Journal of Engineering and Applied Sciences 15 (5): 1170-1174, 2020
ISSN: 1816-949X

Intrusion Detection Systems Data Classification by Possibilistic C-Means Method

Aini Suri Talita and Eri Prasetyo Wibowo
Faculty of Computer Science and Information Technology, Gunadarma University,
Jalan Margonda Raya 100, Depok, 16424 Jawa Barat, Indonesia

Abstract: Internet Detection System (IDS) can be used to detect a malicious attempt on the network or system aims to access restricted information or exploit internal resources, monitors and analyze user activity and maintain data integrity. Based on its detection mechanism, IDS can be grouped into offline and real-time IDS. On offline IDS, a saved labeled data set as in KDD Cup'99 data set are used to measure the fitness factor of the rules on the identifier and we can analyze it to prevent some attacks happen in the future. Some classification methods have been widely used to classify IDS data set. It was used to recognize the pattern of the attacks so that we can differ between normal and unusual behaviors. On this research, we use Possibillistic C-Means (PCM) method as a classifier for KDD Cup'99 data set. Based on the experiment, the best classification results was reach on 13% training data set with accuracy 68,63%. The accuracy is still low since PCM use several values of parameters and it affects the algorithm performances when the chosen values are not the best ones.

Key words: Classification, intrusion detection system, possilibilistic C-means, internal resources, integrity, parameters

INTRODUCTION

The number of intrusion events has grown, especially in this internet of things era. Security becomes a crucial aspect to reduce the risks relating to confidentiality, availability, integrity and non-repudiation. The passive security method such as firewall is not enough anymore against rapid attacks, since, firewall enforces which traffic that allowed in and out of network based on the rules that defined by inspecting the headers but not the contents of the data packets. We need a more active line of defence, Intrusion Detection System (IDS). Intrusion detection system is a software application or tools to monitor the network in order to detect and alert the network administrator if there exist an attempt of malicious behavior from the inside or outside the network. Anderson is the first to introduce IDS in 1980's (Anderson, 1980). IDS can be classified by several factors (Pharate *et al*., 2015). Based on its location, IDS was classified into host-based and network based IDS. Host based network intrusion monitor a system or a computer from internal or external attack, it is a passive system that have to wait for an event as an indication of an attack and cannot proactively prevent it. On network IDS, network and passing traffic were analyzed to detect an attack.

If the system find something identify as an attack, the network administrator will be alerted. It can identify four major type of attacks: DoS (Denial of Service), probe, user to root and remote to local. Based on the detection mechanism, IDS can be classified into: rule engine and artificial intelligence.

On artificial intelligence based IDS, there are two types of methods, supervised and unsupervised learning based methods. If we have the labeled training dataset, then we are able to use the supervised learning one. The training dataset are used to determine the pattern of the attack while the output values are used to calculate the fitness factor for the performance of the identification method. While based on its detection mechanism, IDS can be classified into real time and offline IDS. On offline IDS usually we use saved attack data set as in KDD Cup'99 data set (Hettich and Bay, 1999). Offline IDS help in understanding the attack mechanism and help repairing the damaged cause by a certain type of attack.

The effectiveness of IDS is crucial on cloud environment, since, it can be used to increase the security level of cloud environments. Elsayed and Zulkernine presents novel classification scheme of the state-of-the art of intrusion detection approaches in the cloud environments (Elsayed and Zulkernine, 2015). Their analysis can be used to determine the proper architecture and detection method related to implementation of IDS on cloud environment. While (Chou *et al*., 2008) use decision tree (C4.5) and Naive Bayes method on classifying IDS data set.

Corresponding Author: Aini Suri Talita, Faculty of Computer Science and Information Technology, Gunadarma University, Jalan Margonda Raya 100, Depok, 16424 Jawa Barat, Indonesia

There are several approaches on solving IDS problems. One of the approaches is classification (data mining) approaches. On this approach, we use classification method to recognize the pattern of normal behaviors and attack, so that, the system can differ between malicious behavior and normal ones. Classification technique can classifying data $X = \{x_1, x_2, ..., x_n\}$, $x_i \in R^n$, $i = 1, 2, ..., n$ into several classes. Data with the same characteristics was placed on the same place. Vector Quantization (VQ) is one of the basic principal on classification.

On VQ, each class was represented by a prototype (cluster center), $v_j \in R^d$ often called as medoid, signatureaz or codebook. After we set the initial prototype, we determine the optimization models, consist of membership functions set and prototype set. To solve this type of problem, we can use Alternating Minimization Algorithm (AMA) (Bezdek, 1981) where to find the optimal solution we iteratively updating membership functions set and prototype set by certain formulas.

Some classififcation methods that use VQ and AMA principals are Fuzzy C-Mean (FCM) (Bezdek, 1981), Possibilitic C-Mean (PCM) (Krishnapuram and Keller, 1996), dan generalized fuzzy C-mean (Karayiannis, 2000). One of the fuzzy based classification method that has been used for classifying IDS data set is Fuzzy C-Means (FCM). By using FCM, there is still a possibility that the final membership values do not representing the value of probability that the data belong to some classes. On this research, we try to implement possibilistic C-means, another fuzzy based method on IDS KDD Cup'99 data set.

Fuzzy C-means is one of the most classic and powerful method. But it has several drawbacks, one of them is it does not work well with noise environment. Possibilistic C-Means (PCM) was first introduced by Krishnapuram and Keller (Krishnapuram and Keller, 1993) to overcome this problem. But it also has its own drawbacks, specifically, related with its dependence on parameter and initialization choices. On (Jing, 2009), the combination of Particle Swarm Optimization (PSO) and FCM was used on image segmentation. PSO have the merits of global optimization problem, it is hard to reach local extrema. Local extrema, on some cases lead to the failure on finding global extrema values.

On (Simhachalam and Ganesan, 2014), PCM and FCM are compared on thyroid dataset classification problem. The dataset contains 215 samples and 3 classes, hypothyroid (30 samples), hyperthyroid (35 samples) and normal data (150 samples). Each vector has 5 features, T3- percentage of resin uptake test, total amount of serum thyroxin, total amount of serum tridothyronine, Basal Thyroid-Stimulating Hormone (TSH) and after injection of 200 μg of thyrotropin-releasing hormone maximal absolute difference of TSH value as compared to the basal value. The accuracy of FCM based classifier is 66.97% where from 215 testing set, 144 are correctly classified. While the accuracy of PCM based classifier is 78.13% where 168 data are correctly classified.

Possibilistic C-means and fuzzy C-means are not able to handle uncertainty in dataset. To overcome this problem, we can use the extended version by using type-2 fuzzy logic technique (Karnik and Mendel, 2001; Mendel, 2001). The extension of PCM by using type-2 fuzzy logic technique was given on (Rubio *et al*., 2015) that was performed to describe and deal with uncertainty on the dataset. It includes a secondary membership function to model the uncertainty on the classic PCM. The extension version, IT2PCM was tested on Iris dataset by comparing with IT2FCM performance based on the norm calculated between the real centers of the Iris flower dataset and the defuzzification of the centers found by ITFCM and IT2PCM.

On (Xenaki *et al*., 2016), Sparce Possibilitics C-Means (SPCM), a novel possibilistic clustering algorithm was proposed by introducing sparsity. SPCM use a sparsity constraint that was imposed on the degree of compatibility vectors such that each data is compatible with only a few or even none clusters. On (Koutroumbas *et al*., 2017), the convergence of SPCM was analyzed. One of the result is that SPCM will converge to one of the local minima of the cost function.

On (Kumar and Mathur, 2014), DenOD (Density Based Outlier Detection) was proposed as an efficient outlier detection concept. It was based on unsupervised method and it will implement on IDCC (Intrusion Detection in Cloud Computing) that has 3 components cloud nodes, IDS and end user.

NSL-KDD dataset derived from KDD CUP '99 dataset was used on (Gong *et al*., 2014) that propose a multi-agent intrusion detection architecture along with a feature selection approach on Industrial Control System (ICS) problem. The dataset has several advantages as in: the training set does not include redundant records and its attacks are similar with attacks that ICS faced. Gong *et al*. (2014) deploy the architecture on the second level of ANSI/ISA-99 reference model supervisory control level where the experiment results show that Bayesian network classifier higher classification accuracy by the proposed method. It also improves true positive rate and reduce false positive rate on DoS, probe and R2L attacks on NSL-KDD dataset.

MATERIALS AND METHODS

Intrusion Detection System (IDS) offer additional security measures for networks. It can monitor the networks for some suspicious or unusual behavior. It can also leverage the detection of an attack from external parties to exploiting networks vulnerability or internal

parties that plan to violating security policies. On this research, we evaluate PCM method on KDD Cup'99 data set that contains 4 types of attacks. The first one is Denial of Service attacks (DoS) where attacker attack the system such that legitimate user cannot access the system either by ensuring that there is no memory left to be used or too busy traffic on the network. Some examples on this type of attacks are Mail Bomb, SYN flood and smurf. On probe, attacker scans the network to detect if there exist vulnerability on the network such as on Nmap and Mscan. The third type is User to Root (U2R) where intruder use a valid account of normal user to find the weakness on the system in order to get into the system root. Eject, Ffbconfig and Ps are some examples of U2R attacks. And on Remote to Local attacks (R2L), intruders send some packets into the system through the network to find the system weakness, so that, they can act as a local user as in Guest, IMAP and Dictionary. There are 41 features in total on the data set, some examples is num_failed_logins, duration and protocol_type. But on this research, we only consider 38 real number features.

We can see 41 features of KDD'99 dataset on Table 1. With features no 42 is the its class, normal or one of the attacks. Each KDD'99 data consists of 41 features that given at Table 1.

On this study, we implement possibilistic C-means method that is a modification of FCM to classify IDS dataset KDD Cup' 99. Before we explain PCM in details, we will state fuzzy C-means method that was found by Bezdek (6 di ids jatit). If we have a set consist of m d-dimensional vector, $X = \{x_1, x_2, \ldots, x_m\}$ we define membership matrix $U = [u_{ij}]$, $1 \le i \le n$, $1 \le j \le c$. The entry on the matrix can be seen as a representation of the possibilities of ith-data to be on jth-class. Define a set $V = \{v_1, v_2, \ldots, v_c\}$ as an initial cluster center set. We can choose random vectors as the initial cluster center or another approaches as in choose "the average" vector that we know are in the class. On Eq. 1, we can see the objective function of FCM. The goal is to minimize the function.

$$F(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}(u_{ij})^{m}d^{2}(x_i,v_j) \quad (1)$$

With constraints for i = 1, 2, ... , n, j = 1, 2, ... c and d as the distance or dissimilarity function while m is the fuzziness degree for cluster partition with its value can be any real number more than equal to 1.

$$\begin{aligned}&\sum_{j=1}^{c}u_{ij}=1,\\&\sum_{i=1}^{n}u_{ij}>0,\\&u_{ij}\in[0,1]\end{aligned} \quad (2)$$

We use Eq. 3 and 4 to update cluster center and membership values on FCM.

Table 1: KDD99 data features

No.	Feature name
1	duration
2	protocol_type
3	Service
4	flag
5	src_bytes
6	dst_bytes
7	land
8	wrong_fragment
9	urgent
10	hot
11	num_failed_logins
12	logged_in
13	num_compromised
14	root_shell
15	su_attemted
16	num_root
17	nu_file_creations
18	num_shells
19	num_access_file
20	num_outbond_cmds
21	is_host_login
22	is_guest_login
23	count
24	srv_count
25	serror_rate
26	srv_serror_rate
27	rerror_rate
28	srv_rerror_rate
29	same_srv_rate
30	diff_srv_rate
31	srv_diff_host_rate
32	dst_host_count
33	dst_host_srv_count
34	dst_host_same_srv_rate
35	dst_host_diff_srv_rate
36	dst_host_same_src_port_rate
37	dst_host_srv_diff_host_rate
38	dst_host_serror_rate
39	dst_host_srv_serror_rate
40	dst_host_rerror_rate
41	dst_host_srv_rerror_rate
42	attack_type

$$v_j=\frac{\sum_{i=1}^{n}u_{ij}^{m}xi}{\sum_{i=1}^{n}u_{ij}^{m}},\ j=1,2,\ldots,c \quad (3)$$

And:

$$u_{ij}=\left(\sum_{j=1}^{c}\left(\frac{d(x_i,v_i^t)}{d(x_i,v_j^t)}\right)^{\frac{2}{m-1}}\right)^{-1},\ 1\le i\le n \quad (4)$$

We update the membership matrix and the cluster center until some stopping criteria. It can be maximum number of iteration or while the cluster center already converge.

RESULTS AND DISCUSSION

Fuzzy based classification method has been widely used to solve classification problem related to IDS data set. One of the widely known example is Fuzzy C-Means

(FCM) method. FCM use a probabilistic boundary values where the total membership values on each data for all classes is 1 as in Eq. 6. The boundary values were used to build the membership function by implementing an iterative algorithm. The membership values are not always related with intuitive concept of "similarity" or "possibility" of a certain data belongs to a certain class. On (Krishnapuram and Keller, 1993), proposed a possibilistic based method, Possibilistic C-Means (PCM) that applying probabilistic approaches in the sense that the membership values obtained by the algorithm reflects "similarity" or "possibility" of a certain data belongs to a certain class or not. The objective function on FCM was being modified to derive a new membership function for PCM. If denoted fuzzy partition matrix build on FCM, the for all i, j, entry u_{ij} of U satisfying:

$$u_{ij} \in [0, 1] \tag{5}$$

Where:

$$0 < \sum_{j=1}^{N} u_{ij} < N, \ \forall i, \ \sum_{j=1}^{c} u_{ij} = 1, \ \forall j \tag{6}$$

The values u_{ij} on Eq. 5 represent membership value of data x_j on class β_i, c is the number of classes and N is the number of feature data. Here after, notation β_i is going to be used to represent i-th class and its cluster center. Equation (7) gives the objective function of FCM that was going to minimized where the constraint was given on Eq. 8.

$$J(L,U) = \sum_{i=1}^{c} \sum_{j=1}^{N} (u_{ij})^m d_{ij}^2 \tag{7}$$

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \tag{8}$$

Where: $L = (\beta_1, \beta_2, ..., \beta_c)$ is a c-tuple of prototype as the distance from the j-th data to the i-th cluster center. On this case, u_{ij} is membership value of x_j on cluster β_j and m as the fuzziness degree, $m \in [1, \infty]$.

By simplifying constraints on Eq. 8 results in trivial solution that is minimizing criteria function by setting 0 values on all membership. The membership value for class representation data was expected to be as large as possible and as small as possible for those are not representing data. The new objective function given by Eq. 9:

$$J_m(L,U) \sum_{i=1}^{c} \sum_{j=1}^{N} (u_{ij})^m d_{ij}^2 + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{N} (1-u_{ij})^m \tag{9}$$

Where η_i computed by Eq. (10).

$$\eta_i = K \frac{\sum_{j=1}^{N} u_{ij}^m d_{ij}^2}{\sum_{j=1}^{N} u_{ij}^m} \tag{10}$$

Table 2: Classification results by possibilistic C-means

Training data	Testing data	Accuracy (%)
40.000	40.000	55.19
40.000	102.108	62.57
40.000	160.000	63.38
40.000	199.910	67.04
40.000	240.000	67.77
40.000	300.000	68.63
40.000	340.000	67.04

To get the global minimum value of J_m (L, U) the updating membership value J_m (L, U) function must be on the form as in Eq. (11):

$$u_{ij} = \left(1 + \left(\frac{d_{ij}^2}{\eta_i}\right)^{1/(m-1)}\right)^{-1} \tag{11}$$

So that, on each iteration, the updated value u_{ij} dependant only on distance between between x_j and β_i that intuitively gives a better results. In the sense of similarity with cluster center, the membership value of an object should be determined only by its distance with cluster centers without considering its location relative with other classes. With this formulation, it is possible that the whole optimal membership solution lie on a unit hyper cube, not constricted on a hyper plane, represented by constrain on Eq. 8. For the cluster centers, it was updated by the formula on Eq. (12):

$$\frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m}, \text{ for } i = 1, 2, ..., c \tag{12}$$

On this research we use possibilistic C-means method. It was implemented on Software MATLAB with KDD Cup'99 (Hettich and Bay, 1999) the input data set. The experiment was held on personal computer with i-7, 1 TB hard disk and 16GB RAM. For the parameters value we use m = 5, K = 12 and iteration number T = 50 as the stopping criterion. The classification results was given on Table 2.

As we can see on Table 2, the highest accuracy 68, 63% was achieved on the 6th row with training data approximately 13% of the testing data but in general its accuracy is not high enough. It is possibly caused by the property of PCM algorithm that needs several values of parameters, so that, it is a bit difficult to find the most "appropriate" values of parameters to have a good classification results. Another possibility is that some of the 38 features on the data set are not representing the data well enough. Its redundancy may cause a low accuracy on the algorithm performance. And based on the experiment, we can see that even though we increase the number of the training data, it does not always lead to the increasing of the accuracy.

CONCLUSION

In this study, we implemented PCM algorithm on KDD Cup'99 data set. The classification accuracy still quite low since PCM needs several parameters values that affect the accuracy. On this study with parameters m = 5, K = 12 and iteration number T = 50, the accuracy reaches 68, 63% for 13% of training data.

RECOMMENDATION

For future works, we can use PCM with different values of parameters to reach a better accuracy or we use another classification method that do not need parameters values or robust against choices of parameters values. Some feature selection methods can be applied to find the best feature set that represent the data.

ACKNOWLEDGEMENTS

This research is sponsored by the grant Kemenristek/Dikti Republik Indonesia skema Penelitian Terapan Unggulan Perguruan Tinggi by 7/E/KPT/2019, contract No. 06a.8/LP/UG/III/2019.

REFERENCES

Anderson, J.P., 1980. Computer security threat monitoring and surveillance. Technical Report, James P. Anderson Company, Fort Washington, PA., USA., February 26, 1980.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, USA., Pages: 256.

Chou, T.S., K.K. Yen and J. Lou, 2008. Network intrusion detection design using feature selection of soft computing paradigms. Int. J. Comput. Intell., 4: 196-200.

Elsayed, M. and M. Zulkernine, 2015. A classification of intrusion detection systems in the cloud. J. Inf. Process., 23: 392-401.

Gong, Y., Y. Fang, L. Liu and J. Li, 2014. Multi-agent intrusion detection system using feature selection approach. Proceedings of the 2014 10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, August 27-29, 2014, IEEE, Kitakyushu, Japan, pp: 528-531.

Hettich, S. and S.D. Bay, 1999. The UCI KDD archive [http://kdd.ics.uci.edu]. University of California, Department of Information and Computer Science, Irvine, CA.

Jing, Z., 2009. Image segmentation using possibilistic C means based on particle swarm optimization. Proceedings of the 2009 WRI Global International Congress on Intelligent Systems, May 19-21, 2009, IEEE, Xiamen, China, pp: 119-123.

Karayiannis, N.B., 2000. Soft learning vector quantization and clustering algorithms based on ordered weighted aggregation operators. IEEE. Trans. Neural Netw., 11: 1093-1105.

Karnik, N.N. and J.M. Mendel, 2001. Operations on type-2 fuzzy sets. Fuzzy Sets Syst., 122: 327-348.

Koutroumbas, K.D., S.D. Xenaki and A.A. Rontogiannis, 2017. On the convergence of the sparse possibilistic C-means algorithm. IEEE. Trans. Fuzzy Syst., 26: 324-337.

Krishnapuram, R. and J.M. Keller, 1993. A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst., 1: 98-110.

Krishnapuram, R. and J.M. Keller, 1996. The possibilistic C-means algorithm: Insights and recommendations. IEEE. Trans. Fuzzy Syst., 4: 385-393.

Kumar, M. and R. Mathur, 2014. Unsupervised outlier detection technique for intrusion detection in cloud computing. Proceedings of the International Conference for Convergence of Technology (I2CT), April 6-8, 2014, IEEE., Pune, pp: 1-4.

Mendel, J.M., 2001. Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice Hall, Upper Saddle River, New Jersey, USA., pp: 265-272.

Pharate, A., H. Bhat, V. Shilimkar and N. Mhetre, 2015. Classification of intrusion detection system. Int. J. Comput. Appl., 118: 23-26.

Rubio, E., O. Castillo and P. Melin, 2015. A new interval type-2 fuzzy possibilistic C-means clustering algorithm. Proceedings of the 2015 5th World Joint Conference on Soft Computing (WConSC) and the North American Fuzzy Information Processing Society (NAFIPS'15), August 17-19, 2015, IEEE, Redmond, Washington, USA., pp: 1-5.

Simhachalam, B. and G. Ganesan, 2014. Possibilistic fuzzy C-means clustering on medical diagnostic systems. Proceedings of the 2014 International Conference on Contemporary Computing and Informatics (IC3I'14), November 27-29, 2014, IEEE, Mysore, India, pp: 1125-1129.

Xenaki, S.D., K.D. Koutroumbas and A.A. Rontogiannis, 2016. Sparsity-aware possibilistic clustering algorithms. IEEE. Trans. Fuzzy Syst., 24: 1611-1626.