Data Mining and Expert System Based on Efficient IDS

Seema Ranga
Research Scholar
Department of Computer Science and Engineering
UIET Kurukshetra Haryana

Ajay Jangra
Assistant Professor
Department of Computer Science and Engineering
UIET Kurukshetra Haryana

Abstract

With popularization of internet, internet attack cases are also increasing, thus information safety has become a significant issue all over the world, hence nowadays, and it is an urgent need to detect, identify and hold up such attacks effectively [1]. In this modern world intrusion occurs in a fraction of seconds and Intruders cleverly use the adapted version of command and thereby erasing their footprints in audit and log files. Successful IDS intellectually differentiate both intrusive and nonintrusive records. Most of the existing systems have security breaches that make them simply vulnerable and could not be solved. Moreover substantial research has been going on intrusion detection system which is still considered as immature and not a perfect tool against intrusion. It has also become a most priority and difficult tasks for network administrators and security experts. So it cannot be replaced by more secure systems [2].

Keywords- Ids, Testing, Weka, Redundant, Classification

I. INTRODUCTION

The concept of intrusion detection system was first suggested in a technical report by Anderson [3], he considered that computer audit mechanism should be transformed and capable to give internal risks and threats for computer safety technicians, and suggested that statistics method should be applied to analyze users' behaviour and detect those masquerades that accessed system sources illegally [3]. Intrusion detection system is to supervise and control all cases happening to computer system or network system, analyze every signal arise from related safety problems, send alarms when safety problems occur, and inform related personnel or units to take applicable measures to reduce possible risks. Its framework includes three parts [4]:

1) Information collection: The source of these collected data can be separated into host, network and application, according to the position.
2) Analysis engine: Analysis engine is able to analyze whether or not there are symptom of every intrusion.
3) Response: Take actions after analysis, record analysis results, send real-time alarm, or adjust intrusion detection system, and so on.

II. CLASSIFICATION OF INTRUSION DETECTION SYSTEM

Generally speaking, there are two kinds of categorization methods for intrusion detection system:

1) According to different data sources, intrusion detection system includes host-based IDS and network-based IDS.
2) According to different analysis methods, intrusion detection system includes Misuse Detection and Anomaly Detection.

III. PROBLEM FOUNDATION

Traditional IDS has some limitations: [6]

1) Poor adaptability.
2) Inability to detect novel attacks.
3) High modelling cost.
4) Slow updating speed.
5) Lack of extensibility.

IV. PROPOSED WORK

This paper aim to drawing and develop intelligent data mining intrusion detection system and its interior part a composite detection engine with anomaly detection and misuse detection features and the two detection engines work consecutively to detect the user's activity in turn. The system collects the data of database examination system in real time, analyzes the audit data, judges that it is a normal behavior, abnormal behavior or aggressive behavior and responds to the result obtain by the operation activities and finally reports the result to the manager in a comprehensible form. The model structure is shown as

Figure 1. This part shows the steps that a data mining task is executed on local area network intrusion detection system in Weka software structure. There are four main steps to execute every data mining task with Weka software. In the following steps will show it in detail by a detection task.

1) Initial network data is collected and pretreated when network connection data including particular attributes
2) Network Data packet is generally involved in some vital attributes, such as protocol type, target IP address and flag bit.
3) Subsequently, use association examination data mining algorithm to handle the correlation data and get association rules, thereby obtaining the normal performance patterns which can be use for abnormal intrusion detection.
4) finally use classification algorithm to carry away rule mining to additional distinguish normal behaviors and intrusion behaviors and generate the rules based on misuse detection and temporarily continue to use analysis data mining algorithm to mine intrusion data sets, extract intrusion patterns, make an intrusion data feature detection model and update the model according to newly obtained data continuously, used for misuse detection [6].

V. PROPOSED ALGORITHM

A. K-Means Algorithm

1) Step 1: Initialization of centroid group.
2) Step 2: Assign each connection to the group that has the closest centroid.
3) Step 3: Recalculate the positions of the K centroids.
4) Step 4: Repeat Steps 2 and 3 until the centroids no Longer move[14].

The proposed method described aims to achieve high accuracy, high detection rate and very low or no false alarm rate. This section discusses the limitations of previous existing methods and advantages of proposed method over them. Y-means clustering algorithm has better detection rate and low false alarm rate. But it cannot solve real time anomaly detection, since it cannot update the date set dynamically during the process. The major advantages of K-means are that it is a lightweight, fast iterative algorithm which is easy to understand and implement. However, the major drawbacks are its sensitivity to initial conditions such as the number of partitions and the initial centroids, and it is also sensitive to outliers and noise. A parallel clustering ensemble algorithm forms the clusters more speedily to mass data. It also achieves high detection rate but its false alarm rate is low.

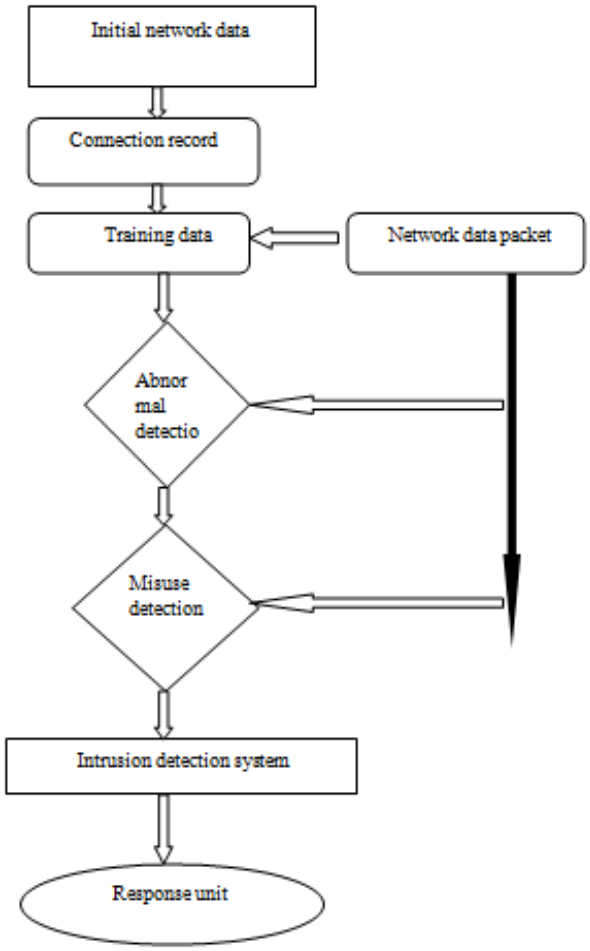

Fig. 1: working flow proposed

– Hybrid learning approach by using K-means clustering and naive bayes classification overcomes the drawback of moderate detection rate and high false alarm rate. A hybrid anomaly detection system that combines kmeans, and two classifiers: k-nearest neighbor and naive bayes overcome the drawback of very high false alarm rate in existing method. With the aim to improve detection rate and decrease false alarm rate, this presents a hybrid approach for intrusion detection system. Feature selection helps in selecting important and relevant features from the data set and reduces the time required to process the data set. Noise and outliers on the data set are reduced by applying filtering method. By using divide and merge and the density of each point the number of the cluster centroids and appropriate initial centroids are calculated automatically. This overcomes the drawback of simple K-means algorithm. Since the single clustering algorithm is difficult to get the great effective detection, clustering ensemble is employed for the effective identification of both known and unknown patterns of attacks to achieve high accuracy and detection rate as well as low false alarm rate.

VI. TOOL USED

A. Introduction to WEKA

WEKA is a data mining organization developed by the University of Waikato in New Zealand that implements data mining algorithms use the JAVA language. WEKA is an open source software issue under General Public License. It is a state of the ability facility for developing machine knowledge techniques and their application to real world data mining problems. It is a group of machine learning algorithms for data mining tasks these algorithms are directly applied to dataset. WEKA equipment algorithms for data pre-processing, categorization, regression, clustering and organization rules; it is also includes visualization tools. The new machine learning scheme can also be developed with this package [7]. Nowadays, the using of intelligent data mining approaches to guess intrusion in local area networks has been increasing rapidly. In this, an improved approach for Intrusion Detection System (IDS) based on combine data mining and expert system is presented and implemented in WEKA [8]. Weka is a gathering of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a data set or called from your hold Java code [9, 10]. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and image. It is also well-suited for developing new machine learning schemes. WEKA consists of Explorer, Experimenter, Knowledge flow, Simple direct Line Interface, Java interface [11]. The WEKA tool incorporates the following steps [12, 13]:

– Analysis and pre-processing of the features in the database and assessing the correctness of the data.
– Classification of the class attributes which divide the set of instances into the appropriate classes.
– Extraction of the potential features to be used for classification.
– Selection of a subset of features to be used in the learning process.
– Investigation of a possible inequity in the selected data set and how it may be counteract.
– Selection of a subset of the instances, i.e. the records that learning is to be based on.
– Application of a classifier algorithm for the learning process.
– Decision on a testing method to estimate the performance of the selected algorithm.

VII. RESULT ANALYSIS

The use of k-means clustering with WEKA the sample data set use for this example is based on the "bank data" available in comma-separated format bank-data.csv. This paper assumes that appropriate data pre-processing has been performed. In this case a version of the initial data set has been formed in which the ID field has been removed and the "children" attribute has been converted to categorical. The resultant data file is “bank.arff” and includes 600 instances. As an illustration of performing clustering in WEKA, we will use its performance of the K-means algorithm to cluster the customers in this bank data set, and to characterize the resulting customer segments.

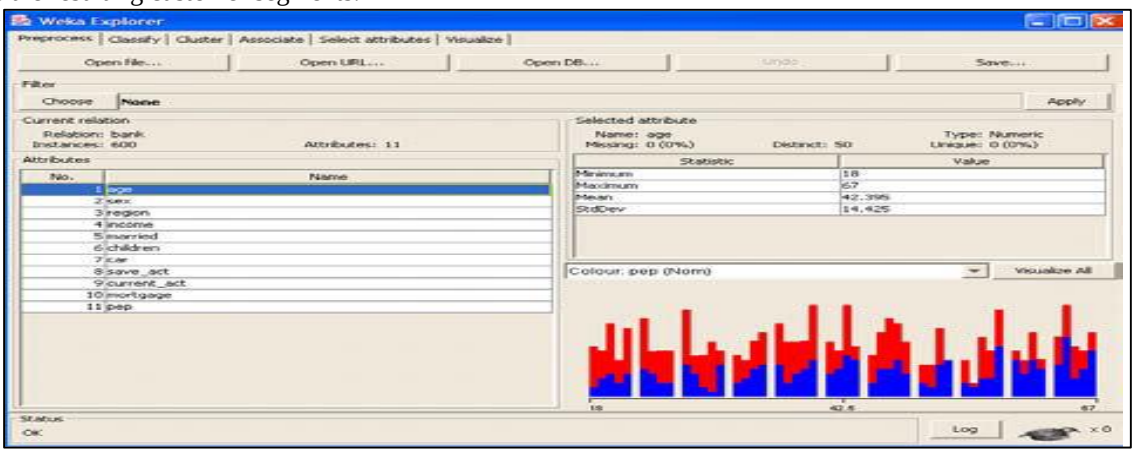

Fig. 2: Data file Bank. Data loaded

Some implementations of K-means only agree to numerical values for attributes [14]. In that case, it may be basic to change the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to regularize the values of attributes that are considered on substantially different scales (e.g., "age" and "income"). While WEKA give filters to complete all of these preprocessing tasks, they are not necessary for clustering in WEKA. This is because WEKA Simple k Means algorithm automatically handle a mixture of categorical and algebraic attributes.

Moreover, the algorithm automatically normalizes numerical attributes when doing distance computation. The WEKA SimplekMeans algorithm use Euclidean distance measure to compute distances between instance and clusters. To perform clustering, choice the "Cluster" tab in the Explorer and click on the "Choose" button. This results in a drop down list of offered clustering algorithms. In this case we select "SimpleKMeans". Next, click on the text box to the right of the "Choose" button to get the pop-up window shown in Figure 5, for editing the clustering parameter.

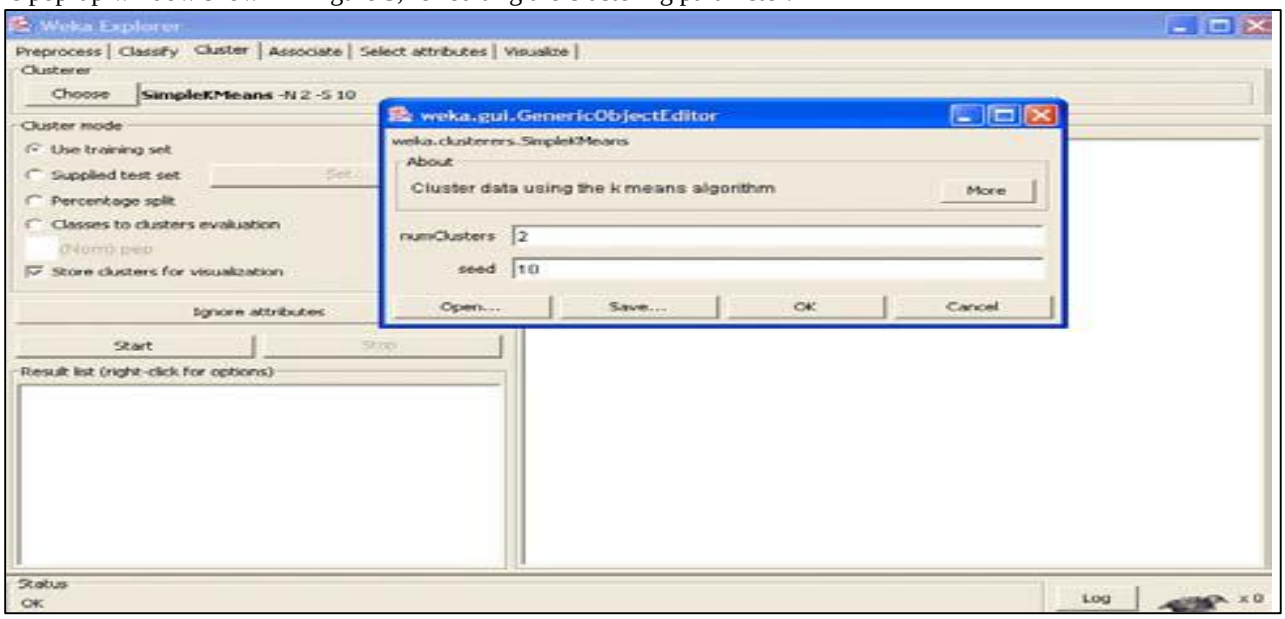

Fig. 3: Selecting Clustering Parameters.

In the pop-up window we enter 6 as the number of clusters[15] (instead of the default values of 2) and we leave the value of "seed" as is. The seed value is used in generate a random number which is, in turn, used for making the initial mission of instances to clusters. Note that, in common, K-means is quite sensitive to how clusters are initially assigned. Thus, it is often needed to try different values and evaluate the results.

Once the options have been specified, we can run the clustering algorithm. Now we make sure that in the "Cluster Mode" panel, the "Use training set" option is selected, and we click "Start". We can right click the effect set in the "Result list" panel and view the results of clustering in a separate window.

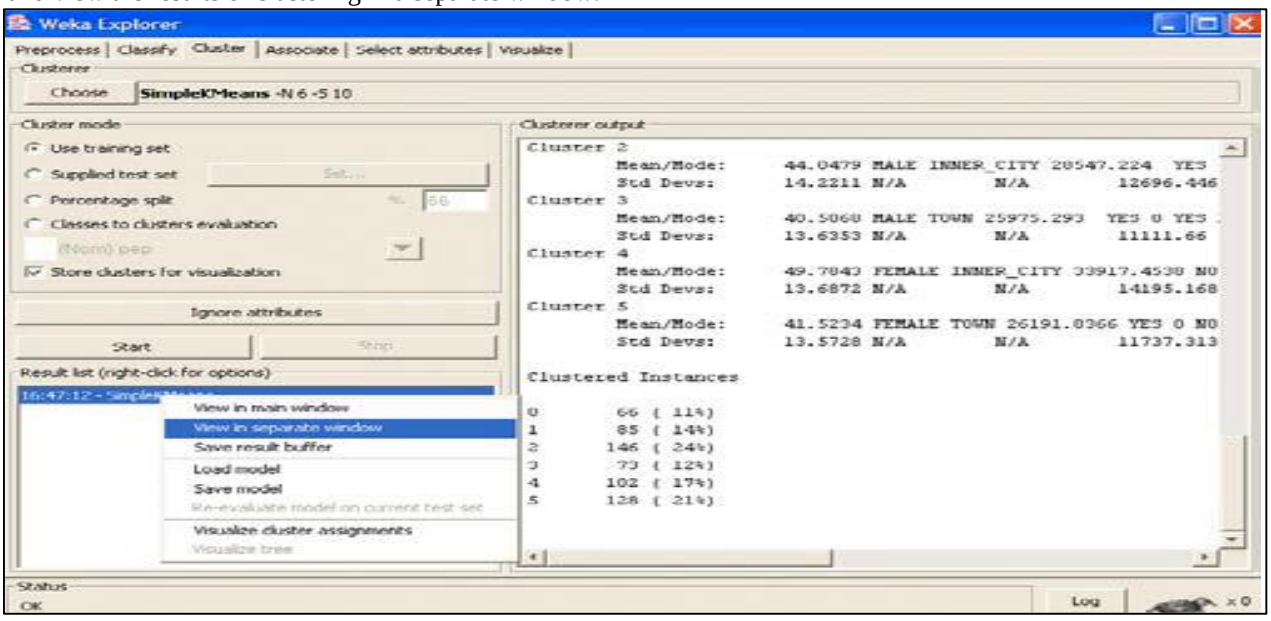

Fig. 4: Clustering in progress.

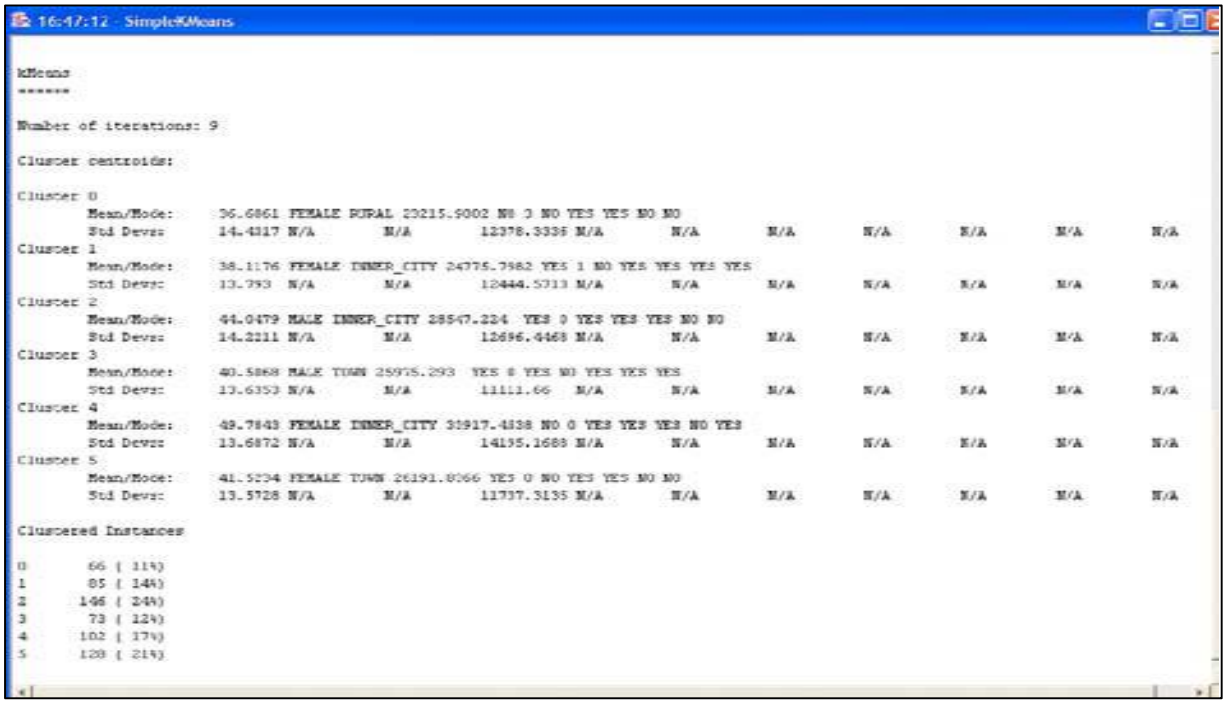

Fig. 5: Clustering data output

Fig. 6:

The result window show the centroid of each cluster as well as statistics on the number and percentage of instances allocate to different clusters. Cluster centroid is the mean vectors for each cluster (so, each dimension value in the centroid represent the mean value for that dimension in the cluster). Thus, centroids can be used to characterize the clusters. For example, the centroid for cluster 1 show that this is a segment of cases representing middle aged to young (approx. 38) females living in interior city with an average income of approx. $28,500, who are married with one child, etc. moreover, this group have on average said YES to the PEP product. Another way of understanding the characteristics of each cluster in during visualization. We can do this by right-clicking the result set on the left "Result list" panel and selecting "Visualize cluster assignments".

You can choose the cluster number and any of the other attributes for each of the three different dimensions accessible (x-axis, y-axis, and color). Different combinations of choices will result in a visual rendering of different relations within each cluster. In the above example, we have chosen the cluster number as the x-axis, the occurrence number (assigned by WEKA) as the y-axis, and the "sex" attribute as the color dimension. This will result in a revelation of the division of males and females in each cluster. For instance, you can note that clusters 2 and 3 are dominated by males, while clusters 4 and 5 are dominated by

females. In this case, by change the color dimension to other attributes, we can see their distribution within each of the clusters. Finally, we may be interested in saving the resulting data set which included each instance along with its assigned cluster.

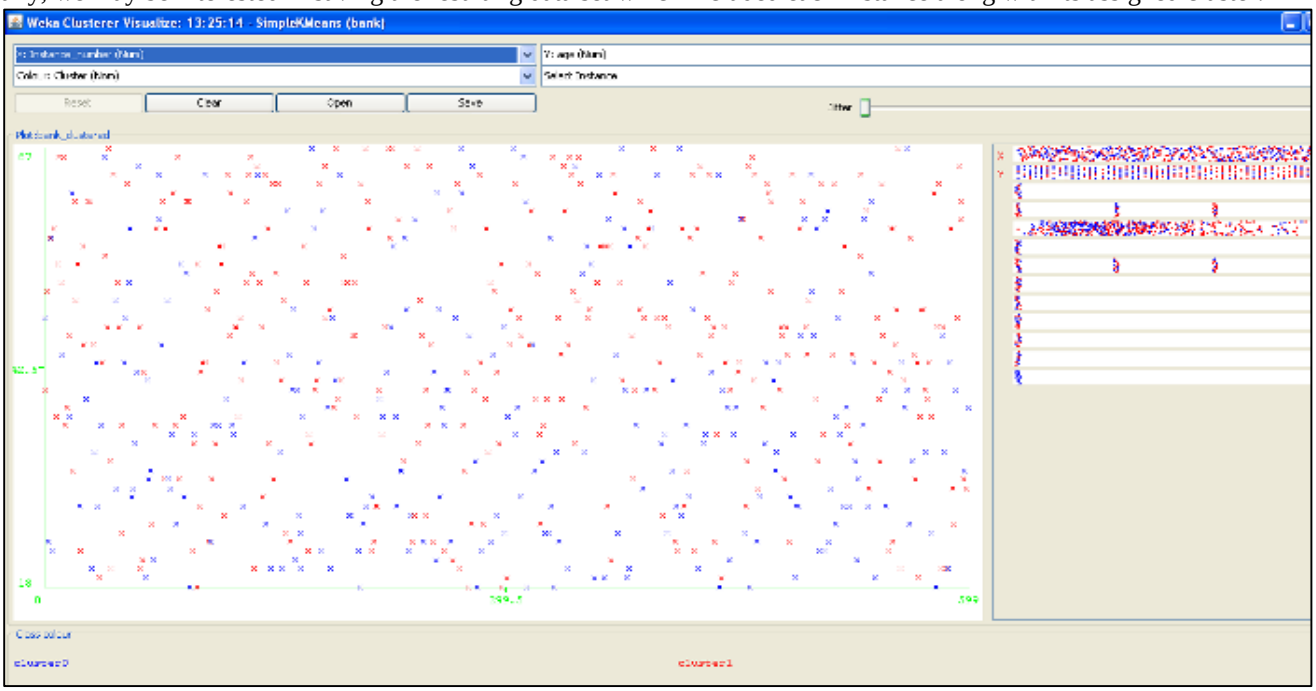

Fig. 7: Visualizing Clusters.

To do so, we click the "Save" button in the visualization window and save the result as the file “bank-kmeans.arff “. Bank-kmeans.arff

- @relation bank_clustered
- @attribute Instance_number numeric
- @attribute age numeric
- @attributesex{FEMALE,MALE}
  @attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
- @attribute income numeric
- @attribute married {NO,YES}
- @attribute children {0,1,2,3}
- @attribute car {NO,YES}
- @attribute save_act {NO,YES}
- @attribute current_ act {NO,YES}
- @attribute mortgage {NO,YES}
- @attribute pep {YES,NO}
- @attributecluster {cluster0,cluster1,cluster2,cluster3,cluster4,cluster5}
- @datA0,48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES,cluster1
- 1,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO,cluster3
- 2,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO,cluster2
- 3,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO,cluster5
- 4,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO,cluster5
- 5,57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES,cluster5
- 6,22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES,cluster0

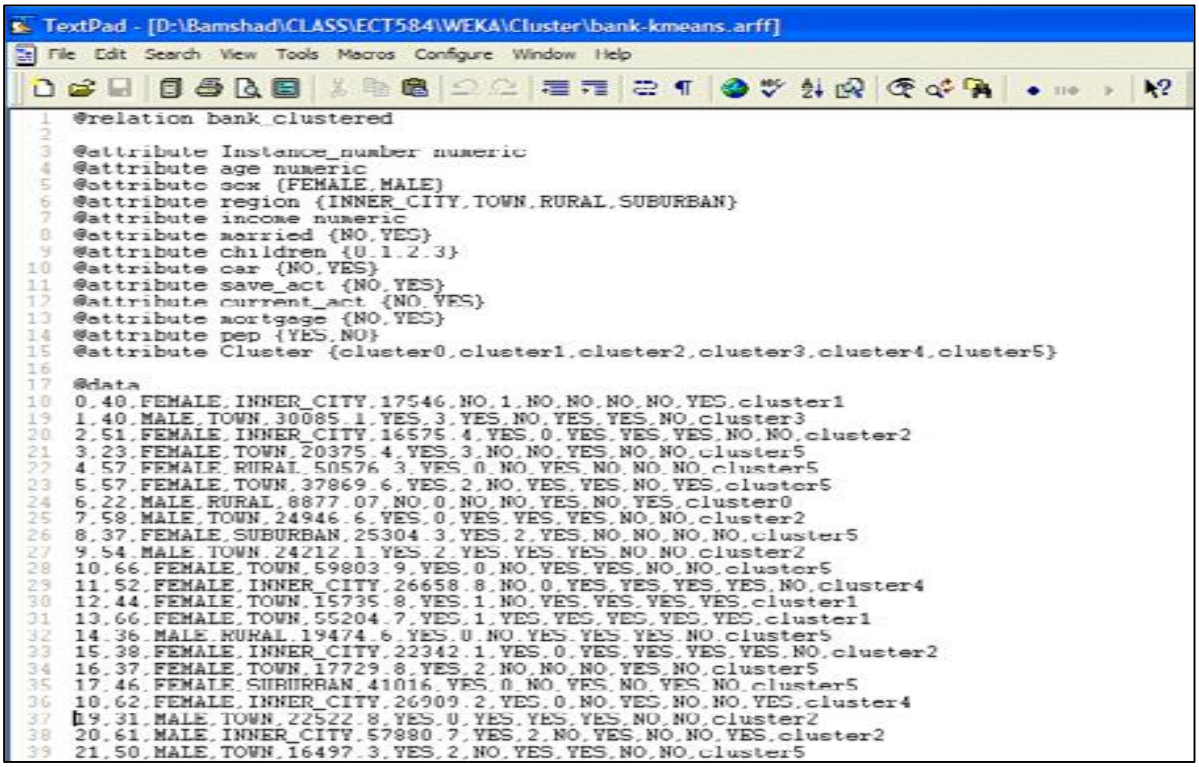

Fig. 8: The top portion of this file bank.k-means.arff

Note to in addition to the "instance_number" attribute, WEKA has also added "Cluster" attribute to the original data set. In the data section, each instance now has its assigned cluster as the last attribute value. By doing some simple management to this data set, we can easily change it to a more usable form for additional analysis or processing. For example, here we have improved this data set in a comma-separated format and sorted the result by clusters. Also, we have added the ID field from the original data set (before sorting). The results of these steps can be seen in the file “bank-kmeans.csv”.

VIII. CONCLUSION

The research compares accuracy, detection rate, false alarm rate and accuracy of other attacks under different proportion of normal information. In this, an improved approach for Intrusion Detection System (IDS) based on combining data mining and expert system is offered and implemented in WEKA. The taxonomy Consists of a classification of the detection principle as well as certain WEKA aspect of the intrusion detection system such as open-source data mining.. The result of the evaluation of the new design produced a better outcome in terms of detection efficiency and false alarm rate from the existing problems. This presents useful information in intrusion detection.

REFERENCES

[1] Su-Yun Wua, Ester Su-Yun Wua” Data mining-based intrusion detectors Crown Copyright” 2008 Published by Elsevier Ltd.
[2] G.V. Nadiammai,“ Effective approach toward Intrusion Detection System using data mining techniques”2013.
[3] Bace, Rebecca G. “ NIST special publication on intrusion detection systems”IEEE2002.
[4] Kalpana Jaswal, Seema Rawat,Praveen Kumar “Design and Development of a prototype Application for Intrusion Detection using Data mining” 2015 IEEE.
[5] S.V. Shirbhate, Dr.S.S.Sherkar, Dr.V.M.Thakare ” Performance Evaluation of PCA Filter In Clustered Based Intrusion DetectionSystem”2014.
[6] Muamer N. Mohammada, Norrozila Sulaimana, Osama Abdulkarim Muhsinb “A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment”2010.
[7] D. Patterson, F. Liu, D. Turner “Performance Comparison of the Data Reduction System. Proceedings of the SPIE Symposium on Defense and Security”2008.
[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, “The WEKA Data Mining Software: An Update” 2009.
[9] Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka
[10] B.X.Wang, D.H.Zhang, J.Wang, et al, “Application of Neural Network to Prediction of Plate Finish Cooling Temperature”, Journal of Central South University of Technology, 2008.
[11] Ian H.Witten and Elbe Frank, "Datamining Practical Machine Learning Tools and Techniques" 2005.
[12] Yang Yong” The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm” © 2011, 164 – 170
[13] Li Hanguang, Ni Yu “Intrusion Detection Technology Research Based on Apriori Algorithm”2012.
[14] Nadya EL Moussaid, Ahmed Toumanari Essi, “Overview of Intrusion Detection Using Data-Mining and the features selection”IEEE2015.