Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 1750 – 1757

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Modeling Optimisation and Computing

A Novel Weighted Fuzzy C –Means Clustering Based on Immune Genetic Algorithm for Intrusion Detection

S.Ganapathy, K.Kulothungan, P.Yogesh, A.Kannan

Department of Information Science and Technology, Anna University, Chennai-600 025, Tamil nadu, India.

Abstract

In this paper, we propose a fuzzy clustering model to find the proper cluster structures from a dataset used for intrusion detection. Since, genetic algorithm is an effective technique to improve the classification accuracy. In this paper, we propose a Novel Weighted Fuzzy C-Means clustering method based on Immune Genetic Algorithm (IGA-NWFCM) and hence it improves the performance of the existing techniques to solve the high dimensional multi-class problems. Moreover, the probability of obtaining the global optimal value is increased by the application of immune genetic algorithm. This proposed algorithm provides a high classification accuracy, stability and probability of gaining global optimum value. The experimental results obtained from this work shows that the clustering results and the proposed algorithm provides better classification accuracy when tested with KDD'99 cup data set.

Keywords: Immune genetic algorithm; Intrusion detection; Fuzzy weighted C means algorithm;

1. Introduction

Intrusion detection systems (IDS) have been used in past research work to prevent the intrusion carried out externally and detects the unauthorized to access the information. In addition, the malicious behaviors of internal users has been monitored and detected by the intrusion detection systems. Moreover, networking devices such as routers and firewalls have been used to maintain user logs to examine the malicious activities. Therefore, the intrusion detection system aims to detect the inappropriate and malicious activities in the internet. The most common approach in the intrusion detection systems is

* Corresponding author. Tel.: 91-9488869712;
E-mail address: ganapathy.sannasi@gmail.com.

doi:10.1016/j.proeng.2012.06.213

statistical method based on anomaly detection and misuse detection. IDS have been classified into two types namely, Host based Intrusion Detection System (HIDS) and Network based Intrusion Detection System (NIDS). HIDS attempts to detect any malicious activity. NIDS monitors the entire network by residing at the server. In General, there are four types of approaches have been used for intrusion detection by various researchers [2, 3, 4, 5, 6]. Such as statistical based, Data mining based, Supervised learning based and unsupervised learning. Recently, various data mining approaches has been proposed for intrusion detection such as Classification, Clustering and Association rule mining techniques are used to create rules that can find the intrusions effectively. These techniques are used to generate rules that can detect the intrusions effectively.

Clustering are used to perform investigative data analysis technique, it attempts to partition a given data set in to dissimilar groups such that data patterns within a group are more similar to one another than those belonging to different groups[7]. Clustering techniques are classified into supervised and unsupervised methods. The unsupervised clustering method is used to detect the underlying structure in the data set for classification [8]. Supervised clustering method involved with the human interaction. The unsupervised clustering techniques are most popular due to the minimal knowledge about the dataset.

Fuzzy clustering [10] is a method to assign a member membership levels, and then using them to include the data elements to one or more clusters. The clustering is to mechanism groups the data from a large data set to a concise representation of a systems behavior clustering. In addition, a form of data compression bind to cannot a large number of samples into a small number to develop prototypes or cluster. In non-fuzzy or hard clustering data is divided into crunchy clusters where each data point belongs to exactly one cluster. In fuzzy clustering, the data points can belong to more than one cluster, and related with each of the points are membership grades which indicate the degree to which the data points belong to the various clusters. Fuzzy clustering partitions a dataset into several groups such that the similarity within a group is larger than that among groups [11]. Data classification is dividing the data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible.

Genetic Algorithms (GA) is computational models that are inspired by evolution of data's or chromosomes, the parameters are encoded into a string like structures called chromosomes. A Genetic Algorithm simply initial to with a total number of chromosomes or data's instead of hearty a randomly or through knowledge based approach. After initialize the chromosome or data is decoded and its fitness value is computed. The genetic operators, such as selection, crossover and mutation are applied probabilistically to generate a new population and their fitness values are recomputed. This process iteratively performed up to the termination criterion is attained then the best chromosome is considered as a solution to the problem concerned. Due to their effectiveness, GAs has been widely used for clustering.

In this paper, we propose a novel weighted fuzzy C- means clustering method based on immune genetic algorithm to detect intruders. In addition, the members of the clusters are classified again to check whether they are intruders or normal users. In this work, KDD Cup data set is used to perform the analysis based on clustering and classification methods. The main advantage of this proposed algorithm is that it checks both anomaly intrusions and misuse based intrusions. This work follows host based intrusion detection method and hence is effective when it is compared with the network based intrusion detection system.

2. Literature Survey

A new weighted fuzzy C-means (NW-FCM) algorithm was proposed by Chih-Cheng Hung et al [1], which is used to improve the performance of both FCM models for high dimensional multiclass pattern recognition problems. The methodology used in NW-FCM is the concept of weighted mean from the non parametric weighted feature extraction (NWFE) and cluster mean from discriminate analysis feature extraction (DAFE).

Chengjie Gu et al [3] proposed a Fuzzy Kernal K Means Clustering method based on Immune Genetic algorithm (IGA-FKKM). Dependence of fuzzy k means clustering on distribution of sample was eliminated with the introduction of kernel function in this approach. Immune genetic algorithm has been used to suppress fluctuations occurred at later evolvement and to avoid local optimum. This algorithm provides global optimum and higher cluster accuracy. Hossen J et al [4] proposed a hybrid fuzzy clustering algorithm which is the combination of fuzzy C-means algorithm and subtractive clustering. Subtractive clustering algorithm is used to increase the speed and FCM provides better accuracy.

Yanfei Zhong and Linagpei Zhang [5] proposed a fuzzy clustering algorithm based on clonal selection for land cover classification (FCSA). Traditional clustering algorithms, such as a fuzzy c means usually require a priori specification of the number of clusters and easily fall into a local optimum. FCSA has attempted to tackle the problems of FCM by use of the clonal selection algorithm to provide near optimal solutions without a priori assumptions of the number of clusters FCSA is designed as a two layer system the classification layer and the optimization layer. The classification layer to find the optimal fuzzy partition to a fixed number of classes by the immune operators of the clonal selection algorithms, clone, selection, mutation operators while in the optimization layer FCSA uses the Xie-Ben index as a measure of the validity of the corresponding partition to find the optimal number of classes.

Li Jian-guo and Guo Jing-wei [6] proposed an improved weighted fuzzy clustering algorithm based on Roughset, by using the methods of attributes contracted in the roughest theory to improve the FCM algorithm. Orinella Cominetti et al [12] proposed a fuzzy clustering algorithm, DiffFUZZY, which utilizes concepts from diffusion processes in graphs and is applicable to a large class of clustering problems than other fuzzy clustering algorithms. Saghamitra Bandyopadhyay [13] presented a short review of some different approaches of GA based clustering methods, two techniques one with fixed number of clusters and another with a variable number of fuzzy clusters.

3. Proposed Method

In this paper, we propose a novel weighted fuzzy c means clustering algorithm based on immune genetic algorithm (IGA-NWFCM) by using the Immune Genetic Algorithm [3] and New Weighted Fuzzy C Means clustering algorithm [1] which is used to optimize the various clustering problems. We propose this IGA-NWFCM algorithm for detecting the intruders in wireless network environment. The proposed IGA-NWFCM is discusses the issues, such as encoding technique, fitness function, genetic operators and immune vaccine. This algorithm encodes the cluster center c_i. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l)$ represent n cluster centers after encoding, where $l = n \times d$, and every d quantitative value from each α represents a d-dimensional cluster center. The fitness function plays a major role in algorithms convergence for performance, because it is the measure standard to evaluate an individual from group of dataset. In this paper, fitness function is defined as [1]

$$f = \frac{1}{J(x,u,v)+1} \tag{1}$$

where x = Element of a cluster, u = Member of the cluster , v = LaGrange's value .

Crossover randomly selects two individuals from parental groups as father individuals at a certain probability, and then produces two offspring individuals by replacing and recombining parts of father individuals randomly. Each offspring individual produced by crossover contains genetic materials of father individuals, but differs from them. The searching capability of IGA has been improved a lot after crossover. In addition, crossover is an important means to achieve a new generation of fine individuals. The frequency of crossover operation is determined by crossover probability P_c. Higher frequency could speed up the convergence to optimum area, but it would also lead to premature convergence if the frequency is too high. The value for P_c is ranges from 0.4 to 0.9. Selection of variation probability P_m is influenced by population size, chromosome length etc. In this paper, the value for P_c and Pm are ranges from 04 to 0.9 and 0.01 to 0.1 respectively.

The proposed work, employs an adaptive method to select immune vaccine $H = \{h_i \mid i=1,2,..,m\}$from optimal individuals genes, in the process of population evolution. Then, this injects vaccine to nβ individuals selected from group $C = (x_1,x_2,\ldots,x_n)$, where β ε (0,1). If the new individual's fitness is not as good as that of prior generation, then the individuals are chosen from prior generation, otherwise, from the new ones. One individual x_i will be selected through the roulette method from the current offspring group P(n) and placed in the next generation group P(n+1).

The roulette method is formulated in the equation 2.

$$p(x_i) = e^{f(x_i)/T_k} \;/ \sum_{i=0}^{n} e^{\frac{f(xi)}{Tk}} \tag{3}$$

where $f(x_i)$ is the fitness value of individual x_i , and T_n is a temperature variable which is gradually tends to 0.

From the above analysis, IGA-NWFCM is discussed as follows.

Novel Weighted Fuzzy C – Means clustering based on Immune Genetic Algorithm

Step 1: Initialize parameters : the number of clusters n, fuzzy degree b, termination condition E, number of individuals in one group p, crossover probability P_c, variation probability P_m, Immune vaccine probability P_v.

```
Step 2: Generate randomly group P(l) with P individuals.
Step 3: Repeat until satisfy the termination condition E,
    a. Decode each individual and calculate the center $c_i$ of each cluster.
    b. Calculate the Euclidean distance $d_H (x_i, g_i)$ between the samples with their cluster center
    c. Calculate the membership $u_{ij} (x_i)$ of each $x_i$.
    d. Calculate the fitness of each individual.
    e. Determine the optimum individual p* from parent groups, and select the immune vaccine
       $H = \{h_i | i=1,2,..,m\}$ from p*.
    f. Obtain group P'(n) by crossing and mutating P(n) with the probabilities of $P_c$ and $P_m$.
    g. Get the group P(n+1) by injecting vaccine and selecting immune.
Step 4: n = n + 1
Step 5: Get $c_i$ and the result of clustering after decoding P*.
```

4. Experimental Results

In order to evaluate *IGA-NWFCM*, the algorithm is tested on a benchmark dataset, the network traffic data from the KDD Cup 1999 Dataset [16]. KDD Cup data set is usually used as a standard dataset to evaluate the performance of clustering algorithm. Network data set includes 100 samples and can be divided into many groups. Each sample has 41 attributes two indicators, *Detection Rate* and *False Alarm Rate*, were used to measure the accuracy of the method. The *Detection Rate* shows the percentage of true intrusions that have been successfully detected. The *False Alarm Rate* is defined as the number of normal instances incorrectly labeled as intrusion divided by the total number of normal instances. A good method should provide a high *Detection Rate* together with a low *False Alarm Rate*.

The KDD dataset includes a wide variety of intrusions together with normal activities simulated in a military network environment. The simulated attacks fall in one of four major categories: DOS (denial of service), R2L (unauthorized access from remote machine), U2R (unauthorized access to local super user privilege) and Probing (surveillance and other probing). In addition, the proposed algorithm evaluates by using the recall and precision. Precision and recall defined as follows,

$$\text{Precision} = [TP / (TP+FN)] * 100 \qquad (4)$$

$$\text{Recall} = [TP / (TP+FP)] * 100 \qquad (5)$$

Table 1 shows the comparison of recall and precision values of the proposed algorithm with the existing algorithms. It is inferred that the proposed algorithm improves the performance of existing algorithms.

Table 1 Performance evaluation of the clustering algorithms

Datasets	FKM (%)		GA-FKM (%)		IGA-FKKM (%)		IGA-NWFCM (%)	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
WWW	85.77	85.14	94.58	93.27	96.35	96.98	97.43	98.12
Mail	84.32	85.63	94.26	94.01	96.27	95.93	97.71	97.23
Database	81.46	82.51	87.35	88.29	91.27	92.56	93.34	94.73
Media	80.54	82.29	86.37	84.62	91.31	90.44	92.92	93.01

Game	79.37	80.25	83.08	82.11	87.49	89.18	89.42	91.23
Average	**82.292**	**83.164**	**89.128**	**88.46**	**92.538**	**93.018**	**94.164**	**94.864**

Figure 1 shows the comparison of false alarm rate analysis between FCM and IGA-NWFCM with full features such IGA-NWFCM, reduces the false alarm rate of FCM with full features and reduced features.

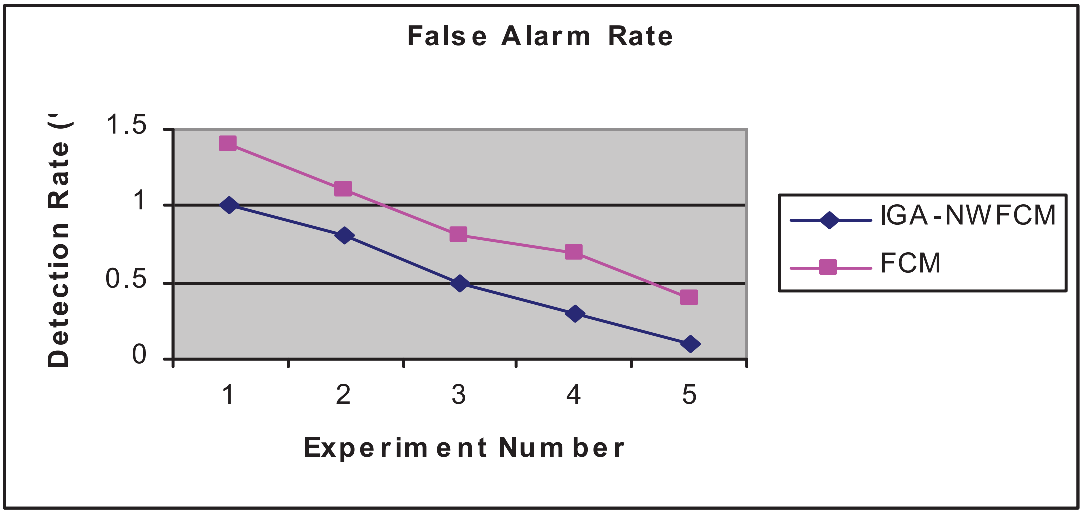

Fig. 1. False Alarm rate Analysis

Four features such as Src_bytes, Dst_bytes, Count, Srv_diff_host_rate selected by using the immune genetic algorithm and also detect the intruders based on these features by using the new weighted fuzzy c means clustering algorithm.

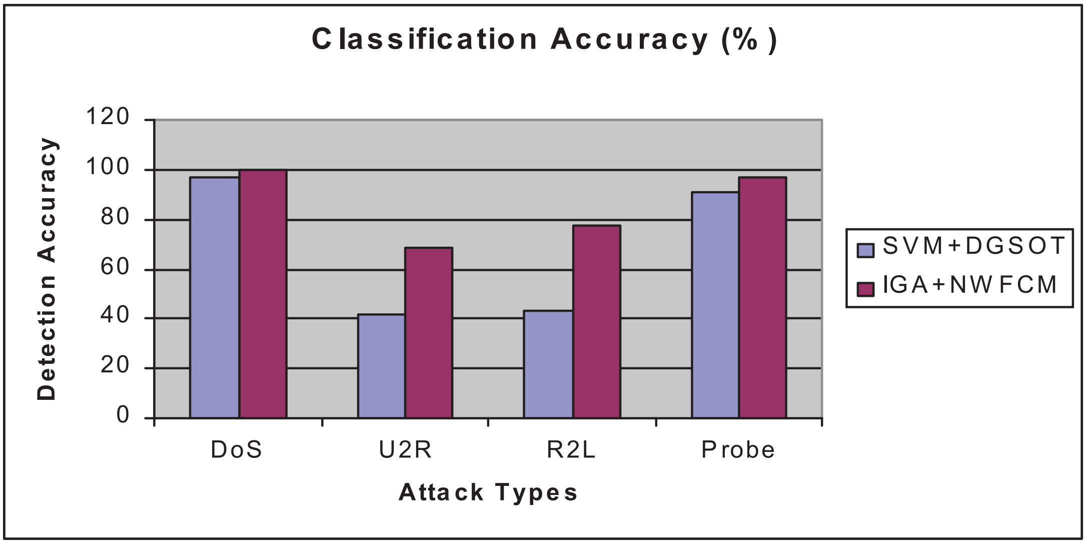

Fig. 2. Performance Analysis of Proposed Algorithm

Figure 2 shows the detection accuracy of Probe, DoS and Other attacks for the same data set. From the experiments conducted, it has been observed that the proposed IGA-NWFCM with pre-processing provides higher detection accuracy when it is compared with SVM-DGSOT. From the figure2, it can be observed that the proposed IGA-NWFCM provides better detection accuracy when it is compared with the existing methods.

5. Conclusion and Future Enhancement

In this paper, we have proposed a Novel Weighted Fuzzy C –Means Clustering Based on Immune Genetic Algorithm for intrusion detection system. The main advantage of this algorithm is that it uses clustering to identify anomaly intrusion and classification to find both anomaly and misuse. A new weighted fuzzy c means clustering module was designed to build the system more accurate for attack detection, using the immune genetic algorithm which is used to improve the performance of the network and also solved the high dimensionality problem in the given data set. The probability of gaining the global optimal value is also increased by the help of immune genetic algorithm. Compared to the existing works, increase the centroid of each cluster, stability and accuracy. The proposed algorithm provides greater classification accuracy, stability, probability and solved the high dimensionality problem of gaining the global optimum value. For future work, we have planned to improve the detection accuracy by using the intelligent agent for decision making.

6. References

[1] Chih-Cheng Hung, Sameer Kulkarni, Bor-Chen Kuo, "A New Weighted Fuzzy C-Means Clustering Algorithm for Remotely Sensed Image Classification", IEEE Journal of Selected Topics in Signal Processing, Vol.5, No.3, pp.543-553, 2011.

[2] Xiaowei Yang, Guangquan Zhang, Jie Lu, Jun Ma, "A Kernal Fuzzy C-Means Clustering-Based Fuzzy Support Vector Machine Algorithm for Classification Problems With Outliers or Noises", IEEE Transactions on Fuzzy Systems, Vol.19, No.1, pp.105-114,2011.

[3] Chengjie GU, Shunyi ZHANG, Kai LIU, He Huang, "Fuzzy Kernal K-Means Clustering Method Based on Immune Genetic Algorithm", Journal of Computational Information Systems, Vol.7, No.1,pp.221-231,2011.

[4] Hossan J, Rahman A, Sayeed S,Samsuddin K, Rokhani F, "A Modified Hybrid Fuzzy Clustering Algorithm for Data Partitions", Australian Journal of Basic and Applied Sciences, Vol.5, No.8, pp.674-681, 2011.

[5] Yanfei Zhong, Liangpei Zhang, "A New Fuzzy Clustering Algorithm Based on Clonal Selection for Land Cover Classification", Mathematical Problems in Engineering, 2011.

[6] 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.

[7] Xuanli L.X and Gerardo B, "A Validity Measure for Fuzzy Clustering", IEEE Transactions Pattern Analysis Mach. Intell., Vol. 13, No.8, pp.841-847, 1991.

[8] Balasko B, Abonyi J, Feil B, "Fuzzy Clustering and Data Analysis Toolbox for Use with Matlab". [Online] Available: http://www.fmt.vein.hu/softcomp/.Zadeh L A, "Fuzzy sets", *Information Control, Vol.8*, pp.338-53, 1965.

[9] Zadeh L A, "Fuzzy sets", *Information Control, Vol.8*, pp.338-53, 1965.

[10] Chen W.J, Giger M.L, Bick U, "A Fuzzy C-Means (FCM)-based Approach for Computerized Segmentation of Breast Lesions in Dynamic Contrast Enhanced MRI Images", Academic Radial, Vol.13, No. 1, pp.63-72, 2006.

[11] Jang J.S, Sun C.T, Mizutani, "Neuro-Fuzzy and Soft Computing – A Computational Approach to Learning and Machine Intelligence", Prentice Hall, 1997.

[12] Orinella Cominetti, Anastasios Matzavinos, Sandhya Samarasinghe, Don Kulasiri, Sijia Liu, Philip K. Maini, Radek Erban, "DifFUZZY: a fuzzy clustering algorithm for complex datasets", International Journal of Computational Intelligence in Bioinformatics and Systems Biology, Vol.1, No.4, 2010.

[13] Saghamitra Bandyopadhyay, "Genetic algorithms for clustering and fuzzy clustering", Vol. 1, No.6, pp. 524-531, 2011.

[14] Wang X, "On the Gradient Inverse Weighted Filter", IEEE Transactions on Signal Processing, Vol.40, No.2, pp.482-484, 1982.

[15] Dai Yongshhoua, Li Yuanyuan, Wei Lei, Wang Junling, Zheng Deling, "Adaptive Immune Genetic Algorithm for Global Optimization to Multivariable Function", Journal of Systems Engineering and Electronics, Vol.18, No.3, pp.655-660, 2007.

[16] KDD Cup 1999 Data, Information and Computer Science, University of California, Irvine. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.