Journal of Theoretical and Applied Information Technology
31st December 2014. Vol.70 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENSEMBLE OF CLUSTERING ALGORITHMS FOR ANOMALY INTRUSION DETECTION SYSTEM

[1] SALIMA BENQDARA, [2] MD ASRI NGADI, [3] JOHAN MOHAMAD SHARIF, [4] SAQIB ALI
[1, 2, 3, 4] Department of Computer Science, Universiti Teknologi Malaysia, Malaysia
E-mail: [1] omqsalima@gmail.com, [2] dr.asri@utm.my , [3] johan@utm.my, [4]saqibali@utm.my

ABSTRACT

Maximizing detection accuracy and minimizing the false alarm rate are two major challenges in the design of an anomaly Intrusion Detection System (IDS). These challenges can be handled by designing an ensemble classifier for detecting all classes of attacks. This is because, single classifier technique fails to achieve acceptable false alarm rate and detection accuracy for all classes of attacks. In ensemble classifier, the output of several algorithms used as predictors for a particular problem are combined to improve the detection accuracy and minimize false alarm rate of the overall system. Therefore, this paper has proposed a new ensemble classifier based on clustering method to address the intrusion detection problem in the network. The clustering techniques combined in the proposed ensemble classifier are KM-GSA, KM-PSO and Fuzzy C-Means (FCM). Experimental results showed an improvement in the detection accuracy for all classes of network traffic i.e., Normal, Probe, DoS, U2R and R2L. Hence, this validates the proposed ensemble classifier.

Keywords: *Intrusion Detection, Ensemble Learning, Voting Ensemble*

1. INTRODUCTION

Confidentiality, Integrity, and availability are the main objectives of computer security. IDS is an automated system which can detect a computer system invasion by using an audit trail provided by the operating system or by using a network monitoring tools. An IDS is a protection system that plays an important role to protect or secure networks. The main target of the IDS is to monitor network events automatically to detect malicious. While designing an IDS, detection accuracy and false positive rate are two important considerations. A single classification technique is not capable of detecting all classes of attacks to achieve acceptable false alarm rate and detection accuracy. Better accuracy rate can be achieved by merging two or more machine learning algorithms to construct the ensemble classifier [1]. In ensemble classifiers, the output of several classifiers used as predictors for a particular problem are combined to improve the accuracy and reduce false alarm rate of the overall system. The core difficulty of ensemble approaches lies in the choice of the algorithms constituting the ensemble and the decision function which combines the results of the different algorithms. Often, the use of more algorithms is seen as advantageous, but it is important to take into account the computational expense added by each new algorithm. The advantage of ensemble approaches is their modular structure, unlike hybrid constructions which are engineered with algorithms having non-interchangeable positions. The ensemble designer can easily replace one or more algorithms with a more accurate one.

Clustering is used in the unsupervised scheme as a machine learning mechanism for discovering patterns that deal with unlabeled data with many different dimensions. Clustering is particularly important in uncovering new attacks which have not been seen before. The major strength of clustering algorithms is that they enable new data to be grouped into relevant coherent groups, thus, resulting in the increased performance of existing classifiers [2].

In this paper clustering ensemble classifiers has been proposed, where each classifier used different learning patterns. The methods organized in this ensemble classifier are KM-GSA [3, 4], KM-PSO [5] and FCM [6].

The rest of the paper is organized as follows: Section 2 discusses the related works on the ensemble approach in IDS. Section 3 and 4 present techniques and data used. Section 5 describes the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

flow of the experiment. The results and discussion of findings are presented in Section 6. Finally, Section 7 concludes the paper.

2. RELATED WORK

Based on a review of the literature [7], detection accuracy is improved by hybrid or an assembly of multiple classifiers.

P. Sadia [8], presented an intrusion detection model with clustering ensemble. The model contained a selection feature that enabled the selection of important attributes from the dataset. A filter method helped in reducing noise and outliers in the data set. Divide and merge helped in calculating the k number of cluster centroids. Results showed that the model achieved high detection rate and low false alarm rate.

Bahri et al. [9] introduced a new approach known as MCSAS. This new approach featured an adaptive strategy for intrusion detection based on a multiple classifier system. MCSAS uses a combined series of multiple classifiers and is intended to reduce the false positives, and the number of undetected attacks, or false negatives. A series of experiments based on the KDD Cup 1999 dataset have proven that the solution performs better, especially in the detection of rare attack types.

Muda et al. [10] proposed an integrated approach by combining K-Means algorithm to form groups of similar data in an earlier stage and Naïve Bayes classifier to classify the clustered data according to attack category. The results showed that the approach achieved better performance over a single Naïve Bayes classifier using KDD Cup 1999 dataset. However, the proposed method suffers from the limitation that it is unable to detect similar attacks such as U2R and R2L.

Folino et al. [11] instead used the entire KDD Cup 1999 dataset and examined the performance of a system composed of several genetic programming ensembles distributed on the network based on the island model. The system showed average performance for the Normal, Probe and DoS classes, but very low for the U2R and R2L classes.

An ensemble model that applied three different learning algorithms (linear genetic programming, neural fuzzy inference and random forest) was proposed by Zainal et al. [12]. A weight is assigned to each training set. This weight indicates a classifier's strength. The same training set is trained by each classifier. In terms of decision making, a class label is assigned by the creator of the classifier based on the weight of the classifier.

A multi-level hybrid model is generated by combining decision tree and Bayesian classification as defined by Xiang et al. [13]. The classifier model is hierarchically structured in the form of class labels in the training set. The results showed that the model improved false negative rate compared to other methods.

Two hybrid approaches for modelling IDS have been proposed by Peddabachigari et al. [14]. One such model utilized an ensemble approach which combined base classifiers and a hierarchical hybrid model known as (DT-SVM), which consisted of a combination of Decision Trees (DT) and support vector machines (SVM). Leaf-node information is initially generated as the training set is passed through the DT classifier. The final output is determined as the leaf-node information is added to the training set which has been trained by the SVM classifier

3. PROPOSED APPROACH

This study has been focused on the design of a new classifier ensemble based on clustering method to improving the classification capability of the IDS system. Ensemble classifiers can be used to improve the accuracy and reduce false alarm rate of the overall system. In this paper a new clustering ensemble classifiers has been designed that consist of KM-GSA, KM-PSO and FCM classifiers. Ensemble classifiers were used to build the individual classifiers and then integrated the outputs of all classifiers to decide final outcome. Different results are obtained from different ensemble classifiers by using different features extracted from the KDD Cup 1999 intrusion detection dataset. Unweight voting scheme is used to select the output of ensemble classifiers. The output is validated by comparing with the highest value produced by the components of ensemble classifier.

4. DATASET

This study used the dataset KDD Cup 1999 [15], which is the largest publicly available and sophisticated benchmark for researchers to evaluate intrusion detection. This study used the 10% of the dataset consisting of 494,020 traffic connections with similar ratios of attacks as in the full dataset [16, 17].

Journal of Theoretical and Applied Information Technology
31st December 2014. Vol.70 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. EXPERMENTAL SETUP

The training and testing data used in this study was comprised of 5,092 and 6,890 records respectively. The composition of these sample data maintains the actual distribution of KDD Cup 1999 data. Experiments presented in this paper are of unsupervised training and its flow is depicted in Figure 1. In this paper, three classifiers are used and each of them is trained using the same training data.

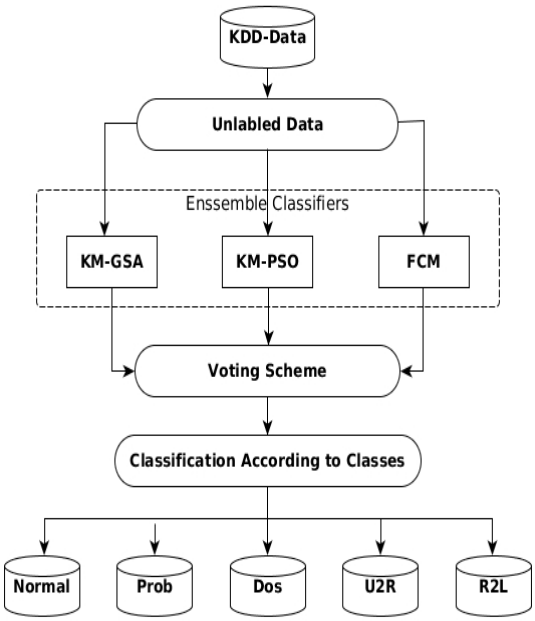

Figure 1: Experimental Flow Diagram

6. RESLUTS AND DISCUSSION

Standard measurements, such as the detection rate (DR), false positive rate (FPR), and detection accuracy rate (ACC), for evaluating the performance of ensemble classifiers model are defined as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP} \quad (1)$$

$$False\_positive = \frac{FP}{FP + TN} \quad (2)$$

$$True\_Positive = \frac{TP}{Total - class - samples}$$

Table 1 summarized the results of individual classifiers and the ensemble classifiers for detection accuracy, true positive rate and false positive rate for all traffic classes.

The results showed that ensemble classifiers achieved the highest detection rate and detection accuracy than individual classifiers in all five classes. Ensemble classifiers achieved highest accuracy with average rate 94.46% and 97.29% for detection accuracy and detection rate respectively. However the best individual classifier, KM-GSA achieved 83.71% and 89.31% for detection accuracy and detection rate respectively. According to the results, ensemble classifiers achieved the lowest false positive rate than KM-GSA in all five classes with average rate 0.07%.

Figure 2 illustrated the accuracy of ensemble classifiers and individual classifiers with respect to the five classes. It is observed that ensemble classifiers showed the highest accuracy as compared to the individual classifiers. The KM-GSA classifier achieved the highest accuracy as compared to KM-PSO and FCM classifiers and lower than ensemble classifiers. The KM-GSA, KM-PSO and FCM classifiers have the similar accuracy by using normal classes. It is observed that all classifiers have close frequency to each other when using U2R and R2L classes. Ensemble classifier showed the highest frequency when using DoS class. It can be concluded that ensemble classifier model have highest frequencies for all five classes. Ensemble classifiers showed better result by improving the detection accuracy for all classes. However each individual classifiers can produce different output results. Thus, the ensemble classifiers improved the detection accuracy for all classes. The experimental results indicate that the detection accuracy is improved by using ensemble classifiers.

Figure 3 and Figure 4 illustrate the comparison in terms of overall accuracy and false positive rate for ensemble classifiers and KM-GSA classifier. Figures 3 illustrates the accuracy of ensemble classifiers and KM-GSA classifier. It is observed that ensemble classifiers achieved the highest accuracy. The detection accuracy for ensemble classifiers had improved by 10.75 % as compared to KM-GSA classifier. Ensemble classifiers showed the better result by improving the detection accuracy because it combined the strong advantage of its each individual classifiers (KM-GSA, KM-PSO, FCM). However each individual classifiers can produce different output results. Figures 4 illustrate false positive rate of ensemble classifiers and KM-GSA classifier. It is observed that ensemble classifiers achieved the lowest false

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

positive. The false positive for ensemble classifiers had reduced by 0.08 as compared to KM-GSA classifier. Ensemble classifiers achieved the lowest false positive because of features of its individual classifiers. Thus, the ensemble classifiers improved overall performance in terms of the detection accuracy rate and false positive.

Table 2 summarized the comparison of various classifier design approaches by different authors with a new proposed ensemble classifiers in building an intrusion detection model. The results in Table 2 showed that the average accuracy of proposed a new ensemble classifiers provided the better result in comparison to single and hybrid classifier. It observed that ensemble classifiers addressed the shortcomings of a single and hybrid classifier.

Table 2: Comparison of various classifier design approaches

Classifier	Classifier design	Average Accuracy (%)
Clustering, SVM [18]	Single classifier	69.8
SVM,NN [19]	Hybrid classifier	86.6
SVM [20]	Single classifier	86.3
SVM[21]	Single classifier	93.0
SVM,SOM [22]	Hybrid classifier	85.3
KM-GSA, KM-PSO, FCM	Ensemble classifiers	94.47

7. CONCLUSION

The aim of this paper is to increase the detection rate and decrease the false positive rate of intrusion detection system using the clustering ensemble classifiers. The ensemble classifier is designed using three classifiers i.e., KM-GSA, KM-PSO and FCM. Furthermore, clustering technique is used to gain benefits from their complementing capabilities. The output of individual classifiers sent to voting scheme to select the final output. The results show that the ensemble classifiers achieve highest detection accuracy and lowest false positive rate for all types of attacks compared to the best individual classifier. Finally, the results concluded that the detection accuracy of IDS has improved significantly by using ensemble classifiers for all traffic classes. Thus, it overcomes the shortcomings of a single classifier which is incapable of detecting all classes of attacks to acceptable false alarm rate and detection accuracy. The input data used in this experiments are 41 dimensional vectors. In future work, dimension reduction and feature selection methods can be used to reduce the training and detection time while maintaining good detection accuracy.

ACKNOWLEDGMENTS:

This work is supported by UTM/RUG/04H11 RMC Universiti Teknologi Malaysia. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFRENCES:

[1] C.Tsai and Lin. C, "A triangle area based nearest neighbors approach to intrusion detection", Pattern Recognition, Vol. 43, No. 1, 2010, pp. 222–229.

[2] K. Wankhade, S. Patka, R.Thool, "An Overview of Intrusion Detection Based on Data Mining Techniques", Proceedings of International Conference in Communication Systems and Network Technologies (CSNT), Gwalior, April

[3] E. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications", Biometrics, Vol. 21, 1965, pp. 768–769.

[4] E.Rashedi, H.Nezamabadi and S. Saryazdi, "GSA: a gravitational search algorithm", Information sciences, Vol. 179, No. 13, 1965, pp. 2232–2248.

[5] D.Van and A. Engel, "Data clustering using particle swarm optimization", Proceedings of International Conference in Evolutionary Computation (CEC), 2003, pp. 215–220.

[6] J.Bezdek, (1981). "Pattern recognition with fuzzy objective function algorithms", Kluwer Academic Publishers, 1981, pp. 1-136-8, 2013, pp. 180 - 185.

[7] A. Mukkamala and A. Abraham, "Intrusion Detection Using an Ensemble of Intelligent Paradigms", Network and Computer Applications, Vol. 28, 2005, pp. 167–182.

[8] P. Sadia, "Intrusion Detection Model Based on Data Mining Technique", Proceedings of International Conference in Advances in Engineering & Technology (ICAET), 2014, pp. 34-39.

[9] E. Bahri, N. Harbi and H. Nguyen Huu, "A Multiple Classifier System Using an Adaptive

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Strategy for Intrusion Detection", Proceedings of International Conference in Intelligent Computational Systems (ICICS'2012), Dubai, Jan 7-8, 2012, pp. 124 – 128

[10] Z. Muda, M. Sulaiman and N. Udzir, "A K means and Naïve Bayes leaning approach for better intrusion detection", information technology journal, Vol. 10, No. 3, 2009, 648–255.

[11] G. Folino, C. Pizzuti, and G. Spezzano, "An ensemble-based evolutionary framework for coping with distributed intrusion detection", Genetic Programming and Evolvable Machines, Vol. 11, 2010, pp. 131–146.

[12] A. Zainal, M.A. Maarof, S.M. Shamsuddin and A. Abraham, "Ensemble of one-class classifiers for network intrusion detection system", Proceedings of International Conference in Information Assurance and Security (ISIAS), Naples, Sept 8-10, 2008, pp. 180 - 185.

[13] C.Xiang, P. Yong and L. Meng, "Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees", Pattern Recognition Letters, Vol. 29, No. 7, 2008, pp. 918-924.

[14] S. Peddabachigari, A. Abraham and C. Grosan, "Thomas: Modeling intrusion detection system using hybrid intelligent systems", Network and Computer Applications, Vol. 30, 2007, pp. 114-132.

[15] KDD, "The 3rd international knowledge discovery and data mining tools competition (KDDCup1999)", California, Irvine: University of California, 2005, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[16] S. Mukkamala and A. Abraham, "Intrusion Detection Using Ensemble of Soft Computing Paradigms", Proceedings of International Conference in Intelligent Systems Design and Applications, Germany, 2003, pp. 239-248.

[17] C .Tsai, Y. Hsu, C. Lin, and Y. Lin, "Intrusion Detection by Machine Learning: A Review", Expert Systems with Applications, Vol. 36, No. 10, 2009, pp. 11994–12000.

[18] L. Khan, M. Awad, and B.Thuraisingham. , "A new intrusion detection system using support vector machines and hierarchical clustering", International Journal on Very Large Data Bases, Vol. 16, No. 4, 2007, pp. 507-521.

[19] G.Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering", Expert Systems with Applications, Vol. 37, No. 9, 2010, pp. 6225–6232.

[20] Venkata and Prasad, "Network Intrusion Detection system based on Feature Selection and Triangle area Support Vector Machine", International Journal of Engineering Trends and Technology, Vol. 3, No. 4, 2012, pp. 466–470.

[21] R. Parimala and R. Nallaswamy, "A study on enhancing classification accuracy of spam e-mail dataset", International Journal of Engineering Trends and Technology, Vol. 21, No. 2, 2011, pp. 15–20.

[22] Fei .W, Yuwen. Q, Yuewei. D and Zhiquan. W, "A Model Based on Hybrid Support Vector Machine and Self-Organizing Map for Anomaly Detection", Proceedings of International Conference on Communications and Mobile Computing (CMC), Shenzhen, April 12-14, 2010, pp. 97 - 101.

Journal of Theoretical and Applied Information Technology
31st December 2014. Vol.70 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Performance of the three classifiers and the ensemble model

Classes	KM-GSA			KM-PSO			FCM			Ensemble Model		
	ACC	FP	DR	ACC	FP	DR	ACC	FP	DR	ACC	FP	DR
Normal	74.9	0.21	87.8	71.8	0.26	65.5	74.1	0.22	65.2	**89.3**	0.1	94.7
Prob	78.7	0.19	85.7	65.3	0.37	85.5	77.7	0.19	66.7	**96.6**	0	99.7
DOS	82.5	0.19	92.7	76.9	0.31	74.3	65.6	0.27	60.9	**99.2**	0	99.2
U2R	88.6	0.11	92.7	83.2	0.16	92.5	84.9	0.15	92.1	**90.6**	0.1	93.6
R2L	93.6	0.06	87.5	88.2	0.11	84.5	90.0	0.09	87	**96.4**	0	99.0
AVG %	**83.7**	**0.15**	**89.3**	**77.1**	**0.24**	**80.5**	**78.4**	**0.18**	**74.4**	**94.4**	**0**	**97.2**

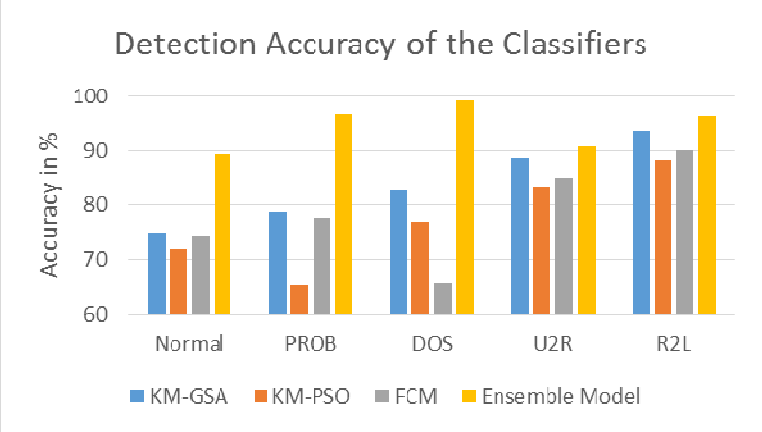

Figure 2: Detection Accuracy of the Classifiers

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

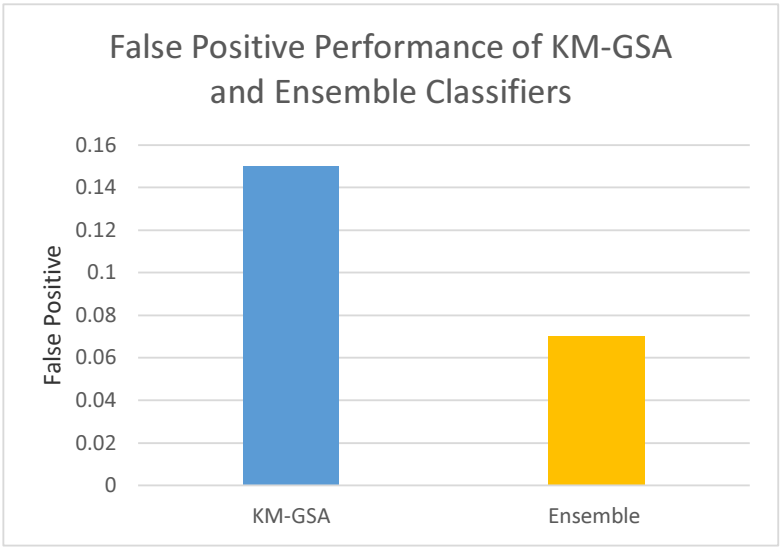

Figure 3: False Positive Performance of KM-GSA and Ensemble Classifiers

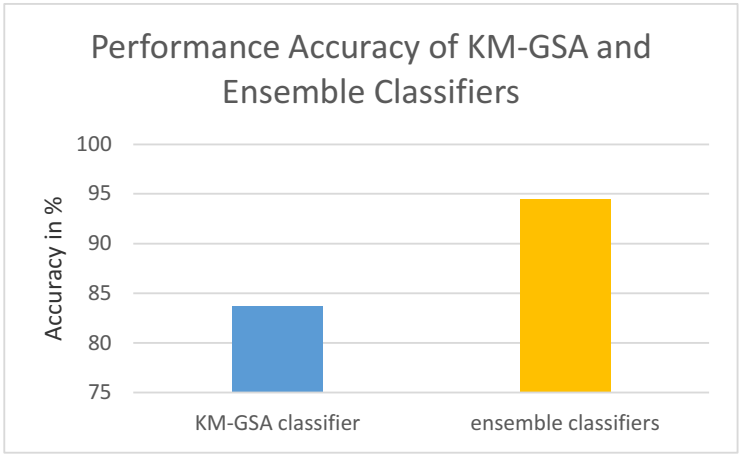

Figure 4: Performance Accuracy of KM-GSA and Ensemble Classifiers