OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Minkowski Distance based Feature Selection Algorithm for Effective Intrusion Detection

Rm. Somasundaram[1], K. Lakshmanan[2], V. K. Shunmuganaathan[3]
[1](Dean, Computer Applications, SNS College of Engineering, Coimbatore, India)
[2](Principal, Sri Durgadevi Polytechnic College, Chennai, India)
[3](Principal, SNS College of Engineering, Coimbatore, India)

Abstract: *Intrusion Detection System (IDS) plays a major role in the provision of effective security to various types of networks. Moreover, Intrusion Detection System for networks need appropriate rule set for classifying network bench mark data into normal or attack patterns. Generally, each dataset is characterized by a large set of features. However, all these features will not be relevant or fully contribute in identifying an attack. Since different attacks need various subsets to provide better detection accuracy. In this paper an improved feature selection algorithm is proposed to identify the most appropriate subset of features for detecting a certain attacks. This proposed method is based on Minkowski distance feature ranking and an improved exhaustive search that selects a better combination of features. This system has been evaluated using the KDD CUP 1999 dataset and also with EMSVM [1] classifier. The experimental results show that the proposed system provides high classification accuracy and low false alarm rate when applied on the reduced feature subsets.*

Keywords: *Feature Selection, Intrusion Detection, Minkowski Distance, Classification, EMSVM.*

I. Introduction

With the rapid advancements in computer networks, the number of attacks by criminals on such computer networks also increases. Intrusion detection is an important technique which protects the computer network and hence is an essential tool for the network security [2]. According to the type of used pattern, intrusion detection techniques are classified into two categories namely misuse detection and anomaly detection [3][2]. Misuse detection is a rule-based approach and uses stored signatures of known attacks to detect intrusions. Therefore, this approach detects known intrusions reliably with low false positive rate. However, it fails to detect novel attacks when they occur. Moreover, the signature database must be updated manually at frequent intervals for future use when new attacks occur. On the other hand, anomaly detection uses normal patterns to model intrusions. In this model, any deviation from the normal behaviors is considered as anomalies [4]. However, it is very difficult to precisely model all normal behaviors. Therefore, it is easy to classify normal behaviors as attacks by mistake which results in high false positive rate. Therefore, the key consent of anomaly detection algorithm is to select an appropriate model to identify normal and abnormal behaviors.

Feature selection is the process of selecting relevant features by removing the irrelevant or redundant features from the original dataset. This method plays an important role in many different areas including statistical pattern recognition, machine learning, data mining and statistics [3]. Generally, feature selection and structure design are performed independently, i.e., one task is performed without considering another task [2][5]. However, since the subset of input features and the structure of neural network are interdependent, they provide a joint contribution to the performance of the classifier. Recently, a variety of new approaches for feature selection and classification have been proposed to solve the above problem [6][7], in which the input feature subset and the network structure are optimized simultaneously. In such a scenario, the relationship between input feature subset and neural network structure are considered resulting in the improvement of performance of the classifier.

In this paper, we propose a new Minkowski Distance and feature ranking based feature selection algorithm for detecting an effective intrusion detection system. This proposed method uses Minkowski distance for feature ranking and performs exhaustive search to choose a better combination of features. From the experiments conducted in this work, it is observed that the proposed feature selection provides optimal number of features and enhanced the classification accuracy to reduce false alarm rate with minimum time for classification.

The remainder of this paper is organized as follows: Section 2 describes the related works. Section 3 depicts the overall system architecture. Section 4 explains the proposed system. Section 5 provides the results and discussion. Section 6 gives the conclusion and future works.

II. Related Works

There are many research works on clustering approach for data analysis. K-means [8] is one of the simple portitioning algorithms available in the literature that solves the clustering problem. Moreover, the K-means algorithm provides a very simple and easy way to classify a given data set through a certain number of k clusters which are fixed in advance. A clustering algorithm that uses Self Organized Maps (SOM) and K-Means [9] for intrusion detection was proposed in the past which doing the SOM finish its training process, the K-means clustering refines the weights obtained by training, and when SOM finishes its cluster formation, K-means again refines the final result of clustering. A parallel clustering ensemble algorithm was proposed by Hongwei Gao et al [10] for IDS which achieves high speed, high detection rate and low false alarm rate. This parallel clustering ensemble is based on the evidence accumulation algorithm and hence combines the results of multiple clustering into a single data partition, and then detects intrusions with PEA algorithm.

Fengli Zhang and Dan Wang [3] proposed an effective wrapper feature selection approach based on Bayesian Networks for network intrusion detection. The authors evaluated the performance of the selected features using a detailed comparison between the wrapper approach and the other four feature selection methods namely Information Gain, Gain Ratio, ReliefF and ChiSquare using the NSL-KDD dataset. Their experimental results illustrate that the features extracted by their approach makes the classifier to achieve high classification accuracy than the other methods. Sannasi Ganapathy et al [2] proposed two new feature selection algorithms namely an Intelligent Rule based Attribute Selection algorithm and an Intelligent Rule-based Enhanced Multiclass Support Vector Machine for effective classification of intrusion data. These intelligent algorithms perform better feature selection and classification in the design of an effective intrusion detection system.

A hybrid learning approach was proposed by Muda et al [5] by using a combination of K-means and naive Baye's classification algorithm. This approach clusters the data to a corresponding group before applying the classification algorithm. A hybrid anomaly detection system was proposed in [11] which combine k-means clustering with two classifiers namely the k-nearest neighbor and naive Baye's classifier. First, it performs the feature selection from intrusion detection data set using an entropy based feature selection algorithm which selects the important attributes by removing the redundant attributes. Next, it performs cluster formation using the k-means clustering algorithm and then it classifies the results by using a hybrid classifier. Ganapathy et al [1] proposed an intelligent multi level classification technique for effective intrusion detection in Mobile Ad-hoc Networks. They use a combination of a tree classifier which uses a labeled training data and an Enhanced Multiclass SVM (EMSVM) algorithm for enhancing the detection accuracy.

III. System Architecture

This intrusion detection system proposed in this paper consists of five major components namely, KDD Cup Dataset, User Interface Module, Feature Selection Module, Classification Module and Decision Making module. The System Architecture is shown in figure 3.1.

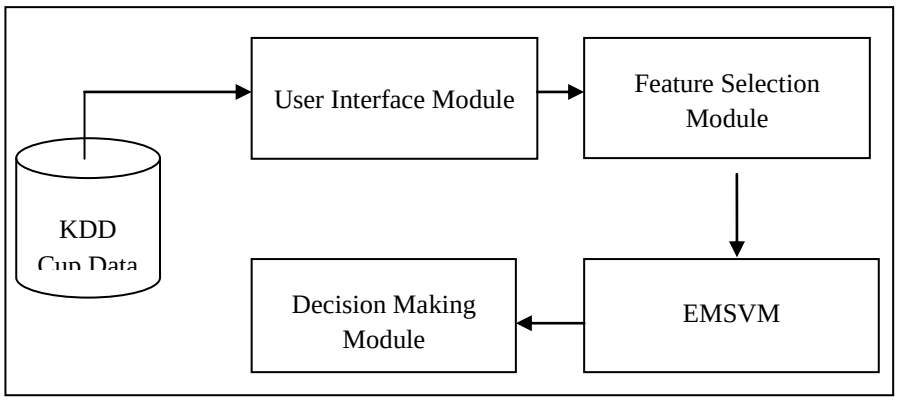

Figure: 3.1 System Architecture

The user interface module collects the necessary data from the KDD'99 Cup dataset. Feature selection module selects the important features using the proposed algorithm. The Classification module classifies the data by using the proposed classification algorithm. Finally, the Decision making module decides whether the particular data is normal or abnormal.

IV. Proposed Work

This section explains the proposed feature selection algorithm. The two important components of the algorithm namely, the feature ranking algorithm and the improved exhaustive search algorithm are examined in the following 2 subsections.

A. Minkowski Distance and Feature Ranking based Feature Selection Algorithm

Generally, feature ranking is performed in two stages, namely, ranking the individual features and ranking by evaluating the subsets of features. Moreover, feature selection is grouped according to the attribute evaluation measure is depending on the type (Filter or wrapper techniques). The filter model relies on common characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one fixed mining algorithm and uses its performance as the evaluation criterion. In this paper, we propose a new feature ranking algorithm based on Minkowski Distance using a single-feature ranking criteria with the filter model.

Phase 1: Feature Ranking Phase

Input : $K = [k_1,k_2,\ldots,k_n]$ - the original feature set

Output: Ranking criterion function, c and ranked features $[l_1, l_2,\ldots l_n]$

1)Initialize: K_i, i=1…n

2)Divide the feature sets into m groups.

3) For each feature $k_i \in K$

a) Compute the scalar Minkowski distance value ($D_{1,k}$) between the mean of a particular group A and the whole set of group, $D_{2,k}$ between the mean of a particular group B and the whole set of group.

b) Calculate and store merit scores f_k using the criterion. $f_k = D_{1,k} + D_{2,k}$

4) Rank features k_i to provide ranked feature set L where $L = [l_1, l_2\ldots l_n]$

Phase 2: Exhaustive Search

Input: $L = [l_1, l2,\ldots l_n]$ the ranked feature set

Output: Evaluation criterion function. f(c,m) and optimal feature subsets R.

1) Initialize: R=[PCR=[ ]; MR=[ ]].
2) Choose a reduced feature subsets which are greater than a threshold value: $F = [k_1, k_2,\ldots, k_m]$, m<=n.
3) Perform a discriminate analysis with F using Minkowski distance in stepwise statistics. To obtain the output by the labeled data (group A and group B): C = classification rate and M = misclassification rate.
4) For each feature $k_i \in F$ and $k_j = \max(F_i)$
   a) Select highest ranked feature and obtain: $R= \{R \cup F_k\}$; $F = \{F - F_k\}$
   b) Perform discriminate analysis with R using Minkowski distance in stepwise statistics. obtain the output by the labeled data from groups A and B using c_k = current classification rate and m_k = current misclassification rate·
   c) Store c_k and m_k into PCR and MR separately.
5) End
6) Obtain the optimal feature subsets R based on evaluation criterion function min (f (c, m)): f(c,m) = sqrt ((1- c) ·100· m ·100)

In [12], a modified greedy search algorithm is used to select the feature subset, which can reduce the iterations to some extent. However, the feature subset is not the optimal. In this work, the evaluation criterion function, which is a quadratic function with the minimum value and hence we can get the optimal feature subset based on the evaluation criterion function. The value range of the evaluation criterion function is between 0 and 100. The final output of this method provides important features for identifying every attack.

B. Enhanced Multiclass Support Vector Machine

In this work, we use the classification algorithm called Enhanced Multiclass Support Vector Machine (EMSVM) [1] for effective classification. In this technique, we test the proposed feature selection algorithm performance with classifier.

V. Results and Discussion

To evaluate our proposed feature selection method, we carried out the implementation of this algorithm by using WEKA software tool on KDD CUP 1999 dataset and calculated the classification rate and misclassification rate. In the experiments, we used Enhanced Multiclass Support Vector Machine (EMSVM) [1] for effective classification of the data set.

A. KDD Cup 1999 Data Set

In the International Knowledge Discovery and Data Mining Tools Competition, only "10% KDD" dataset is employed for the purpose of training [13]. We have selected "10% KDD" as the training data set and "Corrected KDD" as the test data set. Finally, we use EMSVM [1] classifier to validate the algorithm.

B. Experimental Result

The experiments have two phases namely selecting optimal feature subsets for every attack and then classifying the testing data. In the first phase, important attributes from training data of "Corrected KDD" are ranked by the feature ranking values and then an improved exhaustive search algorithm used to select the optimal feature subset. In the second phase, the training data of "10% KDD" was used to train EMSVM classifier and also for testing data of "Corrected KDD" using EMSVM classifier. This algorithm classified the data with full features and also with selected optimal feature subset separately to find the classification rates and false positive rates. Table I gives the optimal number of features selected in this work for all attack types after applying the proposed feature selection algorithm.

Attack Types	**Selected Features**
DoS	1,5,6,10,12, 23, 24, 26, 27, 28, 29, 30, 31,32,33, 34,35, 36, 40,41
Probe	1,5,6,11,12,23,24,31,32,33,34,36,37
R2L	10, 14,17, 19, 26,27,28,29,30,32,35,37,38, 40,41
U2R	1, 5,6,12,13, 23, 24, 27, 28, 30, 31,32,33, 41

Table I: List of Selected Features for Various types of attacks

Table II shows the performance analysis in terms of execution time taken for the proposed Feature Selection Algorithm with classification.

Attacks	**Execution Time Taken (sec)**	
	Malahanobis Distance + Feature Ranking	Minkowski Distance + Feature Ranking
Probe	4.2	2.71
DoS	12	8
Others	4.5	2.16

Table II: Performance of the Proposed Feature Selection Algorithm

From this table, it is observed that the classification accuracy is increased for all types of attacks. Moreover, the training and testing times are reduced for the all the types of attacks namely probe, DoS and others.

Table III shows the comparison of SVM and EMSVM with respect to classification accuracy when the classification is proposed with the selected features obtained from the proposed feature selection algorithms.

Exp. No.	***SVM with Malahanobis Distance Feature Ranking***			***EMSVM with Minkowski Distance Feature Ranking***		
	Probe	DoS	Others	Probe	DoS	Others
1	90.53	89.80	59.62	90.58	90.69	60.52
2	90.78	89.45	61.20	91.21	91.27	62.30
3	90.67	89.70	59.92	91.58	90.49	60.12
4	90.29	89.68	59.70	91.10	91.24	60.13
5	89.83	89.94	60.43	90.30	90.22	61.10

Table III**:** Detection Accuracy Comparisons with Feature Selection

From this Table, it is observed that the classification accuracy is increased in the proposed algorithm when it is compared with the existing algorithms for probe, DoS and others attacks. This is because proposed algorithm performs distance computation with Minkowski distance measures to produce optimal features leading to classification accuracy.

Figure 2 shows the false alarm rate comparison between the proposed feature selection with EMSVM and the existing feature selection with SVM Classifier.

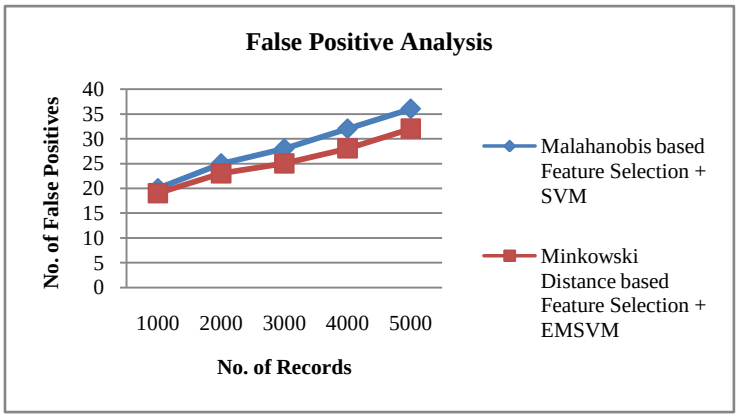

Figure 2: False Positive Analysis

From this figure, it is observed that the false alarm rate is reduced in the proposed model when it is compared with the existing model. This is due to the fact that in the proposed model, classification accuracy is improved by using Minkowski distance.

VI. Conclusion and Future Work

In this paper, we proposed a new intrusion detection system by combining the Enhanced Multiclass Support Vector Machine Algorithm [1] and newly proposed feature selection algorithm called Minkowski Distance and ranking based feature selection algorithm for effective decision making. The experimental results of proposed system provide better detection accuracy and reduced the false alarm rate. In future, the performance of the intrusion detection model can be improved further by adding temporal constraints.

REFERENCES

[1] S.Ganapathy, P.Yogesh, A.Kannan, "An Intelligent Intrusion Detection System for Mobile Ad-Hoc Networks Using Classification Techniques", Computers and Communications and Information Systems, Vol. 148, pp. 117-121, 2011.

[2] Sannasi Ganapathy, Kanagasabai Kulothungan, Sannasy Muthurajkumar, Muthusamy Vijayalakshmi, Palanichamy Yogesh, Arputharaj Kannan, "Intelligent Feature Selection and Classification Techniques for Intrusion Detection in Networks: A Survey", Journal on Wireless Communications and Networking, SprigerOpen, Vol. 271, pp. 1-16, 2013.

[3] Fengli Zhang, Dan Wang, "An Effective Feature Selection Approach for Network Intrusion Detection", Eighth International IEEE Conference on Networking, Architecture and Storage, pp. 307-311, 2013.

[4] Denning D E, "An Intrusion Detection Model", IEEE Transactions on Software Engineering, Vol. 51, no. 8, pp. 12-26, Aug. 2003.

[5] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification", In Proceedings of 7th IEEE International Conference on IT in Asia, 2011.

[6] Guohua Geng, Na Li, Shangfu Gong, "Feature Selection Method for Network intrusion based on Fast Attribute Reduction of Fuzzy Rough Set", 2012 International Conference on Industrial Control and Electronics Engineering, pp. 530-534, 2012.

[7] Janya Onpans, Suwanna Rasmequan,Benchaporn Jantarakongkul, Krisana Chinnasarn, Annupan Rodtook, "Intrusion Feature Selection Using Modified Heuristic Greedy Algorithm of Itemset", 13th International Symposium on Communications and Information Technologies (ISCIT), pp. 627-632, 2013.

[8] Yang Zhong, Hirohumi Yamaki, Hiroki Takakura, "A Grid-Based Clustering for Low-Overhead Anomaly Intrusion Detection", IEEE Conference on Grid Computing and Security, pp.17-24, 2011.

[9] WANG Huai-bin, YANG Hong-liang, XU Zhi-jian, YUAN Zheng, "A clustering algorithm use SOM and K-Means in Intrusion Detection" In Proceedings of IEEE International Conference on E-Business and EGovernment, pp.1281-1284, 2010.

[10] Hongwei Gao, Dingju Zhu, Xiaomin Wang, "A Parallel Clustering Ensemble Algorithm for Intrusion Detection System'', In Proceedings of Ninth IEEE International Symposium on Distributed Computing and Applications to Business, Engineering and Science, pp.450-453, 2010.

[11] Hari Om, Aritra Kundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", In Proceedings of First IEEE International Conference on Recent Advances in Information Technology, 2012.

[12] Janya Onpans, Suwanna Rasmequan,Benchaporn Jantarakongkul, Krisana Chinnasarn, Annupan Rodtook, "Intrusion Feature Selection Using Modified Heuristic Greedy Algorithm of Itemset", 13th International Symposium on Communications and Information Technologies (ISCIT), pp. 627-632, 2013.

[13] KDD Cup 1999 Data, Information and Computer Science, University of California, Irvine.