Improving prediction of intrusion detection of network attacks by computing clustering & classification.

[#1]Niraisha M, [*2]C.Anuradha,
#1 PG Student, Department of Computer Science and Engineering.
Bharath University, Selaiyur, Chennai-600073, India
naga_sharaini@yahoo.co.in
*2 Asst. Professor,Department of Computer Science and Engineering.
Bharath University, Selaiyur, Chennai-600073, India
anuradha.ak23@gmail.com

***Abstract*-IDS which are increasingly a key part of system defense are used to identify abnormal activities in a computer system. In general, the traditional intrusion detection relies on the extensive knowledge of security experts, in particular, on their familiarity with the computer system to be protected. One of the primary challenges to intrusion detection are the problem ofmisjudgment, misdetection and lack of real time response to the attack. In this framework, intrusion detection is achieved using various data-mining techniques andsuggest that a combination of both approaches has the potential to detect intrusions in networks more effectively.**

***Keywords*: Classification, Clustering, IDS, Confusion matrix, External Quality measure**

1. INTRODUCTION

Nowadays, there exists an extensive growth in using Internet in social networking (e.g., instant messaging, video conferences, etc.), healthcare, e-commerce, bank transactions, and many other services. These Internet applications need a satisfactory level of security and privacy. On the other hand, our computers are under attacks and vulnerable to many threats. There is an increasing availability of tools and tricks for attacking and intruding networks. An intrusion can be defined as any set of actions that threaten the security requirements (e.g., integrity,
confidentiality, availability) of a computer/network resource (e.g., user accounts, file systems, and system kernels) . Intruders have promoted themselves and invented innovative tools that support various types of network attacks. Hence, effective methods for intrusion detection(ID) have become an insisting need to protect our computers from intruders. In general, there are two types of Intrusion Detection Systems (IDS); misuse detection systems and anomaly detection systems.

Classification is perhaps the most familiar and most popular data mining technique. Prediction can be thought of as classifying an attribute value into one of a set of possible classes. Clustering is similar to classification in that data are grouped. However unlike classification, the groups are not predefined. Instead, the grouping is accomplished to finding similarities between data according to characteristics found in the actual data. The groups are called clusters.

Amir Azimi[1]A New System for Clustering and Classification of Intrusion Detection System Alerts Using Self-Organizing Maps. Using Self-Organizing Map (SOM), a system is proposed to be able to classify IDS alerts and to reduce false positives alerts. Intrusion incidents to computer systems are increasing because of the commercialization of the Internet and local networks [2]. Intrusion detection has emerged as a significant field of research, because it is not theoretically possible to set up a system with no vulnerabilities [3].

One main confrontation in intrusion detection is that we have to find out the concealed attacks from a large quantity of routine communication activities [4]. Support Vector Machine [5], Fuzzy Logic [6], and Data Mining [7]

The subject is introduced briefly as following, In section 2, formulates the problem. In section 3, the experimental results and analysis. We present the conclusion in section 4.

2. DATA MINING TECHNIQUES

Two common data mining techniques for finding hidden patterns in data are clustering and classification analyses. Although classification and clustering are often mentioned in the same breath, they are different analytical approaches.

Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. This approach of segmenting the database via clustering analysis is often used as an exploratory technique because it is not necessary for the end-user/analyst to specify ahead of time how records should be related together.

There are variety of algorithms used for clustering, but they all share the property of iteratively assigning records to a cluster, calculating a measure (usually similarity, and/or distinctiveness), and re-assigning records to clusters until the calculated measures don't change much indicating that the process has converged to stable segments. Records within a cluster are more similar to each other, and more different from records that are in other clusters. Depending on the particular implementation, there are a variety of measures of similarity that are used (e.g. based on spatial distance, based on statistical variability, or even adaptations of Condorcet values used in voting schemes), but the overall goal is for the approach to converge to groups of related records.

2.1Clustering Problem

- Given a database $D=\{t_1,t_2,\ldots,t_n\}$ of tuples and an integer value k, the *Clustering Problem* is to define a mapping f:Dg{1,..,k} where each t_i is assigned to one cluster K_j, 1<=j<=k.
- A *Cluster*, K_j, contains precisely those tuples mapped to it.
- Unlike classification problem, clusters are not known a priori.

Classification is a different technique than clustering. Classification is similar to clustering in that it also segments customer records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user/analyst know ahead of time how classes are defined. For example, classes can be defined to represent the likelihood that a customer defaults on a loan (Yes/No). It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes. Because each record has a value for the attribute used to define the classes, and because the end-user decides on the attribute to use, classification is much less exploratory than clustering. The objective of a classifier is not to explore the data to discover interesting segments, but rather to decide how new records should be classified.

2.2Classification Problem

- Given a database $D=\{t_1,t_2,\ldots,t_n\}$ and a set of classes $C=\{C_1,\ldots,C_m\}$, the ***Classification Problem*** is to define a mapping f:DgC where each t_i is assigned to one class.
- Actually divides D into ***equivalence*** *classes*.
- ***Prediction*** is similar, but may be viewed as having infinite number of classes.

Confusion Matrix Example

This paper cluster and classify the network based IDSs and then compare the results.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, First we collect the dataset. Apply the dataset in weka tool to find Clustering and Classification results. Finally compare the both results. The datasets for these experiments are from nsl.cs.unb.ca/NSL-KDD network based IDSs.

3.1Dataset Information

Network based IDSs of nsl.cs. Actually 42 attributes are in dataset.2643 instances that dataset. We have to normalize the dataset using preprocessor. Attribute selection using wekatool
Evaluator:weka.attributeSelection.CfsSubsetEval
Search: weka.attributeSelection.BestFirst -D 1 -N 5.

The selected attributes are

5(5,6,12,26,29)

src_bytes

dst_bytes

logged_in

srv.serror_rate

same.srv_rate

3.2 Process

3.2.1 Constructing the confusion matrix

After normalization we have to construct the confusion matrix for the dataset using classification technique in weka tool. Sample confusion matrix for our dataset. This is build by using various classification method results. The Table[1] shows the sample confusion matrix.

TABLE 1:

No	Method	Confusion Matrix after normalize
1	bayes.BayesNet	a b <-- classified as 1366 13 | a = normal 113 1150 | b = anomaly
2	bayesNaiveBayes	a b <-- classified as 1251 128 | a = normal 138 1125 | b = anomaly

3.2.2 Calculate accuracy by classification

The accuracy should be calculated by classification. Classify the dataset using different methods of classification, we get the correctly classified and incorrectly classified details. The Table [2] shows accuracy.

TABLE 2:

	After normalize	
Method	**correctly classified**	**IN correctly classified**
bayes.BayesNet	95.23%	4.77%
bayes.NaiveBayes	89.93%	10.07%
functions.multilayerperception	98.56%	1.44%
functions.SMO	97.24%	2.76%
rules.DecisionTable	98.86%	1.14%
rules.Jrip	99.85%	0.15%
rules.ZeroR	52.20%	47.80%
trees.DecisionStump	91.82%	8.18%

3.2.3 Clustering

After classification we have to apply the different clustering methods. From the cluster we get clustered instances for each method. The Table [3] shows the clustered instances from cluster results

TABLE 3:

Cluster Method	ClusteredInstances					
	0	1	2	3	4	5
EM Cluster	8%	18%	11%	27%	24%	12%
Farthest First	40%	60%	0%	0%	0%	0%
Filtered Cluster	39%	61%	0%	0%	0%	0%
Simple Kmeans	39%	61%	0%	0%	0%	0%
Make dansit Based	40%	60%	0%	0%	0%	0%

3.2.4 Comparison of Classification and Clustering Results

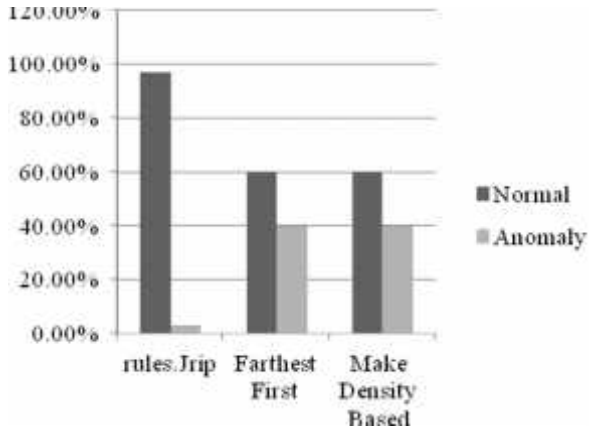
we have to compare the result of classification and clustering. We get best method from the comparison. For our dataset the classification method "rules.jrip" gives us better result. Table [4] shows the comparison.

TABLE 4:

	Method	Normal	Anomaly
classification	rules.Jrip	99.85%	0.15%
Cluster	Farthest First	60%	40%
	Make Density Based	60%	40%

The below graph shows the result.

GRAPH 1:



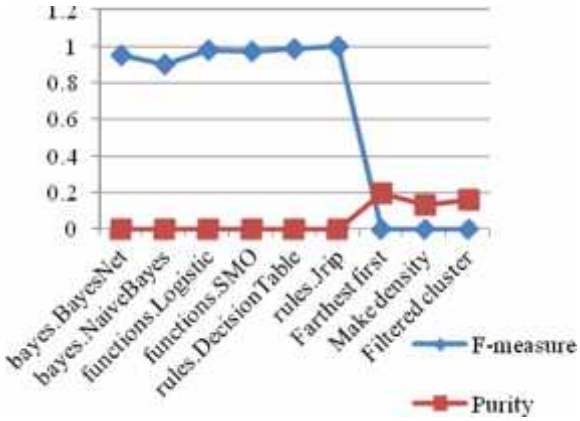
3.2.5 External quality measure

There are two common external quality measures. The first is F-measure, the second is purity. F-measure is a measure that combines the precision and recall ideas from information retrieval. We treat each cluster as if it were the result of a query and each class as if it were the desired set of documents for a query.

Purity assumes that all samples of a cluster are predicted to be members of the actual dominant class for that cluster.

The graph[2] shows the clustering and classification evaluation result.

GRAPH 2:



This method is used to measure the quality. This graph also gives us the rules.jrip method is the best method for intrusion detection in our dataset.

4. CONCLUSION

Intrusion detection systems (IDSs) play an important role in computer security. IDS users relying on the IDS to protect their computers and networks demand that an IDS provides reliable and continuous detection service. Finally Rules.JRIP method in classification is a good technique to address these problems in our data.

5.REFERENCES

[1] A New System for Clustering and Classification of Intrusion Detection System Alerts Using Self-Organizing Maps. Using Self-Organizing Map (SOM), a system is proposed to be able to classify IDS alerts and to reduce false positives alerts. Amir Azimi.

[2] Yao, J. T., S.L. Zhao, and L.V. Saxton, "A Study On Fuzzy Intrusion Detection", In Proceedings of the Data Mining, Intrusion Detection, Information Assurance, And Data Networks Security, SPIE, Vol. 5812, pp. 23-30, Orlando, Florida, USA, 2005.

[3] Susan M. Bridges and RayfordB.Vaughn, "Fuzzy Data Mining And Genetic Algorithms Applied To Intrusion Detection", In Proceedings of the National Information Systems Security Conference (NISSC), Baltimore, MD, pp.16-19, October 2000.

[4] Jian Pei, Upadhyaya, S.J., Farooq, F., Govindaraju, V, "Data mining for intrusion detection: techniques, applications and systems ", in Proceedings of the 20th International Conference on Data Engineering, pp: 877 - 87, 2004.

[5] Shon T, Seo J, and Moon J, "SVM Approach with A Genetic Algorithm for Network Intrusion Detection",Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 3733, pp. 224-233, 2005.

[6] J. Luo, and S. M. Bridges, "Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection",International Journal of Intelligent Systems, Vol. 15, No. 8, pp. 687-704, 2000.

[7] W. Lee, S. Stolfo, and K. Mok, "A Data Mining Framework for Building Intrusion Detection Model", In Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, pp. 120-132, 1999.