Improving the efficiency of KDD cup 1999 data by using Make Density Based Clusterer algorithm in Intrusion Detection system by removing the count attribute

[1]Pratik Jain, [2]Dimple Sukhija, [3]Saksham Agrawal
[1, 2]Assistant Professor, [3]Student
[1, 3]Department of computer science and engineering, IPS Academy, Indore, India,
[2]Indore Institute of Management & Research, DAVV, Indore, India.

Abstract—An Intrusion Detection System screens the network traffic and looks for dubious or untrustworthy movement and known intimidation on the network, and sends up the caution if it comes across such an item. Intrusion detection (ID) as a radiance remains censorial in digital security. To comprehend intrusion detection, initially realize what intrusion is? According to Heady et al., it is defined as "any action that attempts to negotiate the integrity, privacy or accessibility of a resource "for example gaining illegal access, attacking and rendering a system out of service, etc. With the end goal of this article, here it describes intrusion as any unbowed framework or merriment on (at least one) PC or set of connections of computers. This is a delineation of a legal client of a framework attempting to strengthen his advantages to acquire more prominent access to the framework that he presently depended, or a similar client attempting to interface with an unapproved remote port of a server. These are the interruptions that can incite from the rest of the world, a wronged ex-worker who was terminated recently, from your devoted staff. In this section, the average data is revealed as incursion when the case is a false positive. Here they are concentrating on this dilemma with delineation and offering one answer for a similar issue. The KDD CUP 1999 data set is used. In the consequence of this examination, it tends to be seen that if a class has a higher number of counts then this class is opined as an anomaly class. But it will be counted as an anomaly if the right individual is passing the threshold value. An elucidation is proposed to detect the true person and to get rid of fake positives.

Keywords - Ensemble, False alarm rate, K-Means, Data mining, Detection rate, Anomaly Detection System (ADS), false positive, Clustering

I. Introduction

Over the most recent twenty years, with the development of PC innovation, the wellbeing of organization framework has become a vital issue, as PC innovation has been misused by numerous individuals everywhere in the world in a few territories, this leads to network intrusion step by step over the past certain years. It is important to locate a prevailing method to monitor the information as it contains exceptionally susceptive data. Today, there is extremely endless security, for example, information encryption, VPN, and a firewall. They were acceptable inside them. Still, they have worth to utilize yet they are missing to recognize the assaults by a monstrosity. Despite this, intrusion recognition is a moveable one that can give dynamic assurance to the organization's security in invigilating assaults and slug/counter attacks. Network intrusion detection systems (NIDS) for the most part follow one of the three plan models. These regular IDS design categories are signature-based, abnormality-based, and protocol modelling. Every single plan model has its qualities and powerlessness and numerous gadgets are a combination of the three models.

Signature-Based NIDS

This is the conventional plan: generally, all NIDS gadgets have a firm reliance on signature-based detection to some degree. This innovation elucidation bundles for selective examples identified with familiar assaults. Signature-based detection is nearly advantageous to unzip, perceive, and update, and it is suitable it is appropriate at emphatically recognizing known assaults. Regardless of, it has one downside that they may not discover obscure or adjusted intrusions invasions

Categories of Intrusion Detection System

1. Signature-Based Detection Systems

A signature-based intrusion detection system (SBIDS) based on the known signature. This recognition interacts on the continual up-to-date signature as it is much emphatic indisposed known invasions. Also, it is improper to distinguish the new interruptions and book assaults, as it's overall imperfection. The solitary accommodation is that it has a magnificent identification rate than the peculiarity interruption recognition [9].

2. Anomaly Based Detection System

An anomaly-based intrusion detection system (ABIDS) has pulled numerous analysts because of its capacities of identifying novel/fiction assaults. There is a type of unrecognized assault that the AI parcel isn't cognizant of during exercise. For this, the Fiction incursion detection system of working is projected, Anomaly Based IDS has two prime advantages over Signature Based IDS the absolute initial is the capacity to recognize extraneous and "zero days" intrusion. This is finished by comparing the unassuming action with that of deviation from them. The second one is the normal movement profiles are modified for a framework, organization, and hereafter building it much firm for an aggressor to know with certainty what exercises it can remove without getting discovered[11]. The capability of the system relies on how nicely it is instrumented and tested on all protocols. The general drawback of anomaly detection is delimiting its ruleset.

3. Protocol Modeling

Protocol modeling is executed by analyzing network trouble for exceptional protocol bustling and alarming traffic with definitive deputed protocols or protocols that are mysterious to the system. Protocol modeling depends on a variety of several data sources to portray what the usual protocol activity is. Common sources for this data can include protocol specification RFCs, reasonable applications that implement that protocol, and complete analysis of normal network traffic.

II. Literature Tracery

Ugo Fiore et al, in their research discovered, when noise enhances it firstly deems the behavior of the learning method because it could alter the capability of uprooting accurate rules. Effectiveness is assessed with 3 metrics: Maximum rule confidence, Precision, and Recall [1].

D. Denning suggested an algorithm that exploits a feature detection algorithm called symbolic dynamic filtering(SDF)[2]. Francesco M, in his research says it should be concerned to focus on the employment of data mining techniques together with Embattle tree and countenance direct machines for anomaly detection As the consequence of trials shows that the calculation C4.5 has a more prominent capacity than SVM in identifying network abnormality and false alarm rate by using 1999 KDD cup data [3].In Symbolic Dynamic Filtering Algorithm, time-series data is separated for generating symbol sequences that then formulate probabilistic finite-state automata (PFSA) to focus as features for pattern taxonomy [4]. V. chandola et al used various detection frameworks and combined them to form a Hybrid detection framework, which depends on data mining taxonomy and bunching techniques [5]. T. Bhavani et al used Cluster Analysis for Abnormality Detection. K-mean clustering algorithm is used for the detection of anomalous behavior. This clustering algorithm is straightforward and deliberate. It is less PC significant than numerous different algorithms, and thusly it is a superior decision when the dataset is huge [6]. According to S. Lina et al, the definitive number of cluster given by the user are not good calculation for High dimensional dataset, as it leads to non-practical data dealing or its leads to various outlier [7]. B. Thuraisingham suggested Network intrusion detection systems that utilize signature-based methods or data mining-based methods which usually rely on labeled training data. He used Principal Component Analysis (PCA) for data reduction and Fuzzy Adaptive Resonance Theory (Fuzzy ART) for the classifier, for detecting Anomalous behavior in-network for data reduction and Fuzzy Adaptive Resonance Theory (Fuzzy ART) [8]. S. wu et al used New hybrid intrusion detection system using intelligent dynamic swarm-based rough set (IDS-IR) for feature selection and illuminated swarm optimization for intrusion data classification [9]. B. Singh et al studied through simulation and applied it to an industrial case study. The conclusion propounds viable use for decision making in production management. It is a praxis Algorithm for the building of an active network based on work order data [10]. M. Xue et al suggested hybrid views for Intrusion Detection System entrenched on data mining. The most important method is bunching analysis with aim of amended discovery rate and reduction in false alarm rate [11].

A.Samad works on the large comparative study of several anomaly detection programs for identifying different network intrusion [12]. K.Wankhade et al in his paper put side by side Anomaly traffic detection system based on the Entropy of network features and Support Vector Machine (SVM). Before that a hybrid technique that is an invasion of both entropy of network features and support vector machine is compared with individual methods [13]. J. Jonathan presented a new grid-based and density-based clustering algorithm that is expedient for unsupervised anomaly detection [14].

III. Problem Realization

The word Intrusion detection can be explained as detecting any uneven or unwanted access on a network and securing the network and guarding it. Various Intrusion detection systems are there to find if there is any intrusion or not in the network. The method for the detection of the anomalous activity is classified into two groups:-

A. Predetermined intrusion behavior

Firstly, it accumulates the outline of intrusion or the malicious actions, and then it adjudicates the intrusion according to the acquired pattern. It has advanced detection accuracy and has a small bogus alarm rate and hence finds predetermined patterns of intrusions.

B. Predetermined normal behavior

It adjudicates the usual behavior by storing the outline of the user's normal behavior into the database and detects it as an intrusion if the deviation of value is large enough from the normal value [2], [3], [4].

An Intrusion Detection System (IDS) wants to sublimate purity and detection rate as well as inferior false alarm rate. In most cases, the performance of the Intrusion Detection System is evaluated in term of Accuracy (AC), Detection Rate (DR), and false alarm rate (FAR) as in the following formula:

TABLE 1: Confusion Matrix of Intrusion Detection Data

Actual(A)\Predicted (P)	Normal Behaviour (P)	Attack(P)
Normal(A)	TN	FP
Intrusion(P)	FN	TP

- When intrusion detected as intrusion is classified as True Positive(TP)
- When Normal Behavior detected as normal is classified as True Negative(TN)
- When Normal Behavior is detected as Intrusion then it is classified as False Positive(FP)
- When Intrusion Behavior treated as Normal is classified as False Negative(FN)

(1)Accuracy = (TP+TN) / (TP+TN+FP+FN)

(2) Detection Rate = (TP) / (TP+FP)

(3) False Alarm Rate = (FP) / (FP+TN)

The problem arises when Normal data is detected as an intrusion. It is a False Positive case that arises in IDS. First an IDS needs to distinguish how the data is opined as normal or anomaly. It obtained the data of KDD Cup 1999 data. MIT Lincoln Labs arranged and dealt with the 1998 DARPA Intrusion Detection Evaluation Program. The aim of this project is to study and appraise research in intrusion detection. A typical set of data to be reviewed, which comprise a wide range of intrusions simulated in a military network environment, was made available. The version of this Dataset was used in the 1999 KDD intrusion detection contest. So after retracing the data and by comparing the usual behavior and anomalous behavior it wind up with an idea that it takes 41

attributes to ensure whether the input belongs to the normal behavior or anomalous behavior. Following are the attributes to find if it is normal or anomalous behavior (protocol_type,duration, service, flag, srv_serror_rate, dst_bytes, land, wrong_fragment, urgent, hot, num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, is_guest_login, num_access_files, num_outbound_cmds, is_host_login, , 'count', srv_count, serror_rate, src_bytes , rerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same _src_port_rate, dst_host_srv_diff_host_rate, dst_host _serror_rate, dst_host_srv_serror_rate,dst_host_rerror_rate,num_shells dst_host_srv_rerror_rate)

'class' {'normal', 'anomaly'}

Now, with the help of these 41 attributes it can be decided whether the data is standard or anomaly. For example: Consider the four data of KDD Cup 1999 dataset with attributes value as:

Example 2.1
0,udp,other,SF,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,13,1,0.00,0.00,0.00,0.00,0.08,0.15,0.00,255,1,0.00,0.60,0.88,0.00,0.00,0.00,0.00,0.00, normal

Example 2.2
0,tcp,http,SF,232,8153,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,5,5,0.20,0.20,0.00,0.00,1.00,0.00,0.00,30,255,1.00,0.00,0.03,0.04,0.03,0.01,0.00,0.01, normal

Example 2.3
0,tcp,finger,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,24,12,1.00,1.00,0.00,0.00,0.50,0.08,0.00,255,59,0.23,0.04,0.00,0.00,1.00,1.00,0.00,0.00,anomaly

Example 2.4
0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,48,16,1.00,1.00,0.00,0.00,0.14,0.06,0.00,255,15,0.06,0.07,0.00,0.00,1.00,1.00,0.00,0.00, anomaly

The IDS works on 41 attributes to tote up the anomalous behavior. Several values of each attribute have been decided based on certain formulas and if any of the input is differing from the mean value then it is considered anomalous behavior. There is a quandary of false positive in the intrusion detection system. It should discard the count attribute to solve the false positive in IDS. The problem is that there is a need to find out the attack before time.

The main concern is about attributes. As there are a total of 41 attribute in the dataset and hence it takes time to find the anomalous behavior. Hence, there is a need to improve the accuracy by decreasing the number of attributes. The key components must be kept under menstruation while rationalizing attributes. Make Density-based algorithm is one of the most significant clustering algorithms. It is one of the most significant algorithms. It is one scan algorithm that scans the raw data only once. The dense areas of the objects in data space are considered clusters, which are separated by a low-density area. For any point in a cluster, the local density around that point has to surpass some threshold. The algorithm can find randomly shaped clusters easily. It accepts 2 global parameters which are the maximum radius of the area and the least amount of point in that radius. The Flow chart is given as:

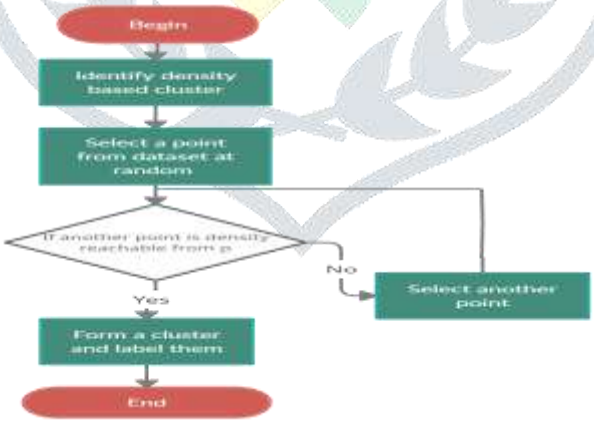

Figure 3.1 Flow chart of make density based clustering algorithm.

IV. Experiment and Results

There is an increase in the rate of false-positive due to the count attribute. To estimate the accuracy of the system the interest is in two general signs of performance: the detection rate and the false positive rate. The total number of the regular case that was (wrongly) classified as intrusions divided by the total number of normal instances can be defined as false-positive rate. The count of intrusion paradigm detected by the system split up by the total count of intrusion paradigms existent in the test set is defined by detection rate. These are nice pointers of execution since they scale which level of intrusions the system can identify and the number of wrong assortments it makes in the process. By calculating these values over the labels data performance of IDS can be measured.

So, the quandary is about authentication. It can enhance authenticity by providing an OTP or one-time password to the user's Email address or contact number. Since the count attribute is of no utilization it can eliminate the count attribute. OTP is the best solution for the existing problem.

Algorithm 1: Registration

● Begin
● Enter the required fields in the registration form which includes email, password as well as username.
● When the user didn't dill the whole form then,
Show "error message" in a dialog box
● Else
Registration Completed.
● End.

Algorithm 2: Login

1. Begin
2. Enter the values in the username and password field and fill in the CAPTCHA.
3. If the username & password entered are right then Login.
4. Else (for attempt=1 to attempt= 10)
// (Where 'attempt' is the number of attempts)
Repeat 1 and 2.
5. Generate a one-time password (OTP) & post it to the contact number or the email id of the client.
6. If OTP is right, do again 1 to 4
7. Else
Display Message "Incorrect OTP".
8. End.

Figure 1 portrays the after effect of utilizing the Make Density Based Clustered algorithm with count attribute. It can be seen that it takes 1.38 seconds to finish the clustering.

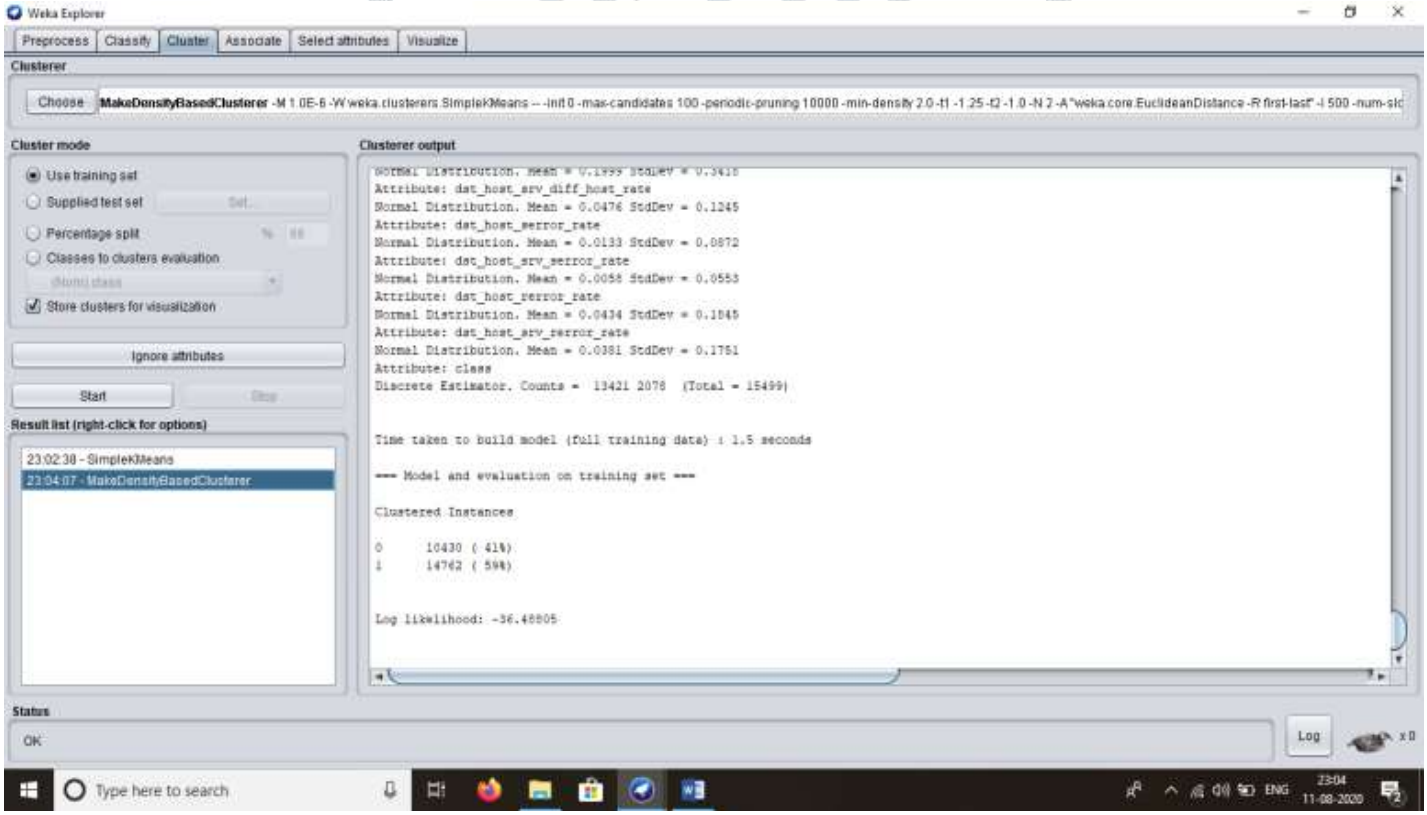

Figure 4.1 Experiment result using Make Density Based Clustered algorithm

Figure 2 portrays the after effect of utilizing the Make Density Based Clustered algorithm with count attribute. It can see that it takes 0.92 seconds to complete the clustering.

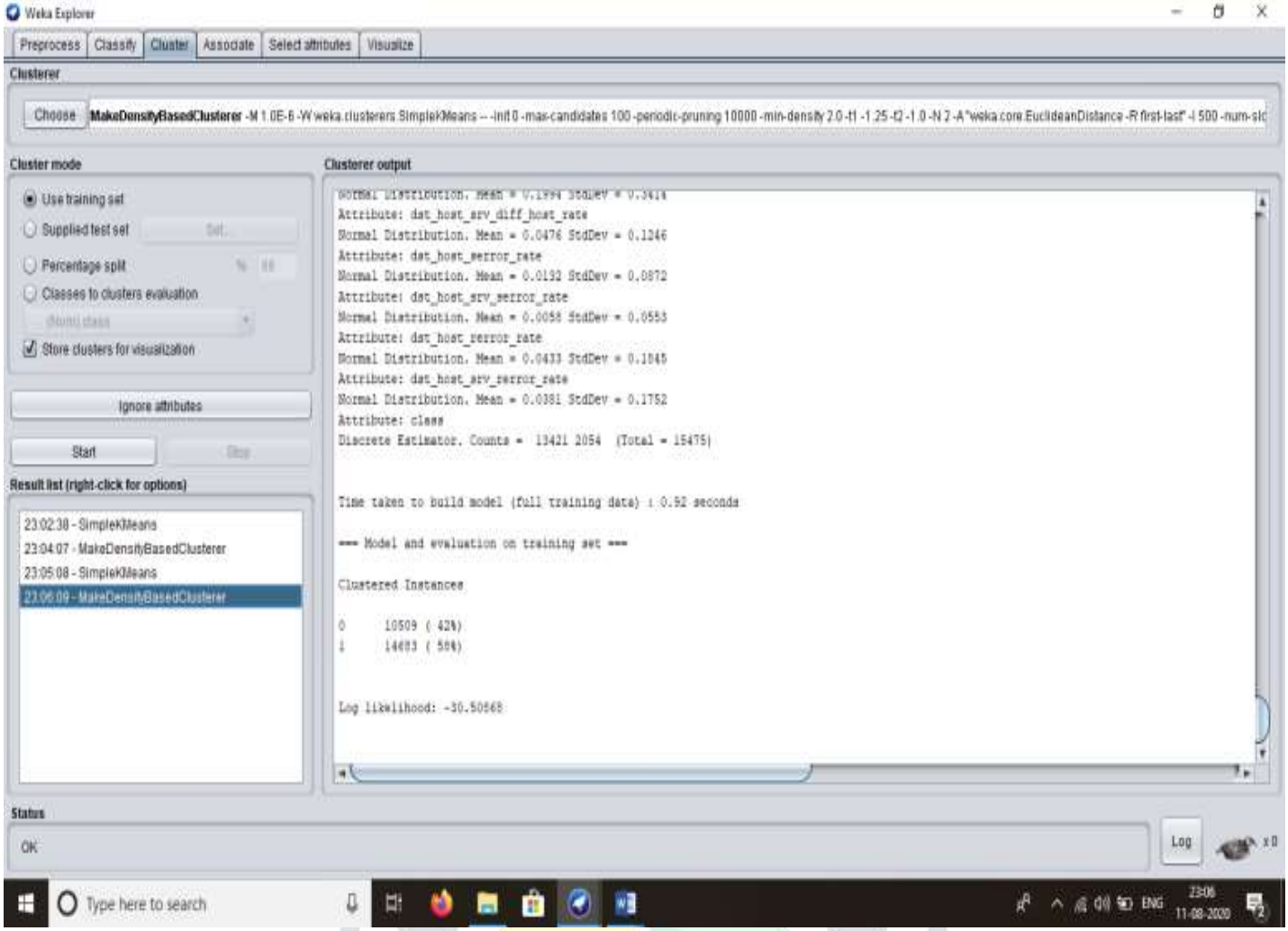

Figure 4.2 Experiment result using Make Density Based Clustered algorithm

Table 4.1:Shows the details of Final Cluster Centeroids with count attribute

Attribute	**Full Data (25192.0)**	Cluster #0 (9695.0)	1 (15497.0)
duration	305.0541	533.1584	162.3509
protocol_type	tcp	tcp	tcp
service	http	private	http
flag	SF	S0	SF
src_bytes	24330.6282	39374.1009	14919.3572
dst_bytes	3491.8472	115.1045	5606.3541
land	0	0	0
wrong_fragment	0.0237	0.0175	0.0276
urgent	0	0	0.0001
hot	0.198	0.0018	0.3208
num_failed_logins	0.0012	0.0002	0.0018
logged_in	0	0	1
num_compromised	0.2279	0	0.3704
root_shell	0.0015	0.0001	0.0025
su_attempted	0.0013	0.0002	0.0021
num_root	0.2498	0.0005	0.4058
num_file_creations	0.0147	0.001	0.0233
num_shells	0.0004	0	0.0006
num_access_files	0.0043	0	0.007
num_outbound_cmds	0	0	0
is_host_login	0	0	0

is_guest_login	0	0	0
count	84.5912	166.3895	33.4178
srv_count	27.6988	9.9234	38.8191
serror_rate	0.2863	0.7253	0.0117
srv_serror_rate	0.2838	0.7212	0.0101
rerror_rate	0.1186	0.2467	0.0385
srv_rerror_rate	0.1203	0.2486	0.04
same_srv_rate	0.6606	0.163	0.9718
diff_srv_rate	0.0624	0.1195	0.0266
srv_diff_host_rate	0.0959	0.0013	0.1551
dst_host_count	182.5321	245.2073	143.3221
dst_host_srv_count	115.063	12.5472	179.1975
dst_host_same_srv_rate	0.5198	0.0554	0.8103
dst_host_diff_srv_rate	0.0825	0.1512	0.0396
dst_host_same_src_port_rat	0.1475	0.0636	0.1999
dst_host_srv_diff_host_rate	0.0318	0.0067	0.0476
dst_host_serror_rate	0.2858	0.7215	0.0133
dst_host_srv_serror_rate	0.2798	0.7179	0.0058
dst_host_rerror_rate	0.1178	0.2368	0.0434
dst_host_srv_rerror_rate	0.1188	0.2478	0.0381
class	normal	anomaly	normal
Time taken to build model (full training data) : 1.5 seconds			
=== Model and evaluation on training set === Clustered Instances 0 10430 (41%) 1 14762 (59%) Log likelihood: -36.48805			

Table 4.2:Shows the details of Final Cluster Centeroids without count attribute

Attribute	**Full Data (25192.0)**	Cluster #0 (9695.0)	1 (15497.0)
duration	305.0541	533.1584	162.3509
protocol_type	tcp	tcp	tcp
service	http	private	http
flag	SF	S0	SF
src_bytes	24330.6282	39374.1009	14919.3572
dst_bytes	3491.8472	115.1045	5606.3541
land	0	0	0
wrong_fragment	0.0237	0.0175	0.0276
urgent	0	0	0.0001
hot	0.198	0.0018	0.3208
num_failed_logins	0.0012	0.0002	0.0018
logged_in	0	0	1
num_compromised	0.2279	0	0.3704
root_shell	0.0015	0.0001	0.0025
su_attempted	0.0013	0.0002	0.0021
num_root	0.2498	0.0005	0.4058
num_file_creations	0.0147	0.001	0.0233
num_shells	0.0004	0	0.0006
num_access_files	0.0043	0	0.007
num_outbound_cmds	0	0	0
is_host_login	0	0	0
is_guest_login	0	0	0
srv_count	27.6988	9.9234	38.8191

serror_rate	0.2863	0.7253	0.0117
srv_serror_rate	0.2838	0.7212	0.0101
rerror_rate	0.1186	0.2467	0.0385
srv_rerror_rate	0.1203	0.2486	0.04
same_srv_rate	0.6606	0.163	0.9718
diff_srv_rate	0.0624	0.1195	0.0266
srv_diff_host_rate	0.0959	0.0013	0.1551
dst_host_count	182.5321	245.2073	143.3221
dst_host_srv_count	115.063	12.5472	179.1975
dst_host_same_srv_rate	0.5198	0.0554	0.8103
dst_host_diff_srv_rate	0.0825	0.1512	0.0396
dst_host_same_src_port_rat	0.1475	0.0636	0.1999
dst_host_srv_diff_host_rate	0.0318	0.0067	0.0476
dst_host_serror_rate	0.2858	0.7215	0.0133
dst_host_srv_serror_rate	0.2798	0.7179	0.0058
dst_host_rerror_rate	0.1178	0.2368	0.0434
dst_host_srv_rerror_rate	0.1188	0.2478	0.0381
class	normal	anomaly	normal
Time taken to build model (full training data) : 0.92 seconds			
=== Model and evaluation on training set === Clustered Instances 0 10509 (42%) 1 14683 (58%) Log likelihood: -30.50868			

Table 4.3 depicts the comparison of results between Simple K-mean algorithm & Make Density Based Clustered algorithm.

Algorithm	Time taken with count attribute	Time taken without count attribute
Make Density Based clusterer	1.5 seconds	0.92 seconds

V. Conclusion

In the current scenario, countless individuals have experienced a lot of difficulties when they have to open an account online or in internet banking and remembering the password of that and also because of having more than one account it is very difficult to deal with so many passwords. In case if 3 wrong attempts are encountered then that account is blocked by that bank's website for the next 24 hours. How this particular problem can be solved is given in this paper. So if this elucidation is followed by a system the dilemma of false positives can be reduced. By getting rid of the count attribute the performance of algorithms improved to a good extent. While putting side by side the rows of table 2 it can be undoubtedly said that the efficiency and accuracy of algorithms to detect an intrusion is increased to a great extent.

REFERENCES

[1] UgoFiore , Francesco, Aniello "Network anomaly detection with the restricted Boltzmann machine" Neurocomputing 122 (2013) 13–23.

[2] Dorothy E. Denning. "An Intrusion- Detection Model" 1986 IEEE Computer Society Symposium on Research in Security and Privacy , pp 118-31.

[3] "Identification of anomalies in processes of database alteration" IEEE 2013 by Francesco Mercaldo,

[4] S. K. Chaturvedi1 , Prof. Vineet R. , Prof. Nirupama T. "Anomaly Detection in Network using Data mining Techniques" International Journal ISSN 2250-2459 Volume 2, Issue 5, May 2012.

[5] *V. Chandola, A. Banerjee, V. Kumar, "Anomaly detection as a survey" ACM Comput. Surv.41 (3)(2009)15:1–15:58.*

[6] T. Bhavani et al., "Data Mining for Security Applications," Proceedings of the 2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing Volume 02, IEEE Computer Society, 2008.

[7] "An intelligent algorithm with feature selection and decision rules applied toanomaly detection" Elsevier 2011 by Shih-Wei Lina, Kuo-Ching Yingb, Chou-Yuan Leec, Zne-Jung Leed.

[8] Bhavani Thuraisingham "Data Mining for Malicious Code Detection and Security Applications" 2009 IEEE/WIC/ACM 2009.

[9] Wang "Information-Theoretic Outlier Detection for Large-Scale Categorical Data" VOL. 25, NO. 3, MARCH 2013 by Shu Wu, Member, and Shengrui.

[10] "Exploiting Anomaly Detections for high Dimensional data using Descriptive Approach of Data mining" IEEE(ICCT) 2013 by Bharat singh,NidhiKushwaha and OP vyas.

[11] M. Xue , C. Zhu, "Applied Research on Data Mining Algorithm in Network Intrusion Detection," jcai , pp.275-277, 2009 International Joint Conference Artificial Intelligence, 2009.

[12] Abdul Samad bin Haji Ismail "A Novel Method for Unsupervised Anomaly Detection using Unlabeled Data" IEEE 2008.

[13] Kapil Wankhade, MrudulaGudadhe, Prakash Prasad, "A New Data Mining Based network Intrusion Detection Model", In Proceedings of ICCCT 2010, IEEE, 2010, pp.731-735.

[14] Jonathan J, Davis , Andrew J. Clark "Data preprocessing for anomaly based network intrusion detection: A review" Elsevier 2011.