K-Means Clustering and Naive Bayes Classification for Intrusion Detection

Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir
Faculty of Computer Science and Information Technology, University Putra Malaysia,
43400 UPM Serdang, Selangor Darul Ehsan, Malaysia.
warusia@fsktm.upm.edu.my

Abstract-*Intrusion detection systems (IDS) effectively complement other security mechanisms by detecting malicious activities on a computer or network, and their development is evolving at an extraordinary rate. The anomaly-based IDS, which uses learning algorithms, allows detection of unknown attacks. Unfortunately, the major challenge of this approach is to minimize false alarms while maximizing detection and accuracy rates. To overcome this problem, we propose a hybrid learning approach through the combination of K-Means clustering and Naïve Bayes classification. K-Means clustering is used to cluster all data into the corresponding group based on data behavior, i.e. malicious and non-malicious, while the Naïve Bayes classifier is used to classify clustered data into correct categories, i.e. R2L, U2R, Probe, DoS and Normal. Experiments have been carried out to evaluate the performance of the proposed approach using KDD Cup '99 dataset. The results showed that our proposed approach significantly improves the accuracy, detection rate up to 99.6% and 99.8%, respectively, while decreasing false alarms to 0.5%.*

Keywords: Intrusion Detection system; Anomaly Detection; Hybrid Learning; Clustering; Classification

1 Introduction

With the rapid growth of network technology, cyber crime incidents also increase accordingly. System vulnerabilities and valuable information attract attackers' attention. The number of attacks through network has risen dramatically in recent years. Gaining unauthorized access to files and network as well as other serious security threats can be detected by employing intrusion detection systems (IDS). IDS identifies any activity that violates the security policy from various areas within a computer and network environment. An IDS is capable of sending early alarms upon risk of exposure caused by attacks, in order to alert the system administrators to execute corresponding response measurements, thus reducing the possibility of more serious damage to the system/organization. IDS can be identified by two techniques, namely misuse, or signature-based detection and anomaly detection [1]. Misuse detection techniques can detect known attacks by examining attack patterns, matching them to the list of signatures, much like virus detection by an antivirus application. However, this type of IDS requires a frequent updating of the signature database with new signatures; otherwise, it fails to detect unknown attacks if the signature is not in its library. Unlike signature-based detection, anomaly-based detection is designed to capture any activities which deviates from the normal usage pattern/profile, and will be considered as intrusion. Although

anomaly detection has the capability to detect unknown attacks, it has the potential to generate high volume of false alarms.

In recent years, data mining approach have been proposed and used as anomaly detection techniques to discover unknown attacks [2]. This approach has resulted in high accuracy and good detection rates but with moderate false alarm rates on novel attacks. In addition, unresolved issues such as incorrectly predicting an intrusion as normal, and normal instances as attacks has become an inevitable limitation in building an effective anomaly detection. Therefore, there is a need to detect and identify such attacks accurately in an interconnected network.

In this work, we propose a hybrid learning approach based on the combination of two data mining techniques, namely K-Means clustering and Naïve Bayes classification to improve current anomaly-based detection capabilities in terms of accuracy, detection rate as well as false alarm rate. The proposed approach is evaluated using KDD Cup '99 benchmark dataset and compared with the single classifier approach and previous findings. The rest of the paper is organized as follows: in Section 2, related works of this field are discussed. We describe the proposed model in Section 3. Experimental results and comparisons are presented in Section 4. Finally, the conclusion and future work are presented in Section 5.

2 Related Work

Data mining is the latest technology introduced in network security to find regularities and irregulaties in large datasets [3,4]. KDD CUP '99 dataset is the dominating evaluation dataset used by most researchers to test their proposed techniques. The best possible accuracy and detection rate can be achieved using hybrid learning approaches [5]. However, the work to improve false alarm rate is an ongoing affair. Different classifiers can be used to form a hybrid learning approaches such as combination of clustering and classification techniques [6]. Clustering is an anomaly-based detection method that is able to detect novel attacks and forming natural groupings of data based on similarities among the patterns [7].

Tsai and Lin employ K-Means clustering to cluster data instances into k-clusters [6]. Next, the research trains the new dataset, which consists of only the centers of cluster with Support Vector Machine (SVM). They managed to obtain high accuracy rate for almost to all attack types. This approach offers high detection rate but comes with high false alarm rate.

Artificial Neural Network (ANN) are widely used and has been successfully applied to IDS to solve many complex practical problems. Thus, Gang, JinXing and Jian [8] propose a novel approach for ANN-based IDS using ANN and Fuzzy Clustering called FC-ANN [8]. Fuzzy clustering is applied to generate different training subset before a different ANN models are trained to formulate different models. Then, a fuzzy aggregation module is employed to aggregate the result. Each subset of the training set have a lower complexity by employing fuzzy clustering and this directly enables the ANN to learn each subset more precisely in order to detect low frequency attacks such as for U2R and R2L attacks. However this approach yield a lower detection rate for Probe attacks compared to Naive Bayes approach.

Cao, Zhong and Feng [9] propose as an algorithm by combining Artificial Immune Network and Radial Basis Function (RBF) Neural Network [8]. In this work, multiple granularities artificial immune network algorithm is employed to first get a hidden neuron candidate. They then train a cosine RBF neural network based on gradient descent learning process, achieving significant pattern classification and accuracy ability. The experimental results indicate that the proposed approach has the ability to get reasonable detection but it can be further improved.

Intrusion detection based on Fuzzy SVMs (FSVM) was proposed by Shaohua et al. [10] to improve the classification accuracy. The purpose of the clustering algorithm is to construct a new training set using centers of clusters. This new set will then be trained with FSVM to obtain a support vector. Although their results have proved that this method has increased the accuracy rate, it is not of an acceptable percentage.

Amiri et al. [11] used a feature selection method to improve existing classifiers' performance by eliminating unimportant features such as for SVM which have heavy computational challenges for large datasets. Thus, the authors recently introduced an improved Least Squares Support Vector Machine called PLSSVM. PLSSVM performs well in classifying Normal and Probe records but misses a large number of dynamic attacks which are very similar to normal behavior, such as DOS and U2R.

Horng [12] proposed SVM-based IDS with BIRCH hierarchical clustering as a preprocessing phase and a simple feature selection procedure to eliminate the unimportant features. The hierarchical clustering algorithm improves the performance of SVM while the simple feature selection procedure aids the SVM model to correctly classify some data. Since this approach could not make distinction between R2L and Normal data, the prediction percentage for this class decreased dramatically.

Various data mining algorithms are compared by Panda and Patra [14] to detect network intrusions. The author concluded that data mining approaches can increase the detection rate as well as reducing the false alarm with reasonable rate, but there are still room for improvement.

A comprehensive set of classifiers for detecting four attack categories which are available on the KDD dataset are evaluated by Huy [15]. The best classifier for each attack category has been chosen and two appropriate classifiers are proposed for their selection models. Nevertheless, the detection rate for R2L attacks can be improved. Meera and Srivatsa [16] proposed the best performed classifier for each category of attacks by evaluating a comprehensive set of different classifiers using the data collected from Knowledge Discovery Database (KDD). However, there are no false alarm and detection rate reported by the author.

In short, various techniques have been proposed in the intrusion detection field and related work; but there are still room to improve the accuracy and detection rate as well as the

false alarm rate. Our proposed approach offers high detection and accuracy with low false alarm rate compared to others in detecting attacks.

3 Hybrid Learning Approach

Anomaly learning approaches are able to detect attacks with high accuracy and high detection rates. However, the rate of false alarms is also high. In order to maintain the high accuracy and detection rate while at the same time reduce the false alarm rate, we propose a combination of two learning techniques.

For the first stage in the proposed hybrid learning approach, we grouped similar data instances based on their behaviors by utilizing a K-Means clustering as a pre-classification component. Next, using Naïve Bayes classifier we classified the resulting clusters into attack classes as a final classification task. We found that data which have been misclassified during the earlier stage may be correctly classified in the subsequent classification stage.

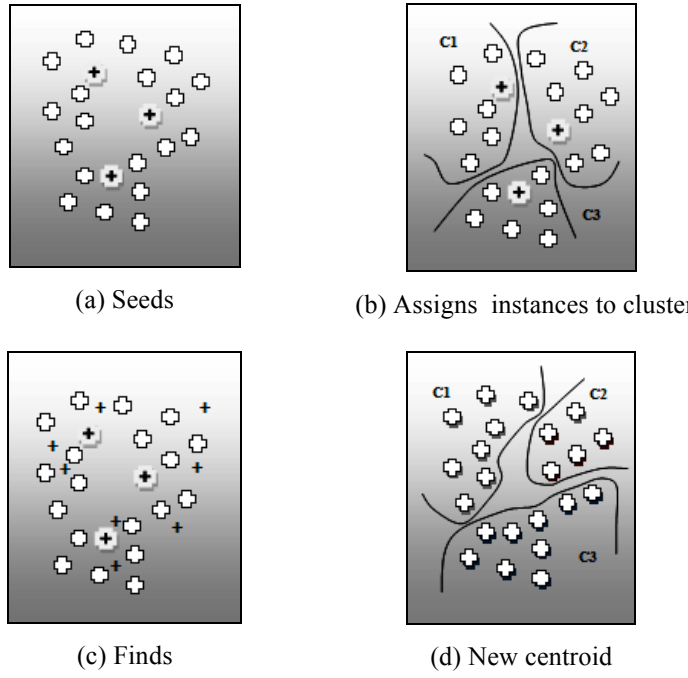
3.1 K-Means clustering

The network intrusion class labels are divided into four main classes, which are DoS, Probe, U2R, and R2L [17]. Figure 1(a) through Figure 1(d) shows the steps involved in the K-Means clustering process. Figure 2 will later show the final overall result with application of the classification approach.

The main goal of utilizing K-Means clustering is to split and group data into normal and attack instances. K-Means clustering methods partition the input dataset into k-clusters according to an initial value known as the seed-points into each cluster's centroids (cluster centers), i.e. the mean value of numerical data contained within each cluster. In our case, we choose k = 3 in order to cluster the data into three clusters (C1, C2, C3). Since U2R and R2L attack patterns are naturally quite similar with normal instances, one extra cluster is used to group U2R and R2L attacks.

Back to Figure 1(b), each input will be assigned to the closest centroid by squaring distances between the input data points and the centroids. New centroids will then be generated for each cluster by calculating the mean values of the input set assigned to each cluster as shown in Figure 1(c).

The steps in Figures 1(b) and (c) are repeated until the result reached a convergence as shown in Figure 1(d).



(a) Seeds (b) Assigns instances to cluster

(c) Finds (d) New centroid

Figure 1. K-Means clustering process

The K-Means algorithm works as follows:

- Select initial centers of the K clusters. Repeat steps 2 through 3 until the cluster membership stabilizes.
- Generate a new partition by assigning each data to its closest cluster centers.
- Compute new clusters as the centroids of the clusters.

3.2 Naïve Bayes classifier

Some behaviors in intrusion instances are similar to normal and other intrusion instances as well. In addition, a lot of algorithms including K-Means are unable to correctly distinguish intrusion instances and normal instances. In order to improve this shortcoming in classification, we combined the K-Means technique with Naïve Bayes classifier. Naïve Bayes has become one of the most efficient learning algorithms [18]. Naïve Bayes are based on a very strong independence assumption with fairly simple construction. It analyzes the relationship between independent variable and the dependent variable to derive a conditional probability for each relationship. Using Bayes Theorem we write:

$$P(H|X) = P(X|H)\,P(H) / P(X) \qquad (1)$$

Let X be the data record. Let H be some hypothesis representing the data record X, which belongs to a specified class C. For classification, we would like to determine P(H|X), which is the probability that the hypothesis H holds, given an observed data record X. P(H|X) is the posterior probability of H conditioned on X. In contrast, P(H) is the prior probability. The posterior probability P(H|X), is based on more information such as background knowledge than the prior probability P(H), which is independent of X. Similarly, P(X|H) is posterior probability of X conditioned on H. Bayes theorem is useful because it provides ways to calculate the posterior probability P(H|X) from P(H), P(X), and P(X|H).

We consider five category classes (C1 = Normal, C2 = DoS, C3 = Probe, C4 = R2L, and C5 = U2R). Given X, we can predict C1, C2, C3, C4, and C5. The Bayes rule is shown in Equation (2).

$$P(C_i|X) = \frac{P(X|C_i).P(C_i)}{P(X)} \quad (2)$$

where C_i represents the category of classes and X is the data record. X may be divided into pieces of instances, say x_1, x_2, ..., x_n which are related to the attributes X_1, X_2, ..., X_N, respectively. The probability obtained is shown in the following Equation (3).

$$P(C_i|X) = \frac{P(x_1|C_i).P(x_2|C_i)\ldots. P(x_n|C_i).\ P(C_i)}{P(X)} \quad (3)$$

The denominator P(X) always constant for all classes. Thus, it can be ignored as in Equation (4).

$$P(C_i|X) = P(x_1|C_i).P(x_2|C_i)\ldots. P(x_n|C_i).\ P(C_i) \quad (4)$$

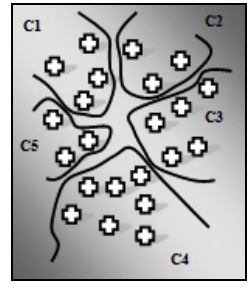


Figure 2. Classifier

Figure 2 shows Naïve Bayes classifier that are used to classify classifies all three clusters as illustrated in Figure 1(d) into more specific categories, which are Probe, Normal, Dos, U2R, and R2L. The combination of these classifiers with the K-Means clustering technique showed an encouraging improvement as compared to previous approaches. The results are surprisingly better in terms of accuracy, detection rate, and false alarm rate.

4 Experiments & Results

4.1 Dataset Description

In our experiments, the KDD Cup'99 benchmark dataset KDD [19] is chosen for evaluation and comparison between the proposed approaches and the previous approaches. The entire KDD data set contains an approximately 500,000 instances with 41 features. The training dataset contains 24 types of attacks, while the testing data contains more than 14 types of additional attack. Further description for the available features and intrusion instances can be found in [20].

In order to demonstrate the abilities to detect different kinds of intrusions, the training and testing data covered all classes of intrusion categories as adopted from [19] as follows:

- **Denial of Service (DoS):** Attacker usually occupies all system sources, disables system resources, and engages all computing or memory resources to be too busy to handle legitimate requests or deny legitimate users from accessing a machine. Examples of attacks are Smurf, Mailbomb, SYN Flooding, Ping Flooding, Process table, Teardrop, Apache2, Back, and Land.
- **Remote to User (R2L):** Attacker sends packets to remote machine over a network and exploits the network vulnerability to gain local access as a user of that machine. Examples of attacks are Ftp_write, Imap, Named, Phf, Sendmail, and SQL Injection.
- **User to Root (U2R):** Attacker takes the advantage of system leak by accessing a normal user's account on the system and exploits system vulnerabilities to get legal administrator access to the system. Examples of attacks are Loadmodule, Perl, Fdformat.

- **Probing:** Attacker performs some preparation step before launching attacks by scanning a network of computers to gather information or to find known vulnerabilities. The attacker will use this information to determine the targets and the type of operating system. Examples of attacks are Nmap, Satan, Ipsweep, Mscan. [19]

Tables 1 and 2 summarize the distribution records for training dataset according to the class type. In order to validate the overall hybrid learning approach, a testing dataset is also used.

Table 1. Sample distribution of the training dataset

Class	No. of Samples	Sample Percentage (%)
Normal	97277	19.69

Probe	4107	0.83
DoS	391458	79.24
U2R	52	0.01
R2L	1126	0.23
Total	494020	100

Table 2. Sample distribution of the testing dataset

Class	**No. of Samples**	**Sample Percentage (%)**
Normal	60593	19.4
Probe	4166	1.33
DoS	231455	74.4
U2R	88	0.028
R2L	14727	4.73
Total	311029	100

4.2 Evaluation Measurement

An efficient IDS requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of IDS is evaluated in terms of accuracy, detection rate, and false alarm rate as in the following formula:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (5)$$

$$\text{Detetion Rate} = (TP) / (TP+FP) \quad (6)$$

$$\text{False Alarm} = (FP) / (FP+TN) \quad (7)$$

Table 3 shows the categories of data behavior in intrusion detection for binary category classes (normal and attacks) in terms of true negative, true positive, false positive and false negative.

Table 3.General Behavior of Intrusion Detection Data

Actual	**Predicted Normal**	**Predicted Attack**
Normal	TN	FP
Intrusions (attacks)	FN	TP

- True positive (TP) when attack data is detected as attack
- True negative (TN) when normal data is detected as normal

- False positive (FP) when normal data is detected as attack
- False negative (FN) when attack data is detected as normal

4.3 Result and Discussion

Table 4 presents the results across all category classes obtained from Naïve Bayes (NB) and proposed hybrid learning approach K-Means with Naïve Bayes (KM+NB) using the training and testing sets. KM+NB have been deployed as in a single running. KM+NB performed better than the single classifier NB in detecting Normal, Probe, and DoS instances. Since Normal, U2R, and R2L instances are similar to each other, KM+NB recorded a comparable result for R2L instances except for U2R instances.

Table 4. Classification result for each category class using training and testing dataset

Dataset	**Training**		**Testing**	
Methods	NB	KM+NB	NB	KM+NB
Normal	91.6	99.6	81	99.5
Probe	99.8	100	95.6	98.3
DoS	94.3	99.5	82.5	99.6
U2R	80	40	80	80
R2L	65.5	61.6	90.3	83.2

Tables 5 and 6 present results across binary category classes obtained from NB and KM+NB using the training dataset. NB is less efficient as the algorithm falsely predicted 818 data as attacks and 471 data as normal as compared to KM+NB with only 40 data and 39 data, respectively.

Table 5. Detection result for the normal and attack classes using training dataset (NB)

Actual	**Predicted Normal**	**Predicted Attack**
Normal	8909	818
Intrusions (attacks)	471	39204

Table 6. Detection result for the normal and attack classes using training dataset (KM+NB)

Actual	**Predicted Normal**	**Predicted Attack**
Normal	9687	40
Intrusions (attacks)	39	39636

In the case of binary category classes prediction for testing dataset, KM+NB performed better than NB as observed in Table 7, where 49 normal data was detected as attack and only 139 attacks data was detected as normal. On the contrary, NB resulted in 1852 false positives and 6448 false negatives as shown in Table 8. In short, NB contributes in increasing false alarm rate as compared to KM+NB.

Table 7. Detection result for the normal and attack classes using testing dataset (NB)

Actual	**Predicted Normal**	**Predicted Attack**
Normal	7875	1852
Intrusions (attacks)	6448	33227

Table 8. Detection result for the normal and attack classes using testing dataset (KM+NB)

Actual	**Predicted Normal**	**Predicted Attack**
Normal	9678	49
Intrusions (attacks)	139	39536

Table 9 shows the measurement in terms of accuracy, detection rate, and false alarm using the training and testing sets of both single classifier and hybrid learning approach. We can see that the single classifier produced a slightly higher accuracy and detection rate but yields high false alarm rates as well. Meanwhile, the hybrid approach recorded high accuracy and detection rate with low false alarm percentage. The clustering techniques used as a pre-classification component for grouping similar data into respective classes helped our hybrid learning approach to produce better results as compared to the single classifier. The hybrid approach also allows misclassified data during the first stage to be re-classified, hence improving the accuracy and detection rate with acceptable false alarms. For instance, the hybrid learning approach enhances the accuracy of the single classifier especially for KM+NB combination, which shows an increase of +16.41% while reducing the false alarm rate up to -18.5%. On the contrary, NB classifier only achieved 83.19% and 19%, respectively. In short, NB suffers high false alarm rate as compared to KM+NB.

Table 9. Single classifiers vs hybrid approach using training and testing dataset

Dataset	**Training**		**Testing**	
Methods	NB	KM+NB	NB	KM+NB
Accuracy	97.39	99.84	83.19	99.6
Detection Rate	97.95	99.89	94.7	99.8
False Alarm	8.4	0.41	19	0.5

Table 10 shows further comparisons made for the proposed hybrid learning approach using the same KDD Cup '99 dataset as in previous researches in terms of accuracy (AC), detection rate (DR), false positive (FP) and false alarm (FA).

Table 10. Further comparison with previous findings

Approaches	**AC**	**DR**	**FP**	**FA**
KM+NB (K-Means+Naïve Bayes)	99.60	99.80	0.09	0.50
Feature Selection + SVM (Amiri, *et al*., 2011)	N/A	98.34	N/A	N/A
BIRCH Clustering + SVM (Horng, 2011)	95.70	N/A	N/A	N/A
ANN + Fuzzy Clustering (Gang, Jinxing and Jian, 2011)	96.71	N/A	N/A	N/A
Fuzzy Clustering + SVM (Shaohua, *et al*., 2010)	91.21	N/A	N/A	N/A
AIN + NN (Cao, Zhong and Feng, 2010)	97.28	N/A	0.18	N/A
Hierarchical Clustering and SVM (Shi, Min and Yuan, 2011)	95.70	N/A	0.70	N/A
TANN (Tsai and Lin, 2010)	96.91	98.95	0.80	3.83
KM-KNN (Tsai and Lin, 2010)	93.55	98.68	0.98	4.79
Hybrid Classifier (Xiang, Yong and Meng, 2008)	96.78	99.21	3.20	3.20
ESC-IDS (Toosi, 2007)	65.48	95.3	N/A	1.9

Amiri et a*l*. [11] proposed the combination of feature selection and SVM method called PLSSVM to improve existing classifiers performance by eliminating the unimportant features. PLSSVM performs with 98.34% as a detection rate which is less accurate than our approach (99.8%).

Although Horng [12]; Gang, Jinxing and Jian [8]; Shaohua et al., [10]; Cao, Zhong and Feng [9] have proved that their respective proposed approaches can obtain reasonable accuracy rates, our results have shown that it can still be improved (99.6%). Anyhow, no detection rate and false alarm has been reported by the authors. In terms of false positive and accuracy rates, our approach achieved better results compared to the recent work by Shi, Min and Yuan [21] which proposed an IDS based on Hierarchical Clustering and Support Vector Machine (SVM).

Tsai and Lin's [6] Triangle Area Nearest Neighbor (TANN) and K-Means with K-Nearest Neighbor (KM-KNN) approaches demonstrated a slightly lower accuracy and detection rates compared to our approach. Moreover, unlike our approach, another potential drawback of this technique is the rate of false alarms. The system proposed by Xiang, Yong and Meng [22] have tendencies to misclassify an normal data as an attack and an attack data as a normal, causing their technique to suffer high false alarms. Meanwhile, the Evolutionary Soft Computing based Intrusion Detection System (ESC-IDS) by Toosi [23] shows serious

shortcomings in its low accuracy rate (65.48%) as well as the tendency to produce high false alarm (1.9%) compared to our approach.

Overall, the proposed approach detected better percentage of attacks than the rest as shown in Table 10 with the accuracy and detection rates above 99.0% , and below 0.5% of false alarm. This is attributed to the combination of K-Means clustering and Naïve Bayes classification that have been used as a pre-classification technique, where similar data are grouped together, and the misclassified data instances during the first clustering stage (using K-Means) were able to be correctly classified in the second stage (using Naïve Bayes). Our proposed hybrid learning approach is proven to be more efficient as compared to previous approaches that are associated with high false alarm rates.

5 Conclusion & Future Work

In this paper, we propose a hybrid learning approach by means of combining K-Means clustering and Naïve Bayes classifiers (KM+NB). The proposed approach was compared and evaluated using the commonly used KDD Cup '99 benchmark dataset. The fundamental solution is to separate instances between the potential attacks and the normal instances during a preliminary stage into different clusters. Subsequently, the clusters are further classified into more specific categories, namely Probe, R2L, U2R, DoS and Normal. Our KM+NB hybrid learning approach significantly reduces false alarm rates with an average below than 0.5%, while keeping the accuracy and the detection rates on average higher than 99%. The approach is able to classify all data correctly, except for attacks of types U2R and R2L. Hence, in future, we are considering the extension of our hybrid IDS by incorporating signature-based detection mechanism, which is better at detecting R2L and U2R attacks.

References

[1] Wenke Lee, J. Salvatore Stolfo, and W. Kui Mok, 1999. A Data Mining Framework for Adaptive Intrusion Detection, Proceedings of the 1999 lEEE Symposium on Security and Privacy, p.I20-132.

[2] Patcha, A., Park, J-M., 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Network*.

[3] Solahuddin, S.,2008. Applying knowledge discovery in database techniques. Modeling Packet Header Anomaly Intrusion Detection Systems. *Journal of Software*, 3(9): p.68-76.

[4] Ming, X., and Changjun, Z., 2009. Applied Research on Data Mining Algorithm in Network Intrusion Detection. International Joint Conference on Artificial Intelligence.

[5] Tsang, C.H., Kwong, S., and Wang, H, 2007. Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. *Pattern Recognition*, 40: p.2373–2391.

[6] Tsai, C.F. and Lin, C.Y, 2010. A triangle area-based nearest neighbors approach to intrusion detection. *Pattern Recognition*, 43(1): p.222-229.

[7] Yang, L. and Li, G., 2007. An active learning based on TCM-KNN algorithm for supervised network intrusion. *Computer and Securtiy*, 26: p.459-467.

[8] Gang, W., Jinxing, H., and Jian, M., 2011. A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. *Expert systems with applications*,376: p.6225–6232.

[9] Cao, L., Zhong, J., and Feng, Y., 2010. Construction Cosine RBF Neural Networks Based on Artificial Immune Networks. *Lecture Notes In Computer Science*, p.134-141.

[10] Shaohua, T., Hongle, D., Naiqi, W., Wei, Z., and Jiangyi, S., 2010. A Cooperative Network Intrusion Detection Based on Fuzzy SVMs. *Journal of Networks*, 5: p.475–483.

[11] Amiri, F., Mohammad, R. Y., Caro, L., Azadeh, S., and Nasser, Y., 2011. Mutual Information-Based Feature Selection for Intrusion Detection System. *Journal of Network and Computer Applications*, 34: p.1184–1199.

[12] Horng, S.J., 2011. A novel intrusion detection system based on hierarchical clustering and support vector machines. *Expert Systems with Applications*. 38(1): p.306-313.

[14] Panda, M. and Patra, M.R., 2008. A comparative study of data mining algorithms for network intrusion detection. In Proceedings of ICETET, India, p.504-507.

[15] Huy Anh, N., and Deokjai, C., 2008. Application of Data Mining to Network Intrusion Detection: Classifier Selection Model. *Lecture Notes in Computer Science*, 5297: p.399-408.

[16] Meera, G., and Srivatsa, S.K., 2010. Classification Algorithms in Comparing Classifier Categories to Predict the Accuracy of the Network Intrusion Detection – A Machine Learning Approach. *Advances in Computational Sciences and Technology*, 3:p.321–334.

[17] Lippmann, R.P., Fried, D.J., Graf, I., Haines, J.W., Kendall, K.R., McClung, D., Weber, D., Webster, S.E., Wyschogrod, D., Cunningham, R.K., and Zissman, M.A., 2000. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX), Los Alamitos, CA, 2: p.12–26.

[18] Harry, Z., and Jiang, S., 2008. Naive Bayes for optimal ranking. *Journal of Experimental and Theoretical Artificial Intelligence*, 20: p.79-93.

[19] KDD (1999). < http://kdd.ics.uci.edu/databases/ - kddcup99/kddcup99.html> [Accessed 5 Jan 2011].

[20] Breiman, L. Et al., 1984. *Classification and regression trees*. Monterey, CA: Wadsworth & Books/Cole Advanced Boks & Software.

[21] Shi-Jinn, H., Ming-Yang, S., and Yuan-Hsin, C., 2011. A novel intrusion detection system based on hierarchical clustering and support vector machines. *Expert Systems with Applications*, 38: p.306–313.

[22] Xiang, C., Yong, P.C., and Meng, L.S., 2008. Design of multiple level hybrid classifier for intrusion detection system using Bayesian clustering and decision tree. *Pattern Recognition Letters*, 29: p.918-924.

[23] Toosi, M., 2007. A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers. *Computer Communications*, 30: p.2201-2212.