Classification of Data Mining and Analysis for Predicting Diabetes Subtypes using WEKA

Ekta
M.Tech. (Software Engineering), Department of Computer Science & Engineering, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra (K.U.K)-136119, Haryana, INDIA
Email: ekta.atreja07@gmail.com
Sanjeev Dhawan
Faculty of Computer Science & Engineering, Department of Computer Science & Engineering, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra (K.U.K)-136119, Haryana, INDIA
Email: rsdhawan@rediffmail.com

Abstract
From the past years Online Social Networks (OSNs) are gaining popularity among the people. Most of the people are using the Social Networking sites to communicate with their loved once, colleagues, friends etc. Data is also growing very faster and there are some issues related to privacy and security in OSNs. Therefore, it is necessary to get the information about the issues and the latest trends in OSNs. In OSNs, the generated datasets have to be analyzed and visualized by users. In order to analyze, classify and visualize the data, WEKA tool is used. In this research paper, holistic approach has been considered to analyze and classify the diabetes dataset for data preprocessing. In order to do so, diabetes.arff dataset is used for data preprocessing and prediction of diabetes. From this research work one can easily analyzed that WEKA tool is quite useful for analyzing the given dataset. Results of the analysis help to predict about the individuals are infected from diabetes or not. Finally, it is analyzed that persons who are suffered from diabetes have age more than 40 and mass more than 35. On the other hand, the persons who are not infected from diabetes have age less than 30 and mass less than 35.

Keywords: Data Analysis, Data mining, Data Preprocessing, Online Social Networks, WEKA.

1. Introduction

In Social Networks there are many users connecting to the each other for sharing their feelings, ideas, and media etc. When users are interacting with each other and data is generating constantly at higher rate and in large volume. Social Networks are not only responsible for the generation of data, in industries and organizations data is generating due to large scale of productions and developments. For making improvements in their products and services Social Networking sites and organizations need to mine the historical data. To analyze data that should not be modified by any other person who does not have authority. Similarly, Dhawan and Ekta [1] discussed various fake profile detection techniques in Social Networks which helped to protect the user's data from being damaged or modified. To analyze the data firstly data should not damaged. By using different techniques data can be protected and data can be analyzed. Abdul and Ali [2] classified the lung cancer using the data mining technique. In another research, Priyanka *et al.* [3] evaluated the performance of the faculty using the data mining technique. Data mining and data visualization is the important aspect for the organizations and Social Networking sites. Sudhir and Kodge [4] mentioned that one of the biggest challenges is data mining. Mining the useful information from the different collected dataset is not an easy task. Hina [5] explained predictive analytics using data mining techniques. WEKA (Waikato Environment for Knowledge Analysis) is one of the most popular tools used in data mining and the visualization of

data. WEKA is written in Java and open source software tool [6].

2. WEKA Toolkit

This section presents the brief overview of WEKA. WEKA (Waikato Environment for Knowledge Analysis) is most popular and widely used open source software written in Java. WEKA was developed at the University of Waikato, New Zealand. WEKA is available free of cost under the GNU General Public License [7]. This tool processes ".ARFF" file as an input in its explorer. After the selection of the dataset in the form .ARFF files it can perform classification, clustering and association etc [8]. To install WEKA tool the latest version of Java in system should be installed. After the installation of Java, WEKA can be downloaded from the official site that is http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

3. Data processing, methodology and results

To predict of the individual is infected by the diabetes or not, the required dataset was downloaded from the available link http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/. This dataset is in the form of ARFF file. This diabetes.arff dataset contains different attributes those can be useful for the prediction of the diabetes. When ARFF file is processed in WEKA which contains list of attributes and parameters as shown in fig. 1.

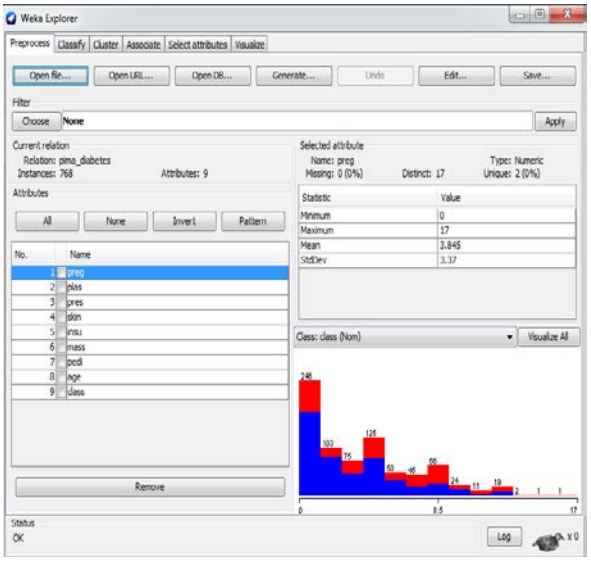

Fig. 1: ARFF file processed in WEKA

In above data file there are different attributes like age, class, mass, preg, plas, skin etc. In WEKA, data can be processed and analyzed using different data mining techniques like clustering, classification, visualization etc. Fig. 3 shows graphical representation of the processed attributes.

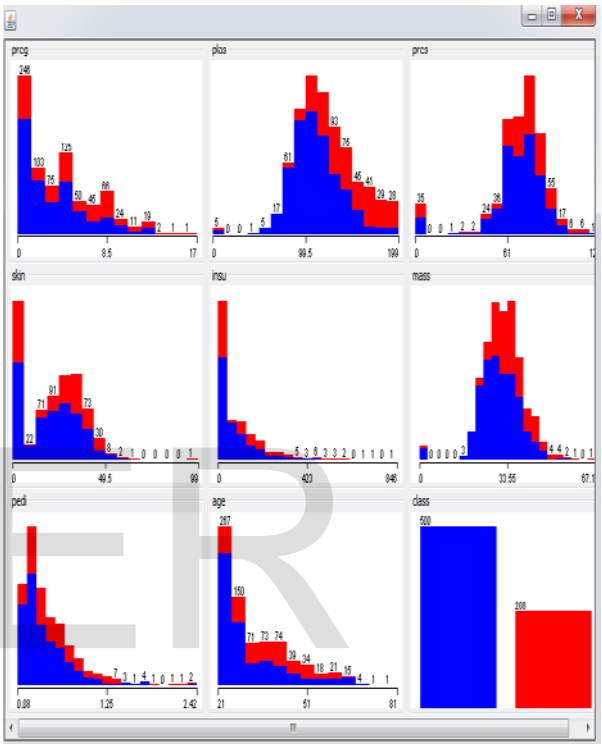

Fig. 2: Graphical representation of the processed attributes

In order to get the information regarding diabetes, the logistic functions in classifiers have to be analyzed. The available results show that the classification is not accurate measure. Therefore, to get accuracy, ranker algorithm is applied which provides equal and refined ranking to all attributes. After the ranking of all attributes, one can omit the lower rank attribute for getting the accurate results. The fig. 3 shows the list which contains lower and higher rank attributes.

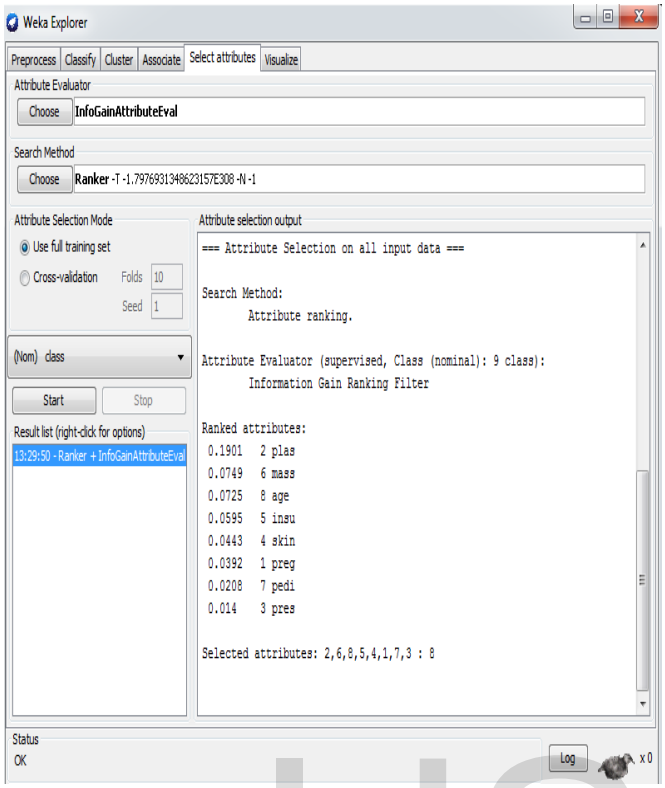

Fig 3: list of lower rank and higher rank attributes

Fig 4 shows the starting of classification process by omitting the lower rank attributes.

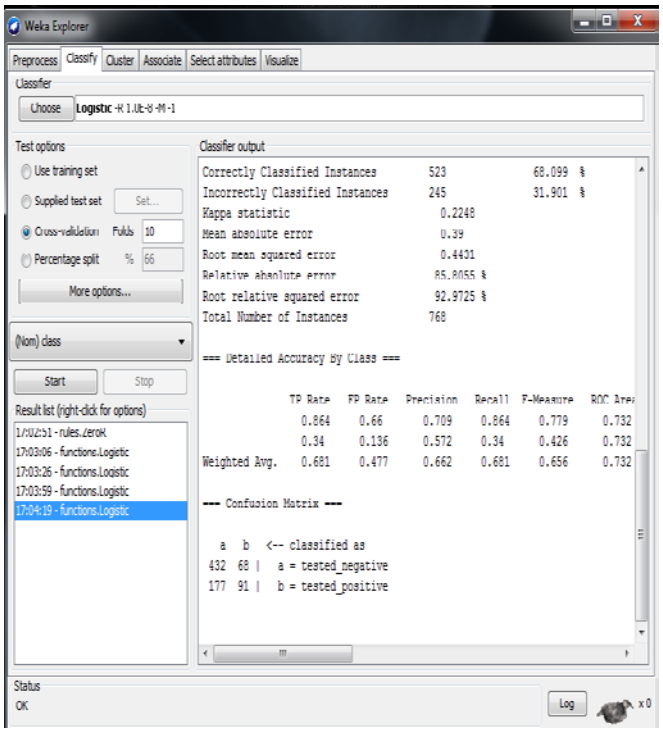

Fig 4: Results after omitting the lower rank attributes

Fig. 4 shows accurate results which are useful to predict about that individual are infected from diabetes or not. The results give the information about the diabetic and non-diabetic individuals. Visualization of the results can be done as shown in Fig. 7. In results there are test positive and test negative. There are two parameters those can be useful to predict about the diabetes like age and mass. The individuals who are not infected from diabetes have age less than 30 and mass less than 35. On the other hand those who are infected from diabetes have age group of more than 40 and mass more than 35.

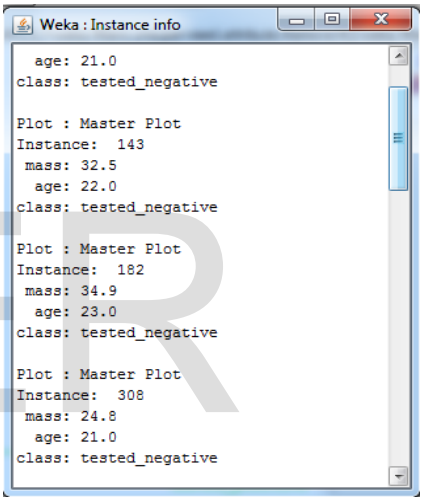

Fig. 5: Negative Individuals

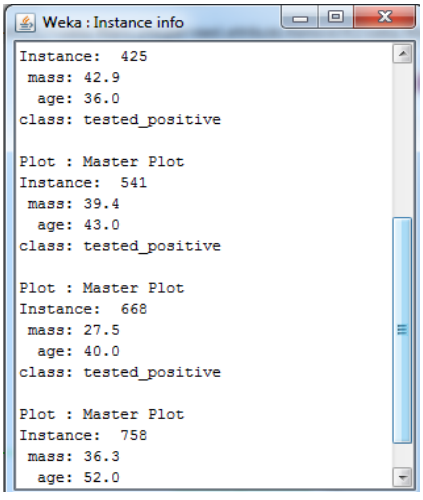

Fig. 6 : Positive individuals

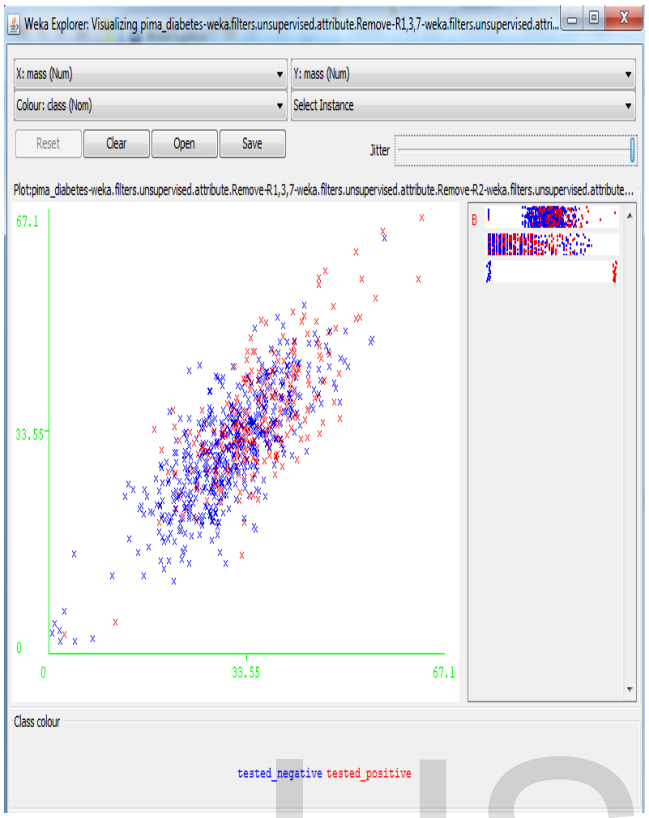

Fig. 7: Visualization of Results

4. CONCLUSION

Data mining helps to extract the information from the dataset. This information can help the industries and organization for improving their services and products. Data mining can be done using WEKA tool in efficiently and accurately. This paper shows the importance of the WEKA tool to analysis for the knowledge discovery about the diabetes. Ranker algorithm was used for the ranking of all attributes. Results help to predict that individuals are infected from diabetes or not.

5. REFERENCES

[1] Sanjeev Dhawan and Ekta, *"Implication of Various Fake Profile Detection Techniques in Social Networks"*, IOSR Journal of Computer Engineering (IOSR-JCE), AETM'16, 2016, pp. 49-55.

[2] M. Venkat Dass, Mohammed Abdul Rasheed and Mohammed Mahmood Ali, "*Classification of lung cancer subtypes by data mining technique*", 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC) , IEEE, pp. 558 – 562.

[3] Priyanka R Shah, Dinesh B Vaghela and Priyanka Sharma, *"Faculty performance evaluation based on prediction in distributed data mining"*, 2015 IEEE International Conference on Engineering and Technology (ICETECH), IEEE 2015, pp. 1 - 5 .

[4] Hina Gulati, *"Predictive analytics using data mining technique"*, 2nd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE 2015, pp. 713-716.

[5] Sudhir and Kodge, *"Census Data Mining and Data Analysis using WEKA"*, (ICETSTM–2013) International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore, pp. 35-40.

[6] WEKA, https://en.wikipedia.org/wiki/Weka_machine_learning, Accessed on 12-April - 2016.

[7] WEKA, http://www.cs.waikato.ac.nz/ml/weka/, Accessed on 12-April-2016.

[8] Arff, http://www.cs.waikato.ac.nz/ml/weka/arff.html, Accessed on 14-April-2016.