National Conference on "Advanced Technologies in Computing and Networking"-ATCON-2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

Analytical Study of Clustering Algorithms by Using Weka

Deepti V. Patange **Dr. Pradeep K. Butey** **S. E. Tayde**

***Abstract* — Emergence of modern techniques for scientific data collection has resulted in large scale accumulation of data pertaining to diverse fields. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorized it, and summarize the relationships identified. The development of data-mining applications such as clustering has shown the need for machine learning algorithms to be applied to large scale data. In this paper we present the comparison of different clustering techniques using Waikato Environment for Knowledge Analysis or in short, WEKA. The algorithm or methods tested are Density Based Clustering, EM & K-MEANS clustering algorithms.**

***Key Words* — Data mining algorithms, Density based clustering algorithm , EM algorithm, K-means algorithms, Weka tools.**

I. Introduction

Data mining is the use of automated data analysis technique to dicover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. There are many data mining techniques are available like classification, clustering, pattern recognition, and association [2]. The data mining tool gather the data, while the machine learning algorithms are used for taking decisions based on the data collected. The two main techniques of Data mining with wide applicability are Clustering. In many cases the concept of classification is confused by means of clustering, but there is difference between these two methods. According to the perspective of machine learning, clustering method is unsupervised learning and tries to group sets of objects having relationship between them [3], where as classification method is supervised and assign objects to sets of predefined classes [4].Given our goal of clustering a large data set, in this study we used k-means[5] algorithm. In our research Weka data mining tool [9] [10] was used for performing clustering techniques. The data set used in this research is of Cyber Crime Attacks, which consists of 3 attributes and 49988 instances.

II. Problem Statement

The problem in particular is a comparative study of performance of integrated clustering techniques i.e.

simple K-Means clustering algorithm integrated with different parameters of Cyber Crime Attacks dataset containing 13 attributes and reduced to only 4 attributes by using attribute selection filter, 49988 instances and one class attribute.

III. Proposed Methodology

The proposed project was implemented in 5 stages .

A. Procuring Data Set

The dataset of Cyber Crime Attacks for the current research work was downloaded from the website www. NSL.cs.ulb.ca/nsl/kdd.

B. Cleaning Data Set

A set of data items, the dataset, is a very basic concept for Data Mining. A dataset is roughly equivalent to a two-dimensional spreadsheet or database table. The dataset for crime pattern detection contained 13 attributes which were reduced to only 4 attributes by using a Java application namely no of attack, protocol, type of attack and number of times the attack happened. This structure of 4 attributes and 50000 instances or records became the final cleaned dataset for the data mining procedures.

C. Processing Data Set

Each and every organization is accession vast and amplifying amounts of data in different formats and different databases at different platforms. This data provides any meaningful information that can be used to know anything about any object. Information is nothing just data with some meaning or processed data. Information is then converted to knowledge to use with KDD.

IV. Proposed System

The data pre-processing and data mining was performed using the world famous Weka Data Mining tool. Weka is a collection of machine learning algorithms for data mining tasks. Weka is open source software for data mining under the GNU General public license. This system is developed at the University of Waikato in New Zealand. "Weka" stands for the Waikato Environment for Knowledge Analysis. Weka is freely available http://www.cs. waikato.ac.nz/ml/weka. The system is written using object oriented language Java. Weka provides implementation of state-of-the-art data mining and machine learning algorithm. User can perform association, filtering, classification, clustering, visualization, regression etc. by using Weka tool.

Fig.1Weka Tool

V. PERFORMING CLUSTERING IN WEKA

For performing cluster analysis in weka. I have loaded the data set in weka that is shown in the figure. For the weka the data set should have in the format of CSV or .ARFF file format. If the data set is not in arff format we need to be converting it.

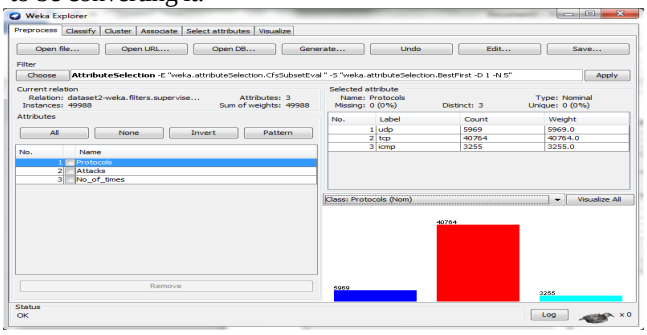

Fig. 2 Results After Attribute Selection Filter

Clustering is an unsupervised method of data mining. In clustering user needs to define their own classes according to class variables, here no predefined classes are present. In weka number of clustering algorithms are present like cobweb, Density Based Clustering, FarthestFirst, SimpleK-Means etc. K-Means is the simplest technique and gives more accurate result than others [13]. ***KMeans algorithm*:**

1) Select number of clusters to be divide.
2) Select initial value of centroid for each cluster randomly from the set of instances.
3) Object clustering
 I. Measure the Euclidean distance (Manhattan distance median value) of each object form the centroid.
 II. Place the object in the nearest cluster, whose centroid distance is minimum.
4) After placing all objects calculate the MEAN value.
5) If changes found in centroid value and newly calculated mean value.
 I. Then make the MEAN value new centroid
 II. Count the repetitions
 III. Go to step 3.
6) Else stop this process.

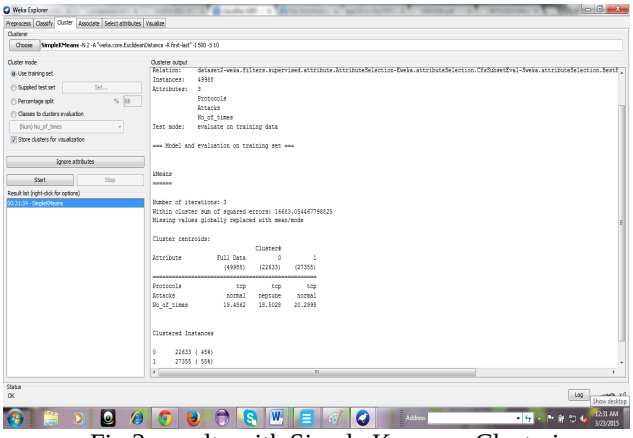

Fig.3: results with Simple K-means Clustering

From the fig.3 the results for SimpleK-means cluster are as follows:

Instances: 49988
Attributes: 3
 Protocols
 Attacks
 No_of_times
Test mode: evaluate on training data
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 16683.054467798825
Missing values globally replaced with mean/mode
Cluster centroids: Cluster#

Attribute	Full Data (49988)	0 (22633)	1 (27355)
Protocols	tcp	tcp	tcp
Attacks	normal	neptune	normal
No_of_times	19.4862	18.5028	20.2998

Clustered Instances

0 22633 (45%)
1 27355 (55%)

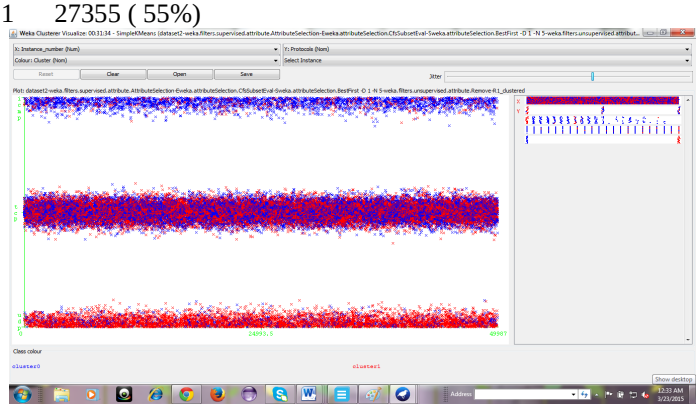

Fig. 4: Cluster Visualized on Protocols

From fig.4 we can see the clusters of three protocols groups.

Density Based Clustering Algorithm

Density Based Clustered algorithm is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996. It is a density –based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. Density Based Clustering[14] is one of the most common clustering algorithms and also most cited in scientific literature.

OPTICS can be seen as a generalization of Density Based Clustering Algorithm to multiple ranges, effectively replacing the the key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts). One of the most well known density-based clustering algorithms is the DBSCAN [15]. DBSCAN separates data points into three classes: **1)Core points:** These are points that are at the interior of a cluster. **2)Border points:** A border point is a point that is not a core point, but it falls within the neighborhood of a core point. **3)Noise points:** A noise point is any point that is not a core point or a border

point. To find a cluster, DBSCAN starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to Eps and Min Pts. The algorithm makes use of a spatial data structure(R*tree) to locate points within Eps distance from the core points of the clusters [16]. Another density based algorithm OPTICS is introduced in [17], which is an interactive clustering algorithm, works by creating an ordering of the data set representing its density-based clustering structure.

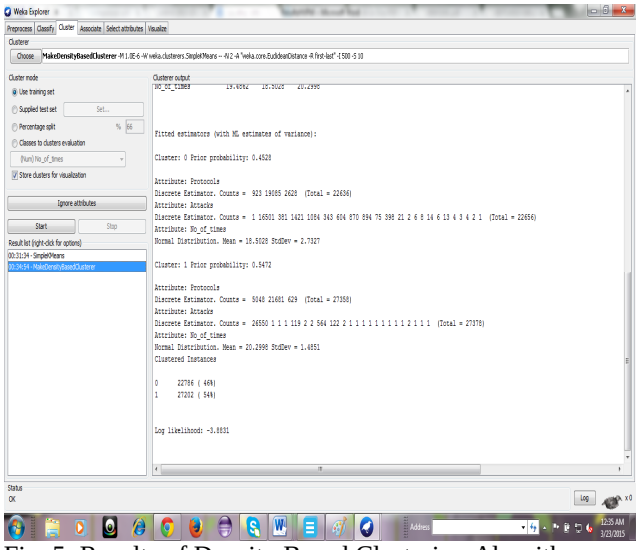

Fig. 5: Results of Density Based Clustering Algorithm

From Density Based Clustering algorithm the following results were discovered by using weka.

Instances: 49988
Attributes: 3
Protocols
Attacks
No_of_times
Test mode: evaluate on training data
MakeDensityBasedClusterer:
Wrapped clusterer:
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 16683.054467798825
Missing values globally replaced with mean/mode
Cluster centroids: Cluster#

Attribute	Full Data	0	1
	(49988)	(22633)	(27355)
Protocols	tcp	tcp	tcp
Attacks	normal	neptune	normal
No_of_times	19.4862	18.5028	20.2998

Fitted estimators (with ML estimates of variance):
Cluster: 0 Prior probability: 0.4528
Attribute: Protocols
Discrete Estimator. Counts = 923 19085 2628 (Total = 22636)
Attribute: Attacks
Discrete Estimator. Counts = 1 16501 381 1421 1084 343 604 870 894 75 398 21 2 6 8 14 6 13 4 3 4 2 1 (Total = 22656)
Attribute: No_of_times
Normal Distribution. Mean = 18.5028 StdDev = 2.7327
Cluster: 1 Prior probability: 0.5472
Attribute: Protocols
Discrete Estimator. Counts = 5048 21681 629 (Total = 27358)
Attribute: Attacks
Discrete Estimator. Counts = 26550 1 1 1 119 2 2 564 122 2 1 1 1 1 1 1 1 1 2 1 1 1 (Total = 27378)
Attribute: No_of_times
Normal Distribution. Mean = 20.2998 StdDev = 1.4851
Clustered Instances
0 22786 (46%)
1 27202 (54%)
Log likelihood: -3.8831

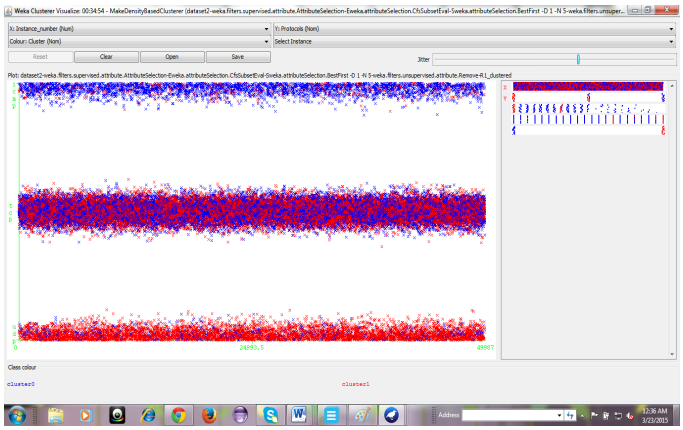

Fig. 6: Cluster visualized on Protocols by Density Based Cluster

EM Algorithm

EM algorithm [19] is also an important algorithm of data mining. We used this algorithm when we are satisfied the result of k-means methods. An Expectation–Maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM [18] iteration alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.The result of the cluster analysis is written to a band named class indices. The values in this band indicate the class indices, where a value '0' refers to the first cluster; a value of '1' refers to the second cluster, etc. The class indices are sorted according to the prior probability associated with cluster, i.e. a class index of '0' refers to the cluster with the highest probability.

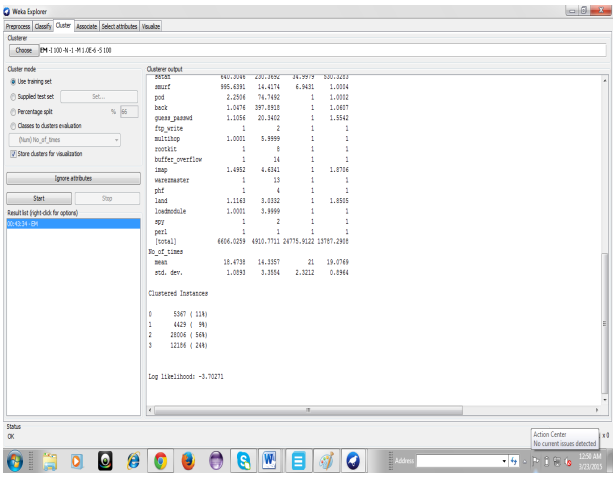

Fig. 7: Results of EM Algorithm by using weka

From the above Fig.7 the following results were discovered, they are

Instances: 49988
Attributes: 3
Protocols
Attacks
No_of_times
Test mode: evaluate on training data
Number of clusters selected by cross validation: 4

Attribute	Cluster 0 (0.13)	1 (0.1)	2 (0.5)	3 (0.28)
Protocols				
udp	3732.1914	419.6163	1819.995	1.1973
tcp	1461.2225	2686.7614	22854.9754	13765.0407
icmp	1392.6121	1784.3933	80.9417	1.0529
[total]	6586.0259	4890.7711	24755.9122	13767.2908
mean	18.4738	14.3357	21	19.0769
std. dev.	1.0893	3.3554	2.3212	0.8964

Clustered Instances
0 5367 (11%)
1 4429 (9%)
2 28006 (56%)
3 12186 (24%)
Log likelihood: -3.70271

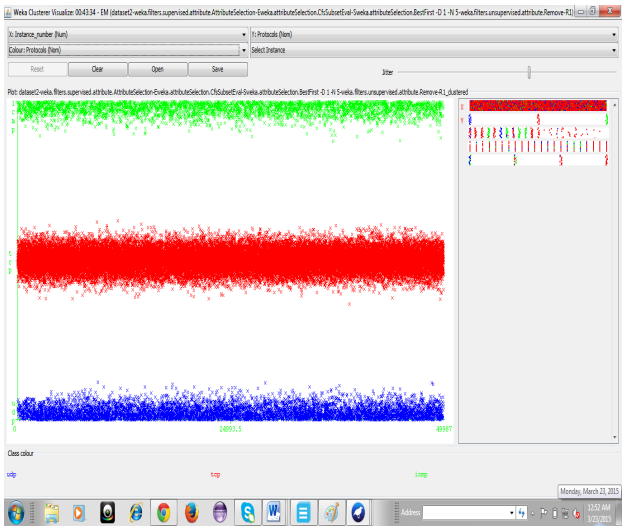

Fig.8 Cluster visualized by EM Algorithm on Protocol

VI. COMPARISON

Above section involves the study of each of the three techniques introduced previously using Weka Clustering Tool on a set of Cyber Crime data consists of 13 attributes and 50000 entries. Clustering of the data set is done with each of the clustering algorithm using Weka tool and the results are:

Table 1: Comparison result of algorithms using weka tool

Name of Cluster	***Instances***	***No. of clusters selected by cross validation***	***Log likelihood***	***Clustered Instances***				***Time to build Model***
				0	***1***	***2***	***3***	
EM	*49988*	*4*	*-3.70271*	*11%*	*9%*	*56 %*	*24 %*	*0.08*
Make Density Based Cluster	*49988*	*2*	*-3.8831*	*46%*	*54%*			*0.04*
K-means	*49988*	*2*		*45%*	*55%*			*0.02*

CONCLUSION

After analyzing the results of testing the algorithms we can obtain the following conclusions:

- The performance of K-Means algorithm is better than EM, Density Based Clustering algorithm.
- All the algorithms have some ambiguity in some (noisy) data when clustered.
- K-means algorithm is much better than EM & Density Based algorithm in time to build model.
- Similarly, in Log likelihood both EM and Density Based Clusters have negative values which does not show its perfection in results.
- Density based clustering algorithm is not suitable for data with high variance in density.
- K-Means algorithm is produces quality clusters when using huge dataset.
- Every algorithm has their own importance and we use them on the behavior of the data, but on the basis of this research we found that k-means clustering algorithm is simplest algorithm as compared to other algorithms and hence k-means is better algorithm to use on this data set.

REFERENCES

[1] Yuni Xia, Bowei Xi —Conceptual Clustering Categorical Data with Uncertainty‖ Indiana University – Purdue University Indianapolis Indianapolis, IN 46202, USA.
[2] Sanjoy Dasgupta —Performance guarantees for hierarchical clustering‖ Department of Computer Science and Engineering University of California, San Diego.
[3] A. P. Dempster; N. M. Laird; D. B. Rubin —Maximum Likelihood from Incomplete Data via the EM Algorithm‖ Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp.1-38.
[4] Slava Kisilevich, Florian Mansmann, Daniel Keim —P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos, University of Konstanz.
[5] Fei Shao, Yanjiao Cao —A New Real-time Clustering Algorithm‖ Department of Computer Science and Technology, Chongqing University of Technology Chongqing 400050, China.
[6] Jinxin Gao, David B. Hitchcock —James-Stein Shrinkage to Improve K-meansCluster Analysis‖ University of South Carolina,Department of Statistics November 30, 2009.
[7] V. Filkov and S. kiena. Integrating microarray data by consensus clustering. International Journal on Artificial Intelligence Tools, 13(4):863–880, 2004.
[8] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. In Proceedings of the thirty-seventh annual ACM Symposium on Theory of Computing, pages 684–693, 2005.
[9] E.B Fawlkes and C.L. Mallows. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 78:553–584, 1983.
[10] M. and Heckerman, D. (February, 1998). An experimental comparison of several clustering and intialization methods. Technical Report MSRTR-98-06, Microsoft Research, Redmond, WA.
[11] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. Computational statistics and data analysis, 14:315–332.
[12] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30.
[13] Microsoft academic search: most cited data mining articles: DBSCAN is on rank 24, when accessed on: 4/18/2010 .
[14] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.
[15] Z. Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery, 2:283–304, 1998.
[16] R. Ng and J. Han. "Efficient and effective clustering method for spatial data mining". In: Proceedings of the 20th VLDB Conference, pages 144-155, Santiago, Chile, 1994.
[17] E. B. Fowlkes & C. L. Mallows (1983), "A Method for Comparing Two Hierarchical Clusterings", Journal of the American Statistical Association 78, 553–569.
[18] M. and Heckerman, D. (February, 1998). An experimental comparison of several clustering and intialization methods. Technical Report MSRTR-98-06, Microsoft Research, Redmond, WA.
[19] A. P. Dempster; N. M. Laird; D. B. Rubin —Maximum Likelihood from Incomplete Data via the EM Algorithm‖ Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp.1-38.

AUTHOR'S PROFILE

Deepti V. Patange, M.Sc., M.Phil (Computer Science) and doing Ph.D. in Computer Science from SGBAU, Amravati under the guidance of Dr. P. K. Butey and working as CHB-Lecturer at Arts, Science & Commerce College, Chikhaldara, Dist. Amravati(M.S.), India from last nine years.
E-mail:deepti_pethkar@rediffmail.com

Dr. Pradeep K. Butey ,Associate Professor, Head, Department of Computer Science, Kamla Nehru Mahavidyalay, Sakkardara, Nagpur-9 and Chairman BOS, Computer Science, RTMNU, Nagpur.

S. E. Tayde , M.Sc., M.Phil (Computer Science) and doing Ph.D. in Computer Science from SGBAU, Amravati under the guidance of Dr. P. K. Butey and working as Lecturer at Department of Computer Science, S.S.S.K.R.Innani Mahavidyalaya, Karanja Lad, Dist. Washim(M.S.), India. E-mail:setayde_24@rediffmail.com