IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume4, Issue-8 E-ISSN: 2347-2693

Review of Data Mining with Weka Tool

Kulwinder Kaur[1*], Shivani Dhiman[2]

[1]*Department of Computer Science Engineering, Indus International University Una, India*
[2] *Department of Computer Applications, Indus International University Una, India*

Available online at: www.ijcseonline.org

Received: 22/Jun/2016 Revised: 10/Jul/2016 Accepted: 16/Aug/2016 Published: 31/Aug/2016

***Abstract*—** Data mining is the process of extract unseen and hidden information from a large amount of data. It is a powerful technology that helps researchers to find the meaningful information by providing different tools and technologies. In this paper we focused on different tools, technologies and application area of data mining. Also discussed the weka tool,how we build data set for weka and how this data set is loaded on weka.

Keywords-DataMining,MachineLearning,Clustering,Classification,WekaTool.

I INTRODUCTION

In this information era, a huge amount of data is collected daily. Analyzing that huge amount of data and extract meaningful information from that data is a necessity to achieve goals. Now we are living in the world where a lot of data (scientific data, medical data, banking data, marketing data & Financial data etc) related to different fields are available but nobody have time to retrieve meaningful information from this data manually. To retrieve this information in easy way, we find shortcut methods to automatically classify it, to automatically summarize it, to automatically discover & characterize trends in it [1].Data Mining discover large datasets to dig out the unknown and earlier weird pattern, relationships and knowledge that are not easy to detect with the algorithms & traditional statistical methods . Data mining has effectively been used in many fields such as marketing, banking, medical, business, fraud detection, weather forecasting etc [2].Data Mining or KDD(knowledge discovery in database) is the process to find the helpful knowledge from a collection of data. This is mostly used data mining technique in this process that includes data preparation and selection, data cleansing, incorporating earlier knowledge on data sets and interpreting perfect solutions from the pragmatic results.

II LITERATURE REVIEW

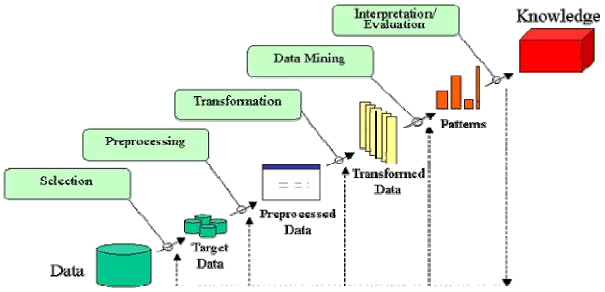


Fig.1: An outline of the steps of the KDD process

Data Selection: In this stage understand the task objectives and requirements and after that select only that data which is helpful to achieve the project goal.

Data Preprocessing: In this phase irrelevant data which is not useful in project that is removed from that data.

Data Transformation or Consolidation: In this phase selected data which is to be transformed into appropriate forms which are used for mining process.

Data Mining: This is the main & crucial phase,in which clever & intelligent techniques are applied on data so as to extract the useful patterns.

Interpretation & Pattern evaluation: In this step data sets are evaluated & represent information.

Knowledge Representation: In this phase, exposed knowledge is visually represented to the user that helps the user to simply understand the data mining result.

III Data Mining Techniques

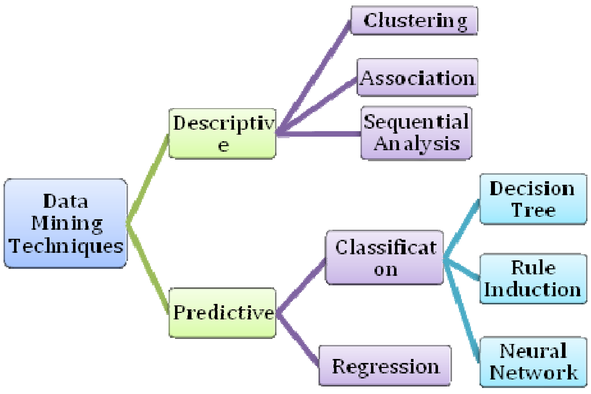


Fig. 2: Data mining different techniques

There are a number of core techniques that are used in data mining describes the type of mining and data recovery operations such as Association, Classification, Clustering, Prediction, Sequential Patterns, Decision Trees.

Association or Relation: Association is the best known data mining technique. Association is the process of finding the relationships between different modules that are present in same database. This technique is used to find out relevant modules from the database such as to find out how the purchase behaviour of one item affects the purchase behaviour of another item. Association rules are created by analyzing the data for frequent if/then pattern and using the criteria support and confidence to identify the relationship.

Classification: Classification is mainly a machine learning technique in data mining that assigns and identify the objects in a group to target categories or classes. In Classification method, we can use mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we build the method by using that method we can learn how to classify the data items into groups. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups [3].

Clustering: Clustering is a process to partitioning a set of data that makes a meaningful or useful cluster of objects which have similar characteristics. The clustering technique defines the classes and puts objects in each class [4].For example in prediction of blood pressure by using clustering we get cluster or we can say that list of patients which have same risk factor means this makes separate list of patients that with related risk factor.

Prediction: Prediction is one of the data mining technique that discovers relation between independent & dependent variable. [4]

IV Data Mining Applications

Data mining is used in various applications some of them are given below:

Sales & Marketing: Data mining enables us to understand the hidden patterns inside the data that helps in planning strategy of marketing.

Health Care Industry: The growth of health industry is increasing day by day. Data Mining helps to store all the data of patients those who are suffering from same type of disease.

Education & Sports: In this field a vast amount of statistics data are collected for each student, teacher, subject and session. Data mining can be used by education organizations in the form of statistical analysis, pattern discovery as well as for prediction.

Telecommunication & fraud detection: In today world where a large amount of data saved on cloud daily, that is the reason to increase the fraud and crime cases. To control these types of fraud cases industries & companines now use Data mining.

V Tools for Data Mining

Data plays a very important role in today's world, most of the data is in structured form and as well as in unstructured form. A lot of the data is in unstructured form and it takes a procedure and system to extract useful information from the data and transform it into understandable and usable form. Number of tools are available for data mining tasks that used artificial intelligence, machine learning and other techniques to extract data. Here are some of the powerful open source data mining tools :

Orange: Orange is an open source data mining tool written in python language. It is a component based & machine learning tool which is used for data visualization. In this tool data mining can be done through visual programming & python scripting.

Rapid Miner: Rapid Miner is written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. Rapid Miner also provides functionality like data pre-processing, visualization, predictive analytics, statistical modelling, evaluation and deployment. Rapid Miner is used in business. Industrial application, research, education etc.

Weka: The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modelling. Its free under the GNU General Public License, which is a big plus compared to Rapid Miner, because users can customize it however they please. WEKA supports several standard data mining tasks, including data pre-processing, clustering, classification, regression, visualization and feature selection. WEKA would be more powerful with the addition of sequence modelling, which currently is not included.

VI Literature survey on Data Mining

In paper [5] the author represent different survey papers in which one or more algorithms used in prediction of heart disease.By applying different algorithms the best results found by the neural networks that gives the 100% accuracy & decision tree gives 99.62 % accuracy of results in perdiction of heart disease.

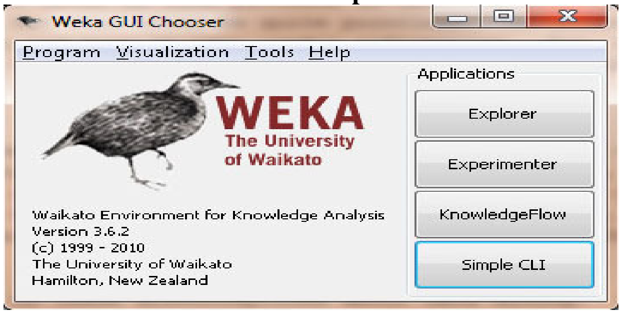
In Paper [6] the aim of the author is to investigate the performance of different classification & clustering methods on a large data set breast cancer. By applying different algorithms the finest results are found by using Bayes Network classifier with the 89.71% accuracy & the time taken to build the model is at 0.19 seconds.

In paper [7] the objective of the paper is to predict the more accurate results in heart disease. In this paper the author apply three algorithms on heart disease data set these are naive bays, j48 decision tree and bagging algorithm and in result the bagging is the one of the successful data mining technique used in to diagnosis of heart disease patients. The results show that bagging algorithm accuracy is 80.03% and time taken to build the model is .05 seconds.

VII Weka

Data mining isn't only the ground of large companies and costly software. In fact, there's a bit of software that does almost all the same things as these expensive pieces of software — the software is called WEKA. WEKA is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results [3].Weka includes so many machine learning algorithms for data mining tasks.

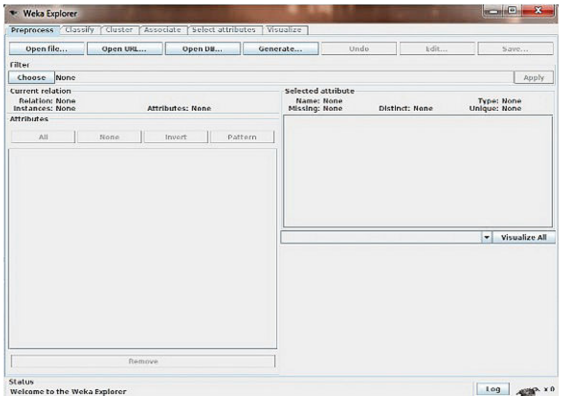
WEKA start-up screen



Waikato Environment for Knowledge Analysis
Version 3.6.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

Fig. 3: Weka startup screen

When you start WEKA, the GUI chooser window open and lets you choose four ways to work with WEKA. According to our data set we choose only the **Explorer** option. This option is more than enough for everything explorer option provides a variety of algorithms to work on data set.



Filter: None

Relation: None
Instances: None
Attributes: None

Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Status
Welcome to the Weka Explorer

Fig. 4: Weka Explore

A. Building the data set for WEKA

WEKA is a tool, that accepts Data set as input in Attribute-Relation File Format (ARFF). In the ARFF data file, you define each column and what each column contains. The ARFF file we'll be using with WEKA appears below [3].

WEKA FILE FORMAT IN ARFF FORM

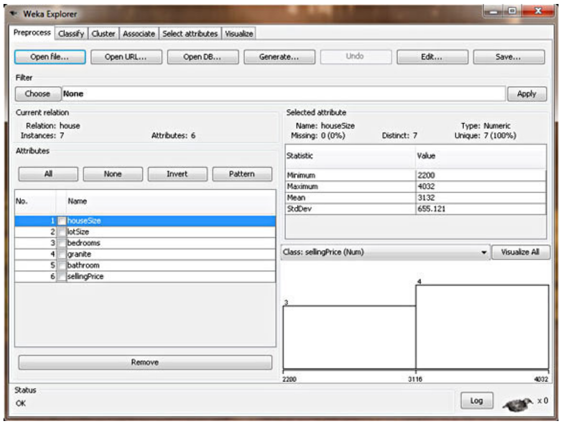
```
@RELATION house

@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE lotSize NUMERIC
@ATTRIBUTE bedrooms NUMERIC
@ATTRIBUTE granite NUMERIC
@ATTRIBUTE bathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC

@DATA
3529,9191,6,0,0,205000
3247,10061,5,1,1,224900
4032,10150,5,0,1,197900
2397,14156,4,1,0,189900
2200,9600,4,0,1,195000
3536,19994,6,1,1,325000
2983,9365,5,0,1,230000
```

B. Loading the data into WEKA

Now the ARFF data file that has been created will be loaded in Weka Tool with the help of following Procedure, Start WEKA, then choose the **Explorer.** You'll be taken to the Explorer screen, with the **Pre-process** tab selected. Select the **Open File** button and select the ARFF file you created in the section above. After selecting the file, your WEKA Explorer will look like this:



Filter: None

Current relation
Relation: house
Instances: 7 Attributes: 6

Selected attribute
Name: houseSize Type: Numeric
Missing: 0 (0%) Distinct: 7 Unique: 7 (100%)

Statistic	Value
Minimum	2200
Maximum	4032
Mean	3132
StdDev	655.121

Attributes

No.		Name
1	□	houseSize
2	□	lotSize
3	□	bedrooms
4	□	granite
5	□	bathroom
6	□	sellingPrice

Class: sellingPrice (Num)

Status
OK

Fig. 5: WEKA with house data loaded

In this way, WEKA allows you to review the data you're working with. In the left section of the Explorer window, it outlines all of the columns in your Attributes and the number of rows of data supplied. By selecting a column, information about the data in that column of your data set will be shown. For example, by selecting the **house Size** column in the left section, the right-section should change to show you additional statistical information about the column. It shows the maximum value in the data set for this column is 4,032 square feet, and the minimum is 2,200 square feet. The average size is 3,131 square feet, with a standard deviation of 655 square feet. Finally, there's a visual way of examining the data, which you can see by clicking the **Visualize All** button.

Conclusion

This paper has attempted to review the extremely dynamic and substantial area data mining. In this paper we discussed the basic process of data mining, Importance of Data mining, Different strategies that are used for data mining like classification, prediction, clustering, and association rules, Different phases in order to find the useful patterns and knowledge. We also discussed about the tools available now used for various operations related to Data mining like WEKA, Orange, Rapid Miner etc.In this paper we briefly discussed weka tool. What type of file weka tool received and how we upload file in weka etc.

References

[1] Jiawei Han, Micheline Kamber " Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Second Edtion-**2006**, ISBN: ISBN: 978-1-5090-0669-4

[2] Ravneet Jyot Singh, Williamjeet Singh "Data Mining in Healthcare for Daibetes Mellitus", International Journal of Science and Research, Volume-**03**, Issue-**07**, Page No (**1993-1998**), **July 2014**.

[3] Mansi Gera, Shivani Goel "Data Mining – Techniques , Methods and Algorithms: A Review on Tools and their Validity", International Journal of Computer Applications, Volume-**113**, Issue-**19**,Page No (**22-29**),**March2015.**

[4] A. Michael, "IBM developerWorks : IBM's resource for developers and IT," 27 April 2010. [Online]. Available: http://www.ibm.com/developerworks/library/ba-data-mining-techniques/.

[5] Beant Kaur, Williamjeet Singh "Review on heart disease prediction using data mining techniques," International Journal on recent and innovation trends in computer and communication , Volume- **2**, Issue-**10**,Page No(**3003-3008**), **October2014**.

[6] Vikas Chaurasia, Saurabh Pal "Data Mining Approach to Detect Heart Dieses," International Jouranal of Advanced Computer Science and Information Technology,Volume-**02**, Issue-**04**, Page No (56**-66**), **2013.**

[7] Mohd Fauzi bin Othman, Thomas Moh Shan Yau "Comparision of different classificaton techniques using WEKA for Breast Cancer," Springer, Volume-**15**, Issue-**04**, Page No (**520-523**), **2007.**

AUTHORS PROFILE

Kulwinder Kaur is currently pursuing M.Tech in Department of Computer Science and Engineeirng, Indus International University Una, India. Her research interests include applications of Data Mining in Medical sector.

Shivani Dhiman, B.Tech, Mtech (Computer Science and Engineering) presently working as Asst. Prof. in the Department of Computer Applications at Indus International University Una, India. Her area of interests include .