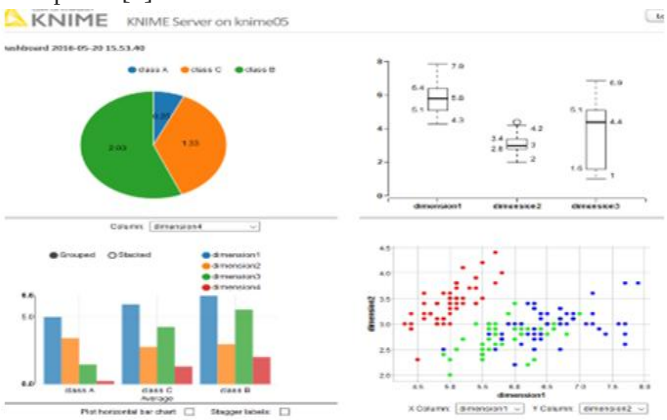

Figure 4 KNIME[12]

D. ***ORANGE*:** Orange is a free and open source component based data mining and machine learning software suite[5]. It contains full set of components which are deployed for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It uses C++ components. Python scripts are deployed to access these components and sometimes orange widgets are also preferred.

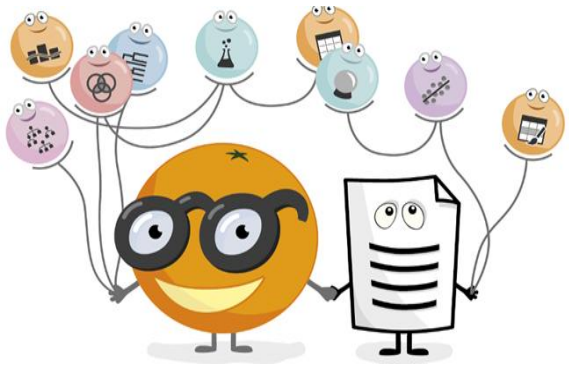

Figure 5 ORANGE[11]

E. **TANAGRA:** It was started in June 2003. The first version was launched in December 2003. Tanagra is the extension of Sipina[15], another free data mining tool which is basically deployed for the supervised learning tasks (classification), especially an interactive and visual construction of decision trees. Sipina is still available online and is maintained. Tanagra is available open source.Tanagra works as the important data mining tools. The user can design a data mining process in the form of a diagram.The output is displayed in a HTML format[14].

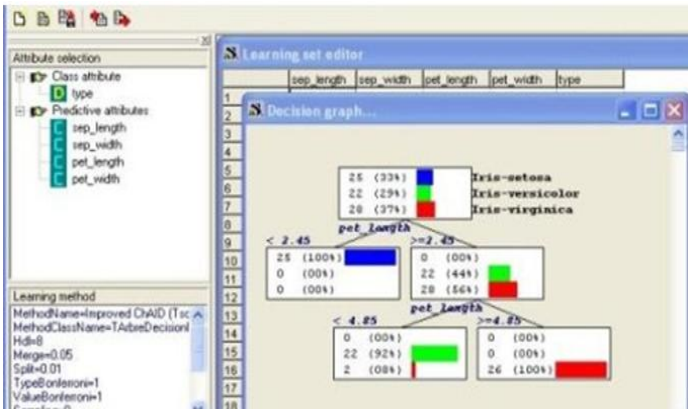

FIGURE 6 TANAGRA[15]

V. COMPARATIVE ANALYSIS OF DATA MINING TOOLS

TOOLS	PROGRAMMING LANGUAGES	FEATURES	OPERATING SYSTEM	LICENSE
WEKA	JAVA	For classification, Clustering and regression mainly	Windows, macOS and Linux	General Public license[10]
RAPID MINER	JAVA	For performing predictive analysis	Windows, macOS and Linux	Free and paid both
KNIME	JAVA	For performing predictive analysis	Windows, macOS and Linux	Free and paid both

ORANGE	C++	Provides good visualized data	Windows, macOS and Linux	General Public license
TANAGRA	R	Clustering, Classification and Association rule mining	Windows	General Public license

Table 1 Comparative Analysis of Tools

VI. CONCLUSION

In this review paper, data mining process, various techniques of data mining and tools of data mining are explained. Comparative analysis of various tools is also represented through table.In future work, author will try to focus on comparison of the tools basis on the implementation of these tools.

REFERENCES

[1] J. August, I. No, and A. Sharma, "Available Online at www.ijarcs.info International Journal of Advanced Research in Computer Science A RESEARCH REVIEW ON COMPARATIVE ANALYSIS OF DATA MINING TOOLS , TECHNIQUES AND PARAMETERS," vol. 8, no. 7, pp. 523–529, 2017.

[2] K. Kaur and S. Dhiman, "Review of Data Mining with Weka Tool," vol. 4, no. 8, pp. 41–44, 2016.

[3] R. Ri and D. W. D. Xvhg, "&rpsdudwlyh 6wxg\ ri 'dwd 0lqlqj 7rrov xvhg iru &oxvwhulqj," pp. 3971–3975, 2016.

[4] H. Solanki, "Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory," vol. 7, no. 1, pp. 23–28, 2013.

[5] V. Gupta and P. Devanand, "A survey on Data Mining : Tools , Techniques , Applications , Trends and Issues .," vol. 4, pp. 1–14, 2013.

[6] https://opensourceforu.com/2017/01/an-introduction-to-weka/

[7] http://www.scientificdatabases.ca/wp-content/uploads/2015/08/Imagen2.png

[8] https://en.wikipedia.org/wiki/Weka_(machine_learning)

[9] https://hortonworks.com/partner/rapidminer/

[10] https://www.1and1.com/digitalguide/online-marketing/web-analytics/a-comparison-of-data-mining-tools/

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

[12] https://www.knime.com/knime-server

[13] https://en.wikipedia.org/wiki/RapidMiner

[14] https://en.wikipedia.org/wiki/Tanagra

[15]http://www.researchpipeline.com/mediawiki/index.php?title=SIPINA