Network Intrusion Detection Using Dynamic Fuzzy C Means Clustering.

Richa Sampat, Shilpa Sonawani
PG Student, Maharashtra Institute of Technology,
University of Pune, India.
richasampat@gmail.com

Abstract

Internet has become a vital part of any organization. But with the growth of internet, intrusion and attacks have also increased. Thus, there arises a need of robust and powerful intrusion detection systems which can detect the attacks. Recently, many novel methods are experimented to build strong IDSs. In this paper, we implement a new Improved Dynamic FCM algorithm and successfully integrate it in WEKA to expand the system functions of the open-source platform, so that users can directly call the FCM algorithm to do fuzzy clustering analysis. Besides, considering the shortcoming of the classical FCM algorithm in selecting the initial cluster centers, we represent this Improved DFCM algorithm which adopts a new strategy to optimize the selection of original cluster centers. A novel classification via dynamic fuzzy c means clustering algorithm has been proposed to build an efficient anomaly based network intrusion detection model. A subset of KDD Cup 99 intrusion detection benchmark dataset has been used for the experiment. The proposed novel concept will be efficient in terms of detection accuracy, low false positive rate in comparison to the other existing methods.

Keywords: Network Intrusion Detection, Fuzzy Clustering, WEKA, Dynamic clustering, Improved Dynamic Fuzzy C Means (IDFCM).

I. INTRODUCTION

With the growth of network technology, nowadays more and more people are learning various ways to attacks through the network resources and carry out extremely destructive attacks. In recent years, the amount of hackers' attacks is growing 10 times per year. Therefore, network security becomes a vital factor of computer technology.

The concept of Intrusion detection system (IDS) proposed by Denning (1987) is useful to detect, identify and track the intruders. Thus, Intrusion Detection is the process of monitoring and analyzing the events occurring in the network in order to detect signs of security problems.

Intrusion Detection is an important aspect as very large amount of sensitive information is stored and processed in networked systems across the globe. IDS' modeling is a critical task and in recent years, various techniques of data mining and soft computing are being explored to strengthen network security. These approaches include neural networks, decision trees, genetic algorithms, support vector machines, Naïve Bayes classification, clustering, fuzzy logic, etc. In this paper, the focus is on fuzzy clustering techniques and its variation that is used for building a robust intrusion detection system.

The rest of the paper is organized as follows. Section 2 gives the details of intrusion detection and its types. Section 3 shows the related work done in this area. Section 4 gives details of proposed framework that uses a metaclassifier and fuzzy clustering algorithm for intrusion detection. Section 5 gives the details of different evaluation criteria on through which the performance of an intrusion detection system is measured. Section 6 shows the experimental results. And Section 7 presents the conclusion.

II. INTRUSION DETECTION

Intrusion is defined as a set of actions that attempt to compromise the integrity, confidentiality or availability of any resource on computing platform. IDS are systems that monitor the network looking for malicious or suspicious behavior in users' activity. According to the different detection methods, we can say that there are two types of intrusion detection systems.

A. Misuse detection

Misuse detection systems are the approach that tries to match user activity to signatures of known attacks that are stored in the database. Such detection systems use a prior defined knowledge to check the new activity happening on the network. It has high speed of detection and low percentage of false alarm. However, it fails in discovering the new attacks that are not defined in the database.

B. Anomaly detection

Anomaly Detection approach works on principle "anomalies are not normal". Such detection approach tries to find whether the change from the normal usage pattern can be called as intrusion or not. Thus, the anomaly detection technique stores the systems normal profile activity and raises an alarm if any abnormal behavior (i.e. intrusive activity) occurs which deviates from normal behavior. Anomaly detection helps in finding new attacks in the network.

This type of IDS can be further divided into two categories: Host-based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System (NIDS). An HIDS resides on a particular host and looks for indications of attack on that host. An NIDS resides on separate machines that look for indications of attack in the whole network.

III. RELATED WORK

In this section, a study of different fuzzy clustering techniques that is used for intrusion detection is presented.

W. Ren [1] has developed a method that applies fuzzy c-means clustering algorithm to detect network intrusion. He carries out fuzzy partition and clustering of data which separates normal data and attack data effectively. His experiment shows the feasibility and validity of fuzzy c-means clustering algorithm.

E. Narayan, *et al* [2] have proposed algorithms on expectation maximization fuzzy c-means clustering (EMFCM). Proposed algorithms provide better result to fuzzy c-means clustering by avoiding the looping problems and saves time. EMFCM clustering algorithm has fast convergence in a few iterations regardless of the initial number of clusters.

H. Wang [3] has proposed a new hybrid fuzzy clustering algorithm that uses Quantum-behaved Particle Swarm Optimization (QPSO) algorithm and combines with fuzzy c-means (FCM) for abnormally detection. This technique avoids the local minimum problems of FCM by including in the algorithm a strong global searching capacity.

F. Guorui [4] in his paper developed a semi-supervised learning algorithm for intrusion detection which is combined with the fuzzy c-Means algorithm. The KDD CUP 99 data set is adopted as the experimental subject. The result proves that the attack behaviors can be more efficiently found from the network data by the semi-supervised FCM clustering algorithm.

J. Visumathi, *et al* [5] proposed a weighted fuzzy c–means clustering based on immune genetic algorithm for intrusion detection system. It solves the high dimensionality problem in the given data set.

T. Fries [6] in his paper presents a fuzzy-genetic approach to intrusion detection that is shown to provide performance superior to other GA-based algorithms. In addition, the method demonstrates improved robustness in comparison to other GA-based techniques.

S. Chittineni, *et al* [7] proposed fuzzy c-means algorithm using neural network algorithm. The proposed work involves two steps. First, an Enhanced K-means Fast Leaning Artificial Neural Network (KFLANN) frame work is used to determine cluster centers. Secondly, fuzzy c-means uses these cluster centers to generate fuzzy membership functions.

In [8], Yu-Ping Zhou has proposed a system in which Principal Component Analysis (PCA) neural network is used to reduce the dimensions of the feature space. A modified fuzzy c-means clustering algorithm is used to cluster the learning data to obtain fuzzy rules. Also a hierarchical neurofuzzy classifier is developed. The experiments and evaluations of the proposed method were performed with the KDD Cup 99 intrusion detection dataset. Results indicate the high detection accuracy for intrusion attacks and low false alarm rate of the reliable system.

In another such paper, the authors propose a method of intrusion detection using an evolving fuzzy neural network. This type of learning algorithm combines Artificial Neural Network (ANN) and Fuzzy Inference Systems (FIS), as well as evolutionary algorithms. The algorithm uses fuzzy rules and allows new neurons to be created in order to accomplish this. They use Snort to gather data for training the algorithm and then compare their technique with that of an augmented neural network.

Li Jian-guo, *et al* [10] proposed an improved weighted fuzzy clustering algorithm based on rough set by using the methods of attributes contracted in the rough set theory to improve the FCM algorithm.

B. Thomas, *et al* [11] proposed a new fuzzy clustering method which is more efficient in handling outlier points than conventional fuzzy c-means algorithm. The new method excludes outlier points by giving them extremely small membership values in existing clusters while fuzzy c-means algorithm tends give them outsized membership values. The new algorithm also incorporates the positive aspects of k-means algorithm in calculating the new cluster centers in a more efficient approach than the c-means method.

IV. FRAMEWORK OF PROPOSED MODEL

We are developing a framework of applying data mining techniques to build intrusion detection models. This framework consists of metaclassifier and fuzzy clustering technique to iteratively do the process of constructing and evaluating detection models. The end system will be accurate and consist of intuitive classification rules that can detect intrusions, and that can be easily inspected and edited when needed.

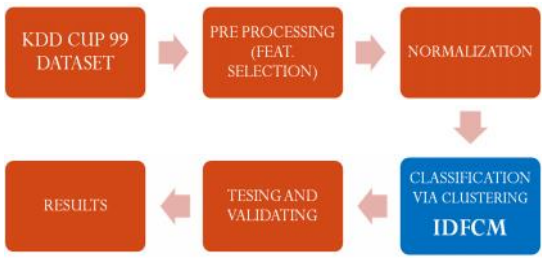

Figure 1 Proposed System Architecture

A. Dataset description

The data set used was the KDD Cup 1999 Data and it contains a wide variety of intrusions simulated in a military network environment. It consists of approximately 4,900,000 data instances, each of which is a connection record obtained from the raw network data gathered during the simulated intrusions. The attacks come under one of the following four categories: DOS - Denial of Service (e.g. a syn flood), R2L - Remote to Local - Unauthorized access from a remote machine (e.g. password guessing), U2R - User to Root - Unauthorized access to super user or root functions (e.g. a buffer overflow attack), Probing - surveillance and other probing for vulnerabilities (e.g. port scanning).

B. WEKA

WEKA (Waikato Environment for Knowledge Analysis) based on JAVA environment is a free, non-commercial and open-source platform aiming at machine learning and data mining. In WEKA, many popular data mining algorithms are implemented. However, the FCM and IDFCM algorithms are not integrated into WEKA. We intend to implement the two fuzzy algorithms in WEKA.

C. Feature Selection

Feature selection is used commonly in data mining. They are techniques used reducing inputs size for processing and analysis. Feature Selection is important because dataset contain far more information than is needed to build the model. Suppose a dataset contains 100 columns to describe the characteristics of a customer but not all columns are needed to build the model. If we keep unwanted columns while building the model, more CPU and memory are required this is an unnecessary overhead. Even if resources are not an issue, then also some noisy or redundant columns might degrade the quality of the system. Thus, feature selection helps to solve the above problems. Feature selection is done for selecting the most relevant features from the dataset. Feature selection is always done before the model is trained.

D. Normalization

Normalizes all numeric values in the given dataset (apart from the class attribute, if set). The resulting values are by default in [0,1] for the data used to compute the normalization intervals.

E. Classification via Clustering

This is a classification technique in WEKA that internally uses a clusterer for classification. Here, firstly the input data is clustered and then the meta-classifier classifies the data into different classes. Internally, any clustering technique can be used to cluster the data. In this paper, dynamic version of fuzzy c means clustering algorithm i.e. IDFCM algorithm is proposed. This IDFCM algorithm is implemented and is used as the clusterer in WEKA.

Fuzzy clustering algorithm: The main idea in fuzzy clustering is the non-unique partitioning of the data in a form of clusters. The data points are assigned membership values for each of the clusters. The fuzzy clustering algorithms allow the clusters to grow naturally. Sometimes the membership value may be zero indicating that the data point is not a member of the cluster under consideration. Many crisp (hard) clustering techniques have difficulties in handling extreme outliers but fuzzy clustering algorithms tend to give them very small membership degree in surrounding clusters. The non-zero membership values, with a maximum of one, show the degree to which the data point represents a cluster. Thus fuzzy clustering provides a flexible and robust method for handling natural data with vagueness and uncertainty. The fuzzy c-means clustering algorithm is one of the most widely used fuzzy clustering techniques.

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. *If* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *then STOP; otherwise return to step 2*

Figure 2 Fuzzy C Means Clustering Algorithm

F. Improved Dynamic Fuzzy C means algorithm

The IDFCM is a modification of the fuzzy c means algorithm, it allows the cluster centers to be adaptively updated as the data points keep streaming in. If a new cluster is formed, then a new cluster center is automatically generated. The intial steps of IDFCM algorithm is similar to the traditional FCM algorithm. To make it dynamic, cluster validity indexing is done. The working of the IDFCM algorithm is shown in the flowchart.

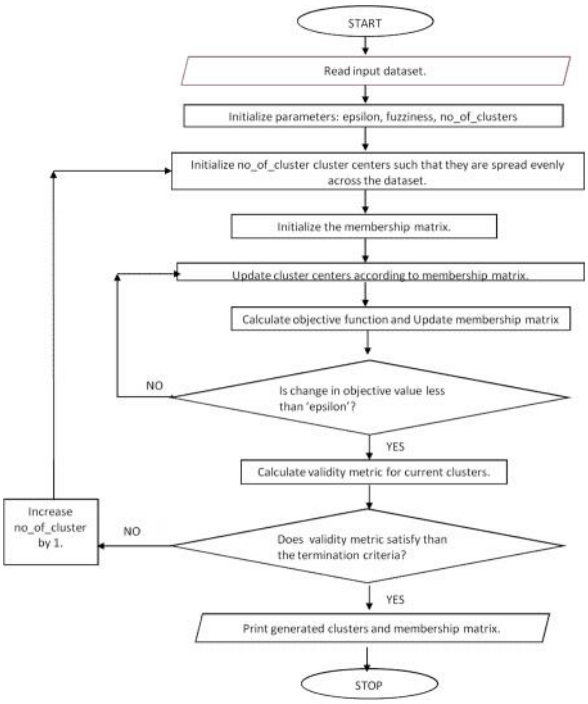

Figure 3 Flowchart of IDFCM algorithm

Validity Indices: A validity index is used to determine how well the data is represented by the selected clusters. There are a number of validity indices available. Here, we have used four widely used validity indices, namely, partition coefficient, partition entropy, FS index and Xie Beni validity index.

V. EVALUATION CRITERIA

Intrusion detection is the process of monitoring and analyzing events that occur in a computer or networked computer system to detect behavior of users that conflict with the intended use of the system. When referring to the performance of IDSs, the following terms are often used when discussing their capabilities:

True positive (TP): classifying an intrusion as an intrusion. The true positive rate is synonymous with *detection rate*, *sensitivity* and *recall*. False positive (FP): incorrectly classifying normal data as an intrusion. Also known as a *false alarm*. True negative (TN): correctly classifying normal data as normal. False negative (FN): incorrectly classifying an intrusion as normal. The performance metrics calculated from these are:

True positive rate

$$(TPR) = \frac{TP}{TP+FN} = \frac{\#\,correct\,intrusions}{\#\,intrusions}$$

False positive rate

$$(FPR) = \frac{FP}{TP+FP} = \frac{\#\,normal\,as\,intrusions}{\#\,normal}$$

VI. EXPERIMENTAL RESULTS

10 fold cross validation method – 25000 Instances						
Exp. No.		Act-ual Clu-ster	Inp-ut Clu-ster	Clu-sters gen-erated	Cor-rect clas-sified	Incor-rect classif-ied
1	Simple K Means	22	12	12	83.77	16.32
	Fuzzy C Means	22	12	12	84.59	15.1
	Improved Dynamic Fuzzy C Means	22	12	**20**	**86.99**	**12.19**
2	Simple K Means	22	12	12	83.77	16.32
	Fuzzy C Means	22	12	12	86.1	12.65
	ImprovedDynamic Fuzzy C Means	22	12	**20**	**88.16**	**11.15**
3	Simple K Means	22	12	12	83.77	16.32
	Fuzzy C Means	22	12	12	86.65	11.93
	ImprovedDynamic Fuzzy C Means	22	12	**20**	**90.45**	**8.23**

Table 6.1 Experimental results with 25k instances

Table 6.1 shows the results obtained when the experiments were done with 25000 instances. The graphs are also shown.

VII. CONCLUSION AND FUTURE WORK

The proposed method is based on the combination of feature selection and improved dynamic fuzzy C means algorithm that improves the performance results of classifiers while using a reduced set of features. It has been applied to the KDD Cup 99 dataset in the intrusion detection field. We used a normalization method on the KDD 99 training dataset and test dataset before applying the proposed scheme to the dataset. The method improves the performance results obtained by fuzzy c means clustering algorithm. Performance analysis is done by comparing the proposed method with other known methods. The proposed method gives impressive detection accuracy and detection rate in the experiment results.

Also, as the well-known open-source machine learning and data mining software, WEKA includes many java packages such as associations, classifiers, clusterers and so on. However, it doesn't implement the traditional fuzzy

clustering algorithm FCM. In our project, we successfully integrate the FCM algorithm and IDFCM algorithm into WEKA. Furthermore, we improve the traditional FCM algorithm in term of the selection strategy of initial cluster centers to fit the characteristics of intrusion data. The Improved DFCM algorithm has smaller errors than the traditional FCM algorithm while maintaining the rapid speed of convergence. Besides, the performance of the improved DFCM algorithm is better than K-means algorithm and traditional FCM algorithm.

REFERENCES

[1] Wuling Ren, *"Application of Network Intrusion Detection Based on Fuzzy C-Means Clustering Algorithm"*, Intelligent Information Technology Application, 2009.
[2] Esh Narayan, Pankaj Singh and Gaurav Kumar Tak, *"Intrusion Detection System Using Fuzzy C_ Means Clustering with Unsupervised Learning via EM Algorithms"*, VSRD-IJCSIT, Vol. 2 (6), 2012, 502-510.
[3] Hao Wang, *"Network intrusion detection based on hybrid Fuzzy C-mean clustering"*, Fuzzy Systems and Knowledge Discovery (FSKD), Seventh International IEEE Conference, 2010.
[4] Feng Guorui, *"Intrusion detection based on the semi-supervised Fuzzy C-Means clustering algorithm"*, 2nd IEEE International Conference on
Consumer Electronics, Communications and Networks (CECNet), 2012
[5] J. Visumathi, Dr. K.L.Shanmuganathan and Dr. K.A.Muhamed Junaid, *"Misuse and Anomaly-based Network Intrusion Detection System using Fuzzy and Genetic Classification Algorithms"*, International Conference on Computing and Control Engineering (ICCCE), 2012.
[6] Terrence P. Fries, *"A Fuzzy-Genetic Approach to Network Intrusion Detection"*, Department of Computer Science Coastal Carolina University Conway, South Carolina.
[7] Suneetha Chittineni, Dr. Raveendra Babu Bhogapathi, *"Neural Network Based Fuzzy C MEANS Clustering Algorithm"*, Available online at www.interscience.in
[8] Yu-Ping Zhou,*"Research on Neuro-fuzzy Inference System in Hierarchical Intrusion Detection"*, Information Technology and Computer Science (ITCS), 2009.
[9] LI Jian-guo, Gao Jing-Wei, *"Research on Improved Weighted Fuzzy Clustering Algorithm based on Rough Set"*, Proceedings of International Conference on Computer Engineering and Technology, pp.98- 102, 2009.
[10] Binu Thomas and Raju G, *"A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining"*, International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.
[11] Om H., *"A hybrid system for reducing the false alarm rate of anomaly intrusion detection system"* IEEE Conference on Recent Advances in Information Technology (RAIT), 2012.