IJETCR
Open Access
ISSN: 2348 - 2117
International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 2 Issue 7; December; 2014; Page No. 81-87

A Pragmatic Approach on Knowledge Discovery in Databases with WEKA

D. Asir Antony Gnana Singh[1], E. Jebamalar Leavline[2]

[1]Department of Computer Science and Engineering,

Bharathidasan Institute of Technology, Anna University, Tiruchirappalli-India

asirantony@gmail.com

[2]Department of Electronics and Communication,

Bharathidasan Institute of Technology, Anna University, Tiruchirappalli-India

jebilee@gmail.com

ABSTRACT

Discovering the knowledge from the databases is very essential in improving the process and the activities in all the fields such as medical, engineering, management, media, agriculture and etc. The data are being generated from all the fields drastically due to the massive growth of information and communication technology. Extracting the knowledge from these massive data is a challenging task among the researchers and data analyst. Knowledge discovery is a sequence of processes in which various tasks are performed on the data. This paper presents a conceptual view of the knowledge discovery from the databases and provides a pragmatic approach on knowledge discovery in databases with WEKA.

INTRODUCTION:

In the recent past, due to the rapid growth of the information technology the data are being generated in a drastic way. Therefore, obtaining the knowledge from the massive data is a challenging task among the researchers. The knowledge discovery is a sequence of processes that is illustrated in Figure 1. This process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. In the data cleaning stage, the data are cleaned by removing the noise and inconsistent data since the real world data are noisy and inconsistent. The data can be in the form of text, image, signals, etc. In the data integration stage, the data are collected from the various sources and stored in a unified scheme by integrating the data received from the various sources. In the data selection stage, the needed data for the analysis are taken from the integrated data in order to do further process. In the data transformation stage, the data are transformed to be compatible to the mining process.

The transformation can be carried out using aggregation or summary or etc. In the data mining stage, the desired data pattern is extracted from the transformed data. The data mining is an important stage in the knowledge discovery process. The extracted data pattern can be used to extract the knowledge from the data. The stages of data cleaning, data integration, data selection, data transformation are known as data preprocessing. The extracted data pattern can be interpreted and analyzed in the in the stage of the data pattern evaluation. Then the knowledge is expressed in the knowledge presentation stage.

Figure 1: The stages of knowledge discovery from the databases

The WEKA is data mining software that consists of preprocessing algorithms such as data filtering algorithms, dimensionality reduction algorithms, etc. and various data mining algorithms for developing the data models for various applications such as clustering, classification, association, and etc. The preprocessing technique improves the quality of the data models which are constructed using the data mining algorithms. In the recent past, various types of classification algorithms are

Corresponding author: D. Asir Antony Gnana Singh

developed by the researchers with different classification techniques such as tree-based, probabilistic-based, rule-based, etc. Likewise, various clustering methods are developed based on the clustering techniques such as distance-based, density-based, etc.

1. CONDUCTING EXPERIMENTS WITH WEKA EXPLORER:

This section expresses the conduction of the experiments with the WEKA explorer. It gives an insight into the predication using classification algorithm with WEKA.

A. Preparation of dataset

The data are prepared in attribute relation file format (arff), since this format files occupy less space in memory even for larger data. The following steps are followed for preparing the data set. The dataset can be classified into two types such as labeled and unlabeled dataset based on the data mining task. The label represents the predictive attribute for an instance of the dataset. The labeled data is used in classification task and the unlabeled data is used for clustering task.

Table: 1 Details of a real-estate labeled dataset

Age	Income	Marital status	Own car	Own motorcycle	Own bicycle	Buy house in apartment
10	less	No	No	no	yes	no
35	more	Yes	Yes	yes	No	yes
40	average	Yes	No	yes	yes	yes
60	more	Yes	Yes	yes	yes	no
25	less	No	No	yes	yes	no

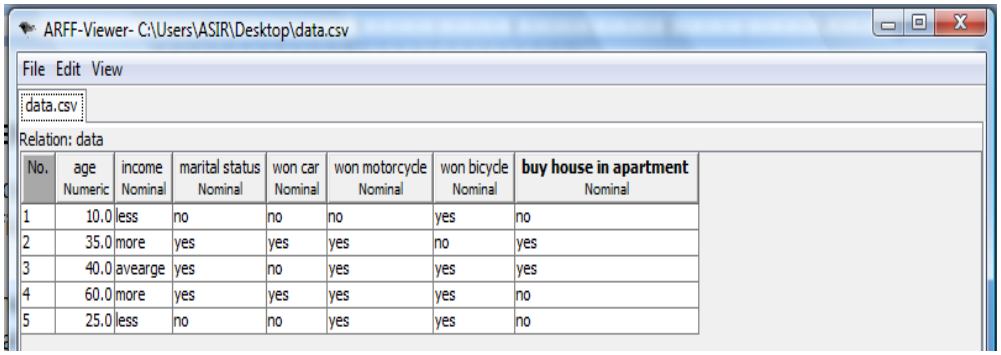
Table 2: The line of codes for arff for the labeled real-estate dataset

```
@relation real-estate_data

@attribute 'age' numeric
@attribute 'income' {less,more,avearge}
@attribute 'marital status' {no,yes}
@attribute 'own car' {no,yes}
@attribute 'own motorcycle ' {no,yes}
@attribute 'own bicycle ' {yes,no}
@attribute 'buy house in apartment   ' {no,yes}

@data
10,less,no,no,no,yes,no
35,more,yes,yes,yes,no,yes
40,avearge,yes,no,yes,yes,yes
60,more,yes,yes,yes,yes,no
25,less,no,no,yes,yes,no
```

The dataset are further classified as test dataset and training dataset. The training dataset can be used for building the classification model or clustering model for predication. Table 1 shows the details of a real-estate labeled dataset. Table 2 shows the line of codes to create the labeled real-estate in arff format. Figure 2 shows the display of the real-estate dataset using WEKA arff-viewer.



No.	age Numeric	income Nominal	marital status Nominal	won car Nominal	won motorcycle Nominal	won bicycle Nominal	buy house in apartment Nominal
1	10.0	less	no	no	no	yes	no
2	35.0	more	yes	yes	yes	no	yes
3	40.0	avearge	yes	no	yes	yes	yes
4	60.0	more	yes	yes	yes	yes	no
5	25.0	less	no	no	yes	yes	no

Figure 2: The display of real-estate dataset using arff-viewer

B. Prediction using classification algorithm

Prediction plays a vital role in the many applications that are used in the day-to-day life such as weather prediction, disease diagnosis, fault detection, etc. In weather prediction application, the natural calamities are predicted using the weather data for saving the human lives from the natural disaster. In disease diagnosis applications, the diseases are identified or predicated using the data collected from the symptoms or tests in order to provide the suitable drugs for the disease. In fault detection application, the faults are identified in various products or processes using the data concerned about that product or process. The fault detections are carried out to identify the location of the fault and rectify the fault before make a rigorous losses. In order to carry out prediction, the following steps are followed.

Step 1: Prepare the training dataset in arff (Attribute-Relation File Format) from the training data as shown in Figure 1.

Step 2: Load the dataset into WEKA explorer using open file option as shown in Figure 3.

Step 3: Build the predictive model using the classification algorithm and save the model as shown in Figure 4 and 5.

Step 4: Supply the unknown dataset as the test dataset that contains no labels for the instances since the labels are to be predicted using the built model. In the WEKA explorer, the labels of the instances are replaced by the symbol "?" such as the instance of the dataset illustrated in Figure 2 is formatted as "10,less,no,no,no,yes,?" using the supplied test set opting in the WEKA explorer as shown in Figure 6.

Step 5: Choose more option and enable the output prediction option as shown in Figure 7.

Step 6: Load the model and re-evaluate the model on current test set as shown in Figure 8 in such a way that the labels of the instances of the given test dataset or unlabeled dataset are predicted.

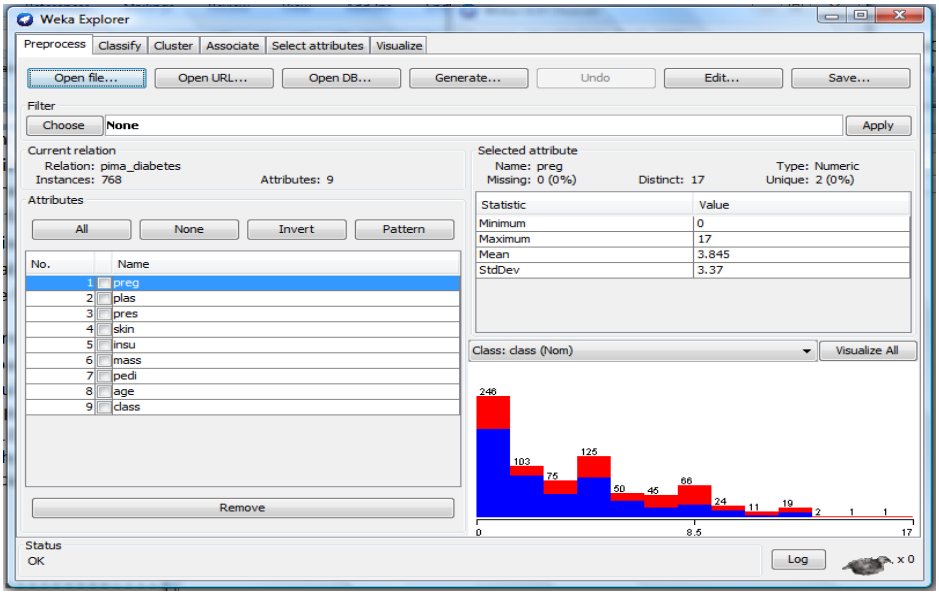


Figure 3: Loading the pima_diabetes dataset on WEKA explorer

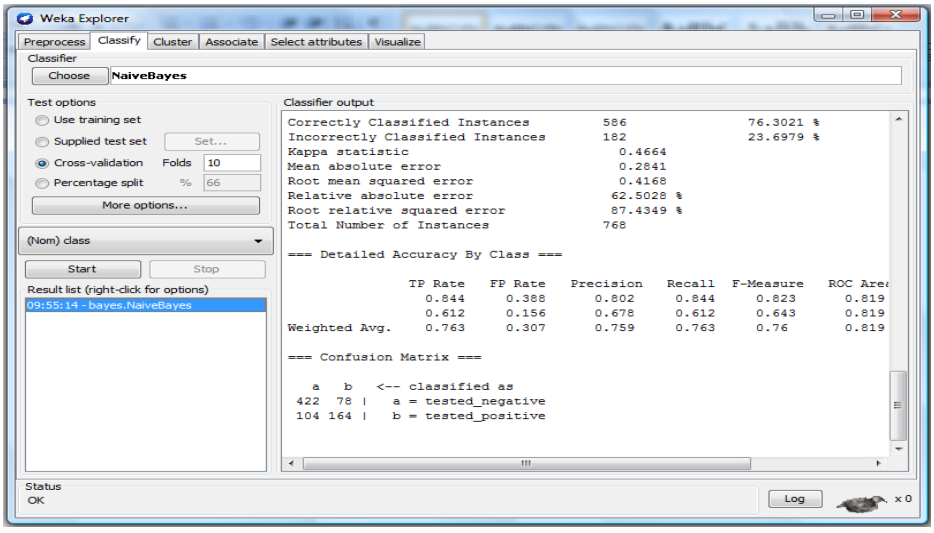


Figure 4: Building a predictive model using Naïve Bayes classifier on pima_diabetes dataset and validating the performance of the classification model using 10 fold cross-validation test options

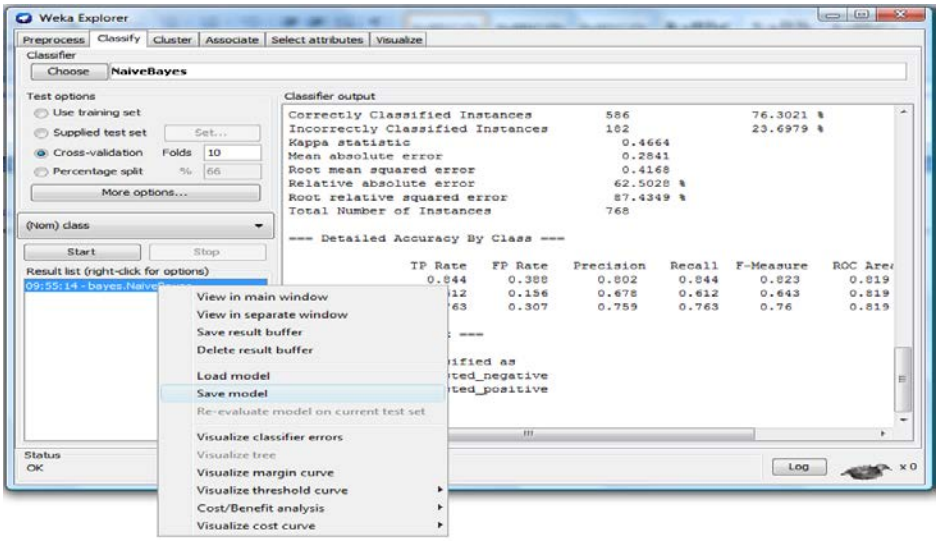


Figure 5: Illustration for saving the built classification model

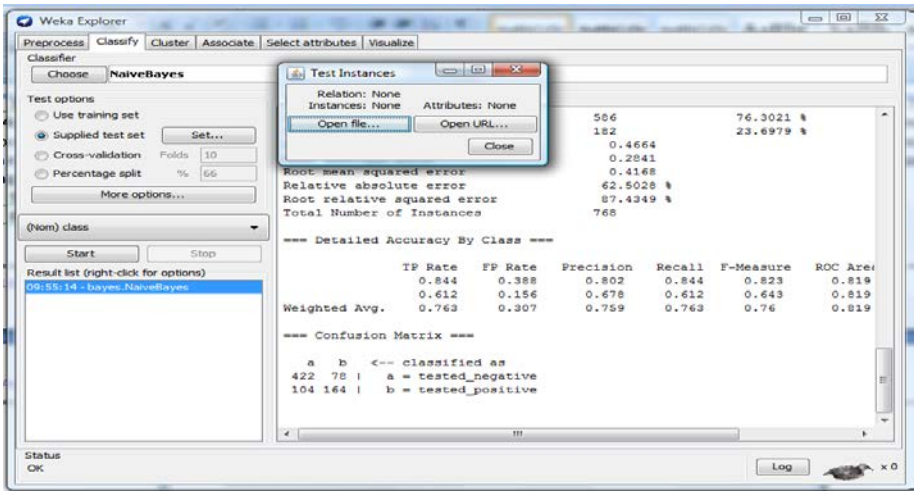


Figure 6: Illustration of supplying the test dataset using test option

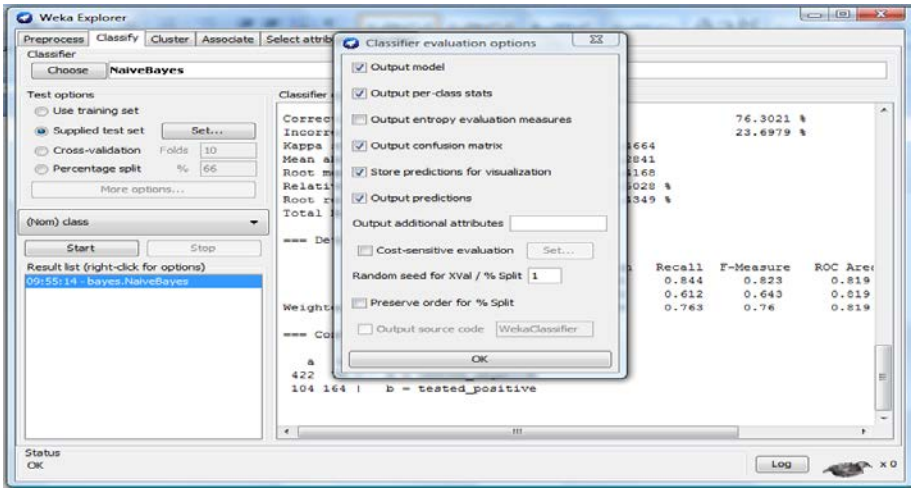


Figure 7: Enabling the output prediction option using more options button

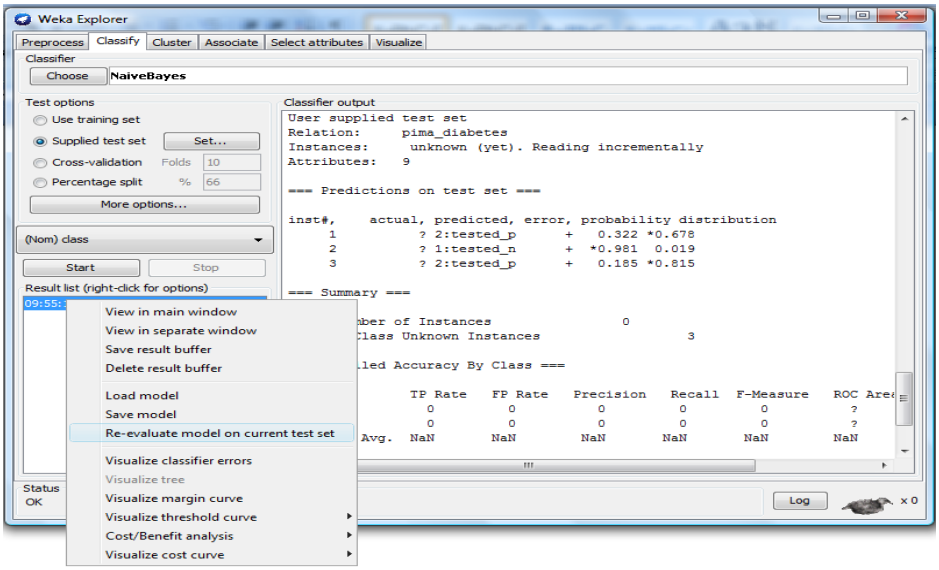


Figure 8: Re-evaluating the model using the re-evaluate model on current set option and display the predicted labels for the given test dataset

2. PREDICTION USING CLUSTERING ALGORITHM:

The clustering algorithm plays a significant role in the applications that are used in the day-to-day life such as fraud detection, outlier detection, etc. In the fraudulent detection application, the suspicious persons are identified with the pattern of abnormal operation or behaviors from the data which are generated by him. In the outlier detection, the distinguishable objects are identified that are deviated from the similar group of objects. The prediction with the clustering algorithm is carried out using the unlabeled dataset. The Table 3 and Table 4 show the dataset with unlabeled data. In this data set the X and Y are the attributes and the values are represented for each data points. The prediction is carried out using the clustering algorithm with the following steps.

Step 1: Prepare the unlabeled dataset in arff file format as shown in Table 4 and load the dataset as shown in Figure 3.

Step 2: Build the predictive model using simple as shown in Figure 9 and save the model as shown in Figure 5.

Step 3: Supply the unknown dataset as the test dataset as shown in Figure 6.

Step 4: Choose more option and enable the output prediction option as shown in Figure 7.

Step 5: Load the model and re-evaluate the model on current test set as shown in Figure 10 to predict the clusters of the given test dataset or unlabeled dataset are predicted.

Table 3: The unlabeled dataset for clustering the data objects.

X	Y
10	61
20	62
22	63
25	64
50	30

Table 4: The line of codes for arff for the unlabeled data

```
@relation 'unlabeled_data'
@attribute x numeric
@attribute y numeric
@data
10,61
20,62
22,63
25,64
50,30
```

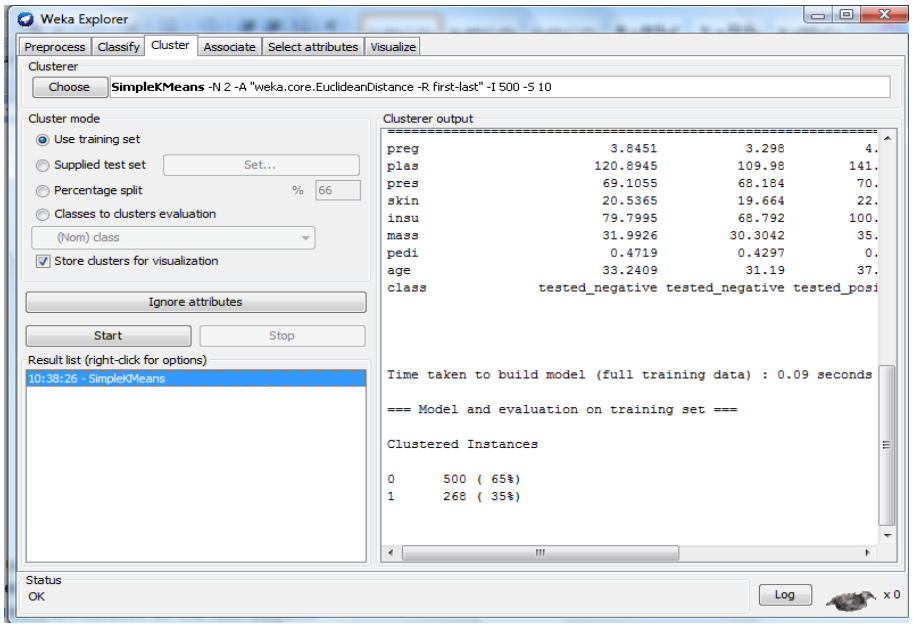


Figure 9: Building a predictive model using simple K means clustering algorithm on pima_diabetes dataset and validating the performance of the clustering model using use training set test option

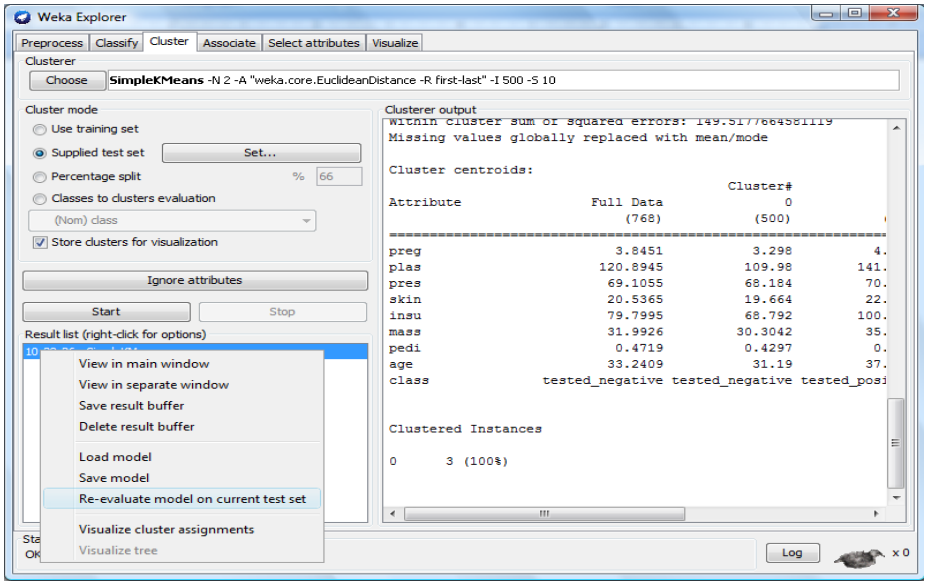


Figure 10: Re-evaluating the model using the re-evaluate model on current test set option and display the predicted cluster for the given test dataset

3. CONCLUSION:

This paper presented a pragmatic approach on knowledge discovery in databases with WEKA. It provides the overview of knowledge discovery process by exploring the details on the various stages in the knowledge discovery process that includes data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Further, it explores a study on conducting experiment with WEKA explorer. This paper also presented the preparation of the dataset and explored the classification and clustering approaches with the various steps that are to be considering while mining the data.

REFERENCES

1. J Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
2. Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.
3. Asir Antony Gnana Singh, S. Balamurugan, E. Jebamalar Leavline, 'An Empirical Study on Dimensionality Reduction and Improvement of Classification Accuracy Using Feature Subset Selection and Ranking', Proceedings of the IEEE International Conference on Emerging Trends in Science, Engineering and Technology, pp. 102–108, 2012.
4. Asir Antony Gnana Singh, S. Balamurugan, E. Jebamalar Leavline , 'Towards Higher Accuracy in Supervised Learning and Dimensionality Reduction by Attribute Subset Selection - A Pragmatic Analysis', Proceedings of the IEEE International Conference on Advanced Communication Control and Computing Technologies, pp. 125–130, 2012.
5. D. Asir Antony Gnana Singh, E. Jebamalar Leavline, Data Mining In Network Security-Techniques & Tools: A Research Perspective. Journal of Theoretical & Applied Information Technology, vol.57, 2013.