Learning a new distance metric to improve an SVM-clustering based intrusion detection system

Roya Aliabkabri sani, Abdorasoul Ghasemi
Faculty of Computer Engineering
K.N. Toosi University of Technology
Tehran, Iran
Roya.AliakbariSani@ee.kntu.ac.ir, arghasemi@kntu.ac.ir

Abstract—**In the recent decades, many intrusion detection systems (IDSs) have been proposed to enhance the security of networks. A class of IDSs is based on clustering of network traffic into normal and abnormal according to some features of the connections. The selected distance function to measure the similarity and dissimilarity of sessions' features affect the performance of clustering based IDSs. The most popular distance metric, which is used in designing these IDSs is the Euclidean distance function. In this paper, we argue that more appropriate distance functions can be deployed for IDSs. We propose a method of learning an appropriate distance function according to a set of supervision information. This metric is derived by solving a semi-definite optimization problem, which attempts to decrease the distance between the similar, and increases the distances between the dissimilar feature vectors. The evaluation of this scheme over Kyoto2006+ dataset shows that the new distance metric, can improve the performance of a support vector machine (SVM) clustering based IDS in terms of normal detection and false positive rates.**

Keywords—Metric learning; Intrusion detection system; Anomaly detection; Clustering Algorithms

I. INTRODUCTION

In recent years, Intrusion Detection System (IDS) has been regarded as one of the important techniques in network security. IDS is a piece of hardware or software, which monitors the traffic of networks and raise alerts if there is a malicious activity. Misused detection and anomaly detection are two mainly different approaches for designing an IDS [1]. In the misused detection approach, IDS uses a set of predefined signatures of attacks for detection of malicious activities. So detection of new attacks is impossible unless we append the signatures of them to the system. The process of constructing these signatures is time consuming. On the other hand, anomaly detection is done by characterizing the normal patterns of the network traffic. This approach is able to detect an unforeseen attack because we can define an attack as a deviation from the normal profile of a system. Since signature based IDSs are easier to develop, they are common methods in the real network architectures, but recent researches concentrate on anomaly detection because they can discover a new attack, if it falls out of the normal traffic pattern [2].

Clustering is one of the most significant techniques for designing an anomaly based IDS. This method partitions similar traffic sessions, according to its features, into the same clusters. These clusters reflect the normal network behavior. During the last decade, many clustering based anomaly detection systems have been proposed [3]. Generally, according to the different assumptions about an anomaly, these intrusion detection schemes are categorized into three classes. In the first category, a data instance, which is not belonging to any cluster, is regarded as an anomaly. In the second category, according to the distance of instances to their closest cluster's center anomaly scores are calculated and the samples with large anomaly scores are considered as anomalies. Finally, in the last category, data instances of small and sparse clusters are marked as anomalies [4]. All of these methods somehow rely on a distance metric. Euclidean distance is the most common distance metric in many algorithms. However, by using the Euclidean distance some features may dominate the others because of different scales. Normalization is a solution for this problem, but it is not sufficient because it does not take into account the importance of features according to the underlying application. Therefore, each application requires its own unique distance function [5].

In this paper, we proposed a metric learning based method to improve the performance of an unsupervised anomaly detection based IDS. By using limited supervision information, we can learn a new distance metric, which has better performance than the Euclidean distance function. The rest of this paper has been organized as follows. Previously clustering based IDSs are reviewed in section II. The background on metric learning is introduced in section III. In section IV, our metric learning based method for improving the performance of a support vector machine (SVM) clustering based IDS is proposed. This method is evaluated in section V over Kyoto2006 data set in terms of normal/attack detection and false positive/negative rates before concluding the paper in section VI.

II. Related work

The enhancement of detection rate is the ultimate goal of proposing a new IDS. Recently, hybrid classifier based approaches have been considered greatly. In these schemes, a combination of several different machine-learning techniques are deployed to design an IDS, which classify the normal and abnormal patterns [6]. Fig.1 shows the overall process of a hybrid scheme that uses two different machine-learning approaches. A set of relevant features can be selected by applying feature selection phase if data instances consist of some redundant features. By ignoring other features, a set of randomly selected instances of dataset with these special features is used for making training data. Sometimes it is required to do a preliminary process on these training data to make them appropriate for the further analysis. For example, some of anomaly detection based IDSs just use the normal profile of the network and this profile is constructed in this step. These data are the input of the basic model, which generate the intermediate results by applying a machine learning method, e.g., a set of optimal parameters or some distinct groups of instances. Finally, by using these intermediate results and applying another machine learning technique, the final model of IDS is constructed.

Several IDSs based on hybrid classifiers have been proposed so far [6]. One of the hybrid IDSs is proposed to improve the performance of SVM algorithm. In this approach, the authors use SVM algorithm along with hierarchical clustering [7]. SVM is a classifier, which can find a discriminate hyper-plane with an optimal margin of separation. Nearest instances to the margin, have the fundamental role in computation of margin. These instances are called support vectors. According to the Fig.1 the basic model of this IDS tries to approximate support vectors by applying hierarchical clustering and the final model is generated by using the results of the basic model. As the author claim, this approach has better performance than pure SVM and similar methods in the term of accuracy, false positive/negative rate and time complexity.

Another hybrid approach is an unsupervised IDS based on K-Means clustering algorithm and iterative dichotomiser 3 (ID3) algorithm. The authors used ARP traffic as normal training data. In the basic model, samples are partitioned into K distinct clusters by applying K-Means algorithm. In the final model, a decision tree per each cluster is found using ID3 algorithm. Applying ID3 on the clusters can solve two essential problems of K-Means clustering, "forced assignment" and "dominance class". A combination of anomaly scores extracted from K-Means and ID3 algorithms, determine final labels of the instances. As the authors report, the proposed IDS has predictive value and specificity of about 98% and 96%, respectively [8].

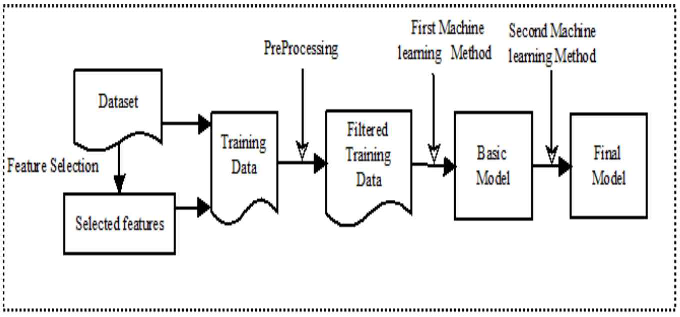

Fig. 1. Overall process of a hybrid classifier based ID

Song et al. [9] have recently proposed a combinatorial unsupervised IDS based on improved K-Means clustering and one-class SVM classifier. The training phase of this algorithm consists of three steps: filtering, clustering and modeling. In the filtering step, by assuming that the number of attacks is far less than normal instances and breaking the interval of each dimension into equal size blocks, attacks are discovered and ignored as a preprocessing step, which is depicted in Fig. 1. In the next step, improved K-means clustering divides the filtered training data into K distinct clusters, so the basic model of Fig.1 is generated. Unlike the original version of K-Means, improved K-Means can determine the number of clusters automatically. Finally, in the last step, K hyper-spheres are obtained by applying one-class SVM to each cluster and the final model of Fig.1 is constructed. A hyper-sphere determines the boundary of a normal cluster. Consequently inside of these borders are the regions of normal traffic. According to the presented process of a hybrid classifier in Fig. 1, the feature selection step is ignored in designing of this IDS.

Testing phase consists of two steps: dividing and classifying. In the first step, each instance is designated to the nearest cluster. Finally, in the last step, if an instance is within the corresponding hyper-sphere of the closest cluster, it is considered as normal otherwise it is an attack. The evaluation, which is done over Kyoto dataset shows that this system can outperform similar method without using some predefined parameters.

Clustering is a machine learning method, which is used by the reviewed IDSs. This scheme is one of the mostly used methods in designing of hybrid classifiers. The process of clustering almost always rely on a distance function. Most of these algorithms do not take into account the importance of distance function and they use the common distance metrics like Euclidean. However, by using the Euclidean distance function, different scaling of features lead to the domination of some features by another [5]. Therefore, deploying of these common metrics lead to the low performance of the clustering based hybrid IDSs. On the other hand, appropriate distance function can improve the efficiency of these systems. So, at the following of this paper, we try to find an appropriate distance function, which is used to improve an SVM-clustering based IDS.

III. Background on metric learning

Metric learning is the process of discovering a new distance function respect to some supervision that is better than the original distance. A set of Supervision is extracted from the ideal distance of some instances.

At the following, $d(x,y)$ refers to the original distance of samples x and y and $\widetilde{d}(x,y)$ is the learned distance. In metric learning, general form of a new distance function is as follow:

$$\widetilde{d}(x,y) = d(f(x), f(y)) \tag{1}$$

It shows that the process of metric learning can be summarized in two simple steps. First, a mapping function according to the supervision must be found. Second, the original distance function is applied to the mapped samples. Since a unique mapping function is learned for all data instances, this method is called global metric learning. There are two classes of global metric learning: Linear and non-linear. In this paper, we focus on linear global metric learning. In this scheme, the mapping function can be expressed in the form of a transformation matrix *G*. Since the Euclidean distance is the most popular distance functions, it is considered as the original metric. Therefore, the new linear distance function can be written in the general form of:

$$\widetilde{d}(x,y) = \|Gx - Gy\|_2 \tag{2}$$

In the literature of metric learning, a new distance function in the form of below is called "Mahalanobis distance" because this formulation is so similar to the general form of Mahalanobis distance function if we replace matrix *A* with the covariance inverse matrix of data i.e. $\sum^{-1}$.

$$\widetilde{d}(x,y) = d_A(x,y) = (x-y)^T A(x-y) \tag{3}$$

For satisfaction of a non-negativity condition of a metric, matrix *A* must be positive semi-definite. Therefore, all of the eigenvalues of *A* are positive and we can write it in the form of $A = GG^T$. Replacement of *A* with GG^T, leads to the transformation of Equation 3 into the square of Equation 2. Therefore, the "Mahalanobis distance" is one of the linear distance functions.

Metric learning requires some supervision information. It can be in the form of relative distance or similarity/dissimilarity constraints. Relative distance constraints are a set of tuples, (x_i, x_j, x_k), which indicate that x_i must be similar to x_j and dissimilar to x_k by applying the new distance metric, i.e.,:

$$d_A(x_i, x_j) < d_A(x_i, x_k) \tag{4}$$

Similarity/dissimilarity constraints are two sets of pairwise constraints

- Similarity constraints are a set of pairs, (x_i, x_j), which should be close to each other in the new distance metric
- Dissimilarity constraints are a set of pairs which are unlike and should not be close in the new distance metric [5]

Using a set of supervision, which is a function of ideal distance, is the main difference between metric learning and dimensionality reduction techniques, e.g., principal component analysis (PCA) or linear discriminant analysis (LDA). Although, these methods are relevant to "Mahalanobis distance" but this dependency is through the projection, which is computed according to their algorithms. For example, PCA is a linear transformation, which ignores the classes of samples and tries to project data into a new coordinate system in which the greatest variance lies on the first coordinate and the second greatest variance in the second coordinate and so on. It does not care about the discrimination of data. Despite, LDA, which tries to model the discrimination of different classes, is based on the assumption that the distribution of data are normal while it is not true about the traffic of a network [10]. However, the focus of metric learning is to learn the underlying metrics in a special application by using a set of supervision.

Selection of an adequate objective function along with one of the previously mentioned supervisions as constraints, turn the problem of learning a new distance metric into a convex optimization problem. Note that, some of the proposed methods in the field of dimensionality reduction also can be define by an optimization problem and they can be classified as a metric learning problem in the certain condition, e.g., Fisher discriminant analysis (FDA) [11].

In the following, a convex optimization problem for metric learning is introduced. The summation of the squared distance between all pairs in the similarity set (*S*) can be considered as the objective function of this problem. The pairs of similarity set should be close to each other in the new distance metric. Therefore, this summation has to be minimized, i.e.:

$$Minimize_A \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2 \tag{5}$$

To avoid the obvious solution *A=0,* the following constraint is added to above problem.

$$\sum_{(x_i, x_j) \in D} \|x_i - x_j\|_A \geq 1 \tag{6}$$

It means that the summation of distances between all pairs in dissimilarity set (*D*) has to be equal or more than 1. Here 1 is just a constant and can be replaced by any other value. Matrix

A has to be positive semi-definite to satisfy the non-negativity condition of a metric. Consequently, the optimization problem will be in the form of a semi-definite programing (SDP) with two constraint in form of below [12]:

$$\begin{aligned} &Minimize_A \quad \sum_{(x_i,x_j)\in S} \left\| x_i - x_j \right\|_A^2 \qquad (7) \\ &s.t. \quad \sum_{(x_i,x_j)\in D} \left\| x_i - x_j \right\|_A \geq 1 \\ &A \geq 0 \end{aligned}$$

IV. PROPOSED METHOD

In this paper, we try to learn an appropriate distance function by deploying metric learning methods to improve the performance of an SVM-clustering based IDS, that is proposed by Song et al. [5]. In the rest of the paper, this IDS is called "Song Model".

Song Model deploys the Euclidean distance function in the clustering phase. However, according to the metric learning concept, using an exclusive distance function in the clustering phase of Song Model can improve the performance of this model. Therefore, we try to apply the conception of metric learning in the field of IDSs.

Fig. 2. depicts the overall process of the proposed method. First of all, the Song Model is constructed using a set of unlabeled data according to a 3-step training phase. This process is introduced in section II. Let $X = \{x_1,...,x_n\}$ be a set of supervision data, which is used to evaluate the Song Model. Y_i' is the label that is assigned to x_i by the Song Model and Y_i is the real label of this instance. According to the result of a simulation, Song Model can detect almost all attacks. Therefore, we try to improve the detection rate of normal traffic. For this reason and because of avoiding a large set of supervision information, misclassified normal supervision data are just used to make supervision information. In other words, if the Song Model assigns the label of -1 (i.e., attack label) to the instance x_i (i.e., $Y_i' = -1$), while it is normal (i.e., $Y_i = 1$), this instance will be selected to make supervision information .The process of making supervision information is described in details in the following subsection.

Based on this supervision information a convex optimization problem is defined. This problem is the basis of a metric learning method because the new distance function is derived from this optimization problem. Subsection *B* describes this process in details.

An appropriate distance function has great influence on the performance of clustering methods. Therefore, this new metric is applied to the clustering step of Song Model while the centers of clusters are supposed to be fixed. Deploying of this new metric leads to the alternative clusters. In other words, a data instance may be assigned to a different cluster using this new metric. Consequently, the border of corresponding SVM model of a cluster is changed so that more normal instances will fall within this border.

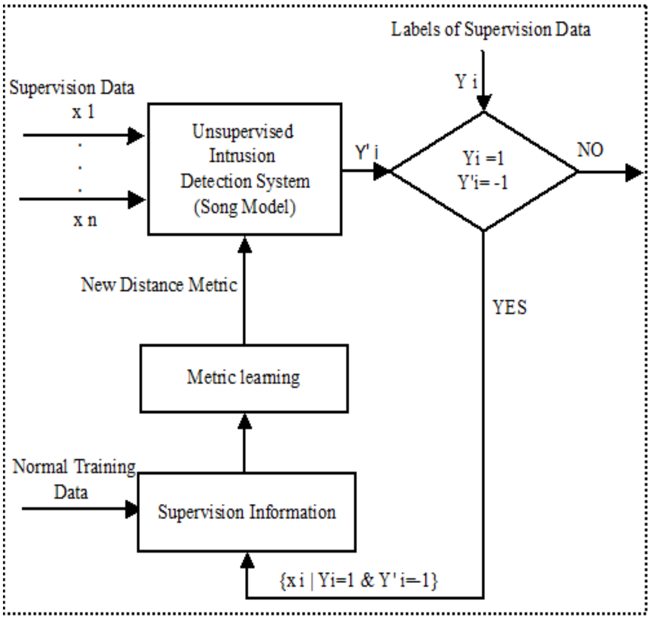

Fig. 2. Overall process of proposed mode

A. *Supervision Information*

In the proposed method, supervision information is in the form of similarity/dissimilarity constraints. A set of misclassified normal supervision data along with some of normal training data are used to form supervision information.

Similarity constraints consist of two different sets of pairs:

- Normal supervision data with wrong labels and their nearest support vector machine of SVM models

A misclassified normal supervision instance should lies within the boundary of its nearest cluster. Therefore, this instance and the nearest support vectors of the SVM model of that cluster should be close to each other using the new distance metric.

- Normal training data and the center of their corresponding clusters

In order to better discriminate between clusters, instances of a cluster should be close to its center using the new distance metric.

Dissimilarity constraints consist of a set of pairs:

- Normal training data and the centers of non-corresponding clusters

This set of pairs also leads to the better separation of different clusters. It means that, the instances of each cluster should be far from the centers of the other clusters using the new distance metric.

To avoid excessive number of constraint 1% of normal training data are randomly selected to make supervision information.

B. *Learning a new distance function*

Suppose that *S* and *D* are two set of similarity and dissimilarity pairs of instances, which are determined using the above procedure i.e.,:

$$S = \{(x_i , x_j)|\ x_i \text{ and } x_j \text{ are similar}\} \qquad (8)$$
$$D = \{(x_i , x_j)|\ x_i \text{ and } x_j \text{ are dissimilar}\}$$

Where x_i and x_j are two samples of either supervision data or normal training data.

Based on this similarity and dissimilarity sets, an optimization problem is defined for learning a new distance metric according to Equation 7. Using this new metric reflects the importance of different features when the distances between samples are measured.

The solution of this SDP problem is the desired matrix *A*. It can be solved using the YALMIP toolbox [13]. YALMIP is a modeling language, which is used as a MATLAB toolbox. The syntax of the YALMIP is consistence with the standard syntax of MATLAB and it is easy to use. An SDP problem can be modeled and solved using the external solvers of YALMIP like SeDuMi.

Consequently, the solution of this problem leads to the new distance function according to the Equation 3.

V. Experimental Result

A. Dataset Description

KDD cup 99 is the most common dataset for evaluating of IDSs. However, most of the 41 features of KDD have redundant information [14]. To overcome this problem a new data set is collected by Kyoto University, which is called Kyoto 2006+ dataset. Each instance in this dataset has 24 features. The first 14 features are directly extracted from KDD dataset according to the feature ranking methods. The other 10 additional features are added to track what is going on in the network. The samples of Kyoto 2006+ are the real connections of the network [5]. In this paper, we use the first 13 continues features of Kyoto dataset.

B. Data Preparation

The ratio of attacks (α) in the real network is very smaller than the normal traffic. In this paper, we suppose that α is equal to 1%. Therefore, 10 different sets of training data including 58294 instances from traffic of November 1-3 2007 are selected randomly and fairly. These sets consist of 582 attacks (1%) and 57712 normal instances (99%).

The proposed model is evaluated by a set of test data including the traffic of 10 days: December 1, 8, 15 and 22 2007, January 10, 17 and 23 2008, February 9, 16 and 23 2008. Finally, a set of supervision data is randomly selected from the traffic of December 4, 11, 18 and 25 2007, January 3, 13 and 20 2008, February 3, 12 and 19 2008. The ratio of supervision data to the original training data is supposed to be 10%.

C. Evaluation and Results

Fig. 3 shows the performance of Song Model and proposed method for different values of parameter *v*. *v* is a one-class SVM parameter that indicates the ratio of instances located out

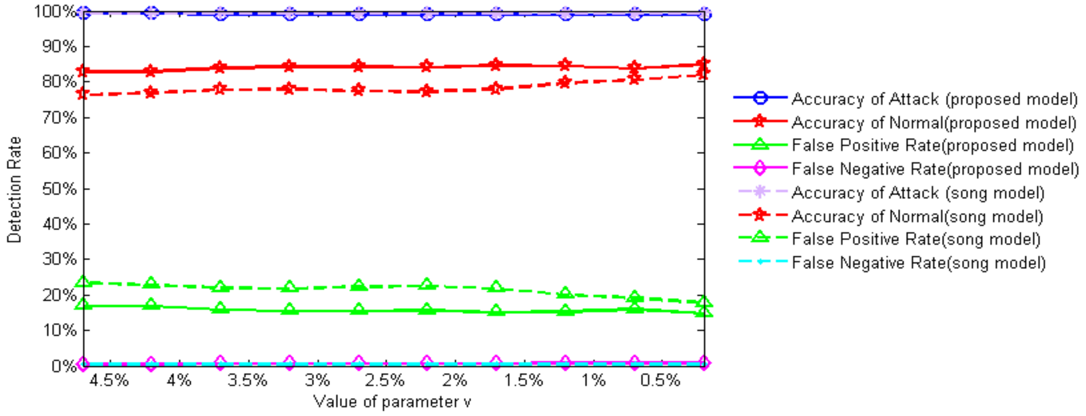

Fig. 3. Performance comparison of two methods

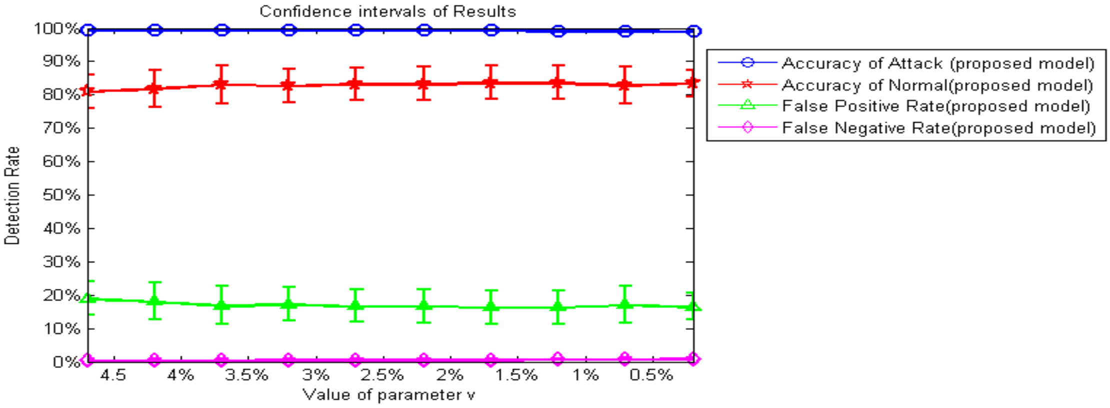

Fig. 4. Sensitivity analysis of proposed method

Table I: Normal detection rate of Song Model and proposed method

Parameter ν	***Normal detection rate of Song Model***	***Normal detection rate of proposed model***
0.002	0.8205	0.8507
0.007	0.8071	0.8383
0.012	0.7979	0.8453
0.017	0.7813	0.8473
0.022	0.7730	0.8417
0.027	0.7761	0.8434
0.032	0.7800	0.8437
0.037	0.7788	0.8410
0.042	0.7704	0.8309
0.047	0.7649	0.8299

-side the borders of SVM models. As shown in Fig.3, by increasing the value of ν, the detection rate of normal is decreased because the corresponding hyper-spheres became smaller and more normal traffic fall out of these borders. As shown in Table I and Fig.3, the normal detection rate of the proposed method is improved about 3% to 7% in comparison with the Song Model. This improvement is greater for larger values of parameter ν. It is reasonable because the improvement is usually done via changing the shape and the size of hyper-spheres to encompass more normal traffic. Therefore, the improvement is greater for the small sized SVM models because they are more flexible for these changes. In other words, when more normal samples fall out of the boundaries of SVM models, the effect of an appropriate distance function will be increased. Table I compares the normal detection rate of Song Model and proposed method according to the different values of parameter ν. There is a tradeoff between normal detection rate and false positive rate. So, it is obvious that the false positive rate of the proposed method will be smaller in comparison with the Song Model. The detection rate of attacks in both methods are almost the same. It means that the alternative clusters do not contain more attacks in comparison with the initial clusters. Since there is a tradeoff between attack detection and false negative rate, this criterion is approximately identical for both methods too. Note that Fig. 3 shows the average evaluation results of two methods over 10 different sets of training data.

Fig. 4 depicts the sensitivity of proposed method to the changes of normal network behavior using the confidence interval. Each confidence interval determines the changes of each criterion for different network behavior.

VI. Conclusion

In this paper, we focus on learning a new distance metric to improve the performance of an SVM-clustering based IDS. In the proposed method, we extract the similarity/dissimilarity sets of supervision information from a small set of supervision data. This information is used to define an SDP problem. We solve this problem using the SeDuMi solver of YALMIP toolbox. The solution is a semi-definite matrix A, which is used to make a new distance metric function. The proposed method applies this new metric to the clustering step of the Song Model while the center of clusters are supposed to be fixed. Consequently, a set of alternative clusters is found. Therefore, the SVM models of clusters change indirectly, so that more normal patterns will fall within the boundaries of these models.

We evaluate the proposed method over a set of real connections of network, which is called Kyoto 2006+ dataset. The results show that the proposed method can outperform the Song Model in the term of normal detection rate and false positive rate.

Refrences

[1] P. G. Teodoro, J. D. Verdejo, G. M. Fernandez and E. Vazquez, "Anomaly-based network intrusion detection: techniques, systems and challenges," Computer & Security, vol. 28, no. 1-2, pp. 18-28, February-March 2009.

[2] A. Patcha and J. M. Park, "An overview of anomaly detection techniques: existing solution and latest technological trends," Comptuter Networks, vol. 51, no. 12, pp. 3448-3470, Aguest 2007.

[3] M. H. Bhuyan, D. K. Bhattacharyya and J. K. Kalita, " Network anomaly detection: methods, systems and tools," IEEE Communication Survey & Tutorials, vol. 16, no. 1, pp. 303-336, February 2014.

[4] V. Chandola, A. Banerjee and V. Kumar , "Anomaly detection: A survey," ACM Computing Surveys (CSUR),vol. 41, no. 3, July 2009.

[5] B. Kulis, "Metric Learning: A Survey," Foundations & Trends in Machine Learning, vol. 5, no. 4, pp. 287-364, 2012.

[6] C. Tsai, Y. Hsu, C. Lin and W. Lin, "Intrusion detection by machine learning: A review," Expert Systems with Applications, vol. 36, no. 10, pp. 11994-12000, December 2009.

[7] L. Khan, M. Awad and B. Thuraisingham , "A New Intrusion Detection System Using Support Vector Machines and Hierarchical Clustering, " The International Journal on Very Large Data Bases, vol. 16, no. 4, pp. 507-521, October 2007.

[8] Y. Yasami and S. Pour Mozaffari, "A novel unsupervised classification approach for network anomaly detection by k-Means clustering and ID3 decision tree learning methods," The Journal of Supercomputing , vol. 53, no. 1, pp. 231-245, July 2010.

[9] J. Song, H. Takakura, Y. Okabe and K. Nakao, "Toward a more practical unsupervised anomaly detection system," Information Sciences, vol. 231, pp. 4-14, May 2013.

[10] K. Fukunaga, introduction to statistical pattern recognition, 2nd ed., San Diego: Academin Press, 1990.

[11] B. Alipanahi, M. Biggs and A. Ghodsi, "Distance Metric Learning VS. Fisher Discriminant Analysis, " in Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2008.

[12] E. P. Xing, A. Y. NG, M. I. Jordan and S. Russell, "Distance metric learning, with application to clustering with side-information," in Advances in Neural Information Processing Systems (NIPS), vol.15, 2002.

[13] J. Lofberg, "YALMIP : a toolbox for modeling and optimization in MATLAB, " in IEEE International Symposium on Computer Aided Control Systems Design, Taipei, 2004.

[14] S. Mukkamala and A. H. Sung, "Identifying significant features for network forensic analysis using artificial intelligent techniques," International Journal of Digital Evidence, vol. 1, no. 4, pp. 1-17, 2003