Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM

Annie George
Dept. Of Computer Science
Amrita Vishwa Vidyapeetham
University, Amritapuri,
Kollam , Kerala

ABSTRACT

Anomaly detection has emerged as an important technique in many application areas mainly for network security. Anomaly detection based on machine learning algorithms considered as the classification problem on the network data has been presented here. Dimensionality reduction and classification algorithms are explored and evaluated using KDD99 dataset for network IDS. Principal Component Analysis for dimensionality reduction and Support Vector Machine for classification have been considered for the application on network data and the results are analysed. The result shows the decrease in execution time for the classification as we reduce the dimension of the input data and also the precision and recall parameter values of the classification algorithm shows that the SVM with PCA method is more accurate as the number of misclassification decreases.

Keywords

Intrusion Detection, Anomaly Detection, Principal Component Analysis, Support Vector Machine

1. INTRODUCTION

As computer networks become an important part of the current world, the threats to it also increase day by day. To detect various threats, intrusion detection systems are needed [1]. There are two types of intrusion detection systems: host-based and network-based. Host-based technology examines events like what files were accessed and what applications were executed. Network-based technology examines events as packets of information exchange between computers. One of the main problems for NIDSs is to build effective behaviour models to distinguish normal behaviours from abnormal behaviours by observing data [2, 3].

There are two types of intrusion detection approaches, misuse detection where we model attack behaviour or features using intrusion audit data and anomaly detection, which is to model normal usage behaviours. Usually in the commercial NIDS, the signature or misuse based approach is followed but anomaly based approach is efficient using the machine learning methods. There are many data mining and machine learning methods used for network intrusion detection. Unsupervised methods such as clustering and supervised methods such as Naïve Bayes, Support Vector Machine are used. But comparisons of the results of using an unsupervised dimensionality reduction method along with the supervised SVM method to SVM without dimensionality reduction is not considered much.

A framework for network anomaly detection using machine learning techniques is explained in this paper, which includes comparison of dimensionality reduction and classifier construction algorithms. KDD99 benchmark dataset is taken to evaluate the performance of our system [4, 5]. Also the KDD99 dataset mapping is done for obtaining the feature vector format. Comparison of anomaly detection using machine learning algorithms is done which shows the need of PCA along with SVM.

2. SYSTEM DESIGN

The first intrusion detection model based on data mining was proposed by Denning [1] and many research works have been devoted to the construction of effective intrusion detection models.

The KD99 dataset for intrusion detection is meant for data mining algorithms, and was established by the Third International Knowledge Discovery and Data Mining Tools Competition [4]. In the KDD99 data set, each data record corresponds to a set of derived features of a connection in the network data. Each connection is labelled either as normal or as an attack, with exactly one specific attack type.

In this paper, we will compare various machine learning algorithms that can be used for anomaly detection. The technical challenges in NIDSs based on machine learning methods are dimensionality reduction and classification. There are three main parts depicted in a framework in figure 1 for an intrusion or anomaly detection tool: pre-processing of network data, feature extraction, classification

The KDD99 IDS dataset have been used as it is the most widely used benchmark dataset for intrusion detections with data mining. The benchmark KDD99 dataset consisting of connection records for each network connection represented by the forty two most important features derived from the network data [5]. From this labelled connection records we need to map the labels to numeric values so as to make it suitable to be the input of our machine learning algorithms. Also assign target class to the connections according to class label feature, which is the last feature in the connection record and assigns a target class zero for normal connection and a one for any deviation from that.

Dimension reduction is to be done for the given set of forty two features in the dataset. The PCA algorithm is considered [6]. Input is the set of connections represented by the forty two features. This module is important as we can represent the original features using a reduced feature set with maximum variance, which is explained later in this paper.

After the dimension reduction using PCA, the reduced set of features that are linear combination of original features is obtained. The classifier solves the anomaly detection problem using the Support Vector Machine algorithm [9] with the output from the PCA algorithm. Comparison of the classification output using support vector machine (SVM) without dimension reduction and classification using the original data is also done.

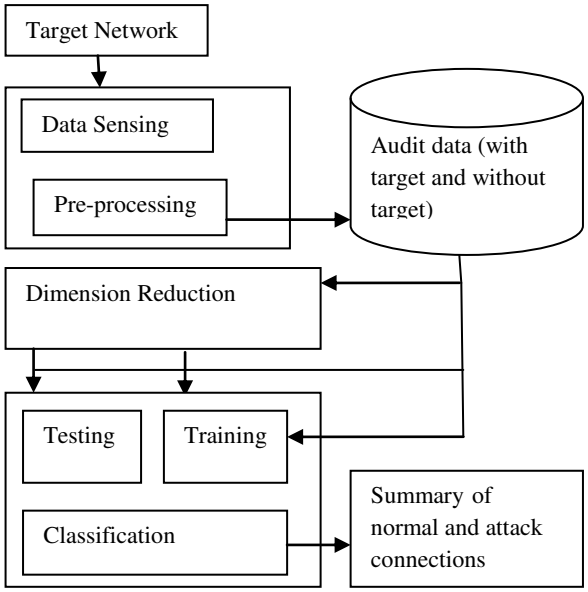

Figure 1. An IDS framework

3. MAPPING CONNECTION RECORD

As the connection record in the KDD99 dataset consists of both nominal and binary data, a mapping mechanism is needed for the nominal data to make it suitable for the algorithms. Some of the nominal attributes in the record are the protocol name, the service and the class label. The mapping of these attributes is done using a dictionary with a unique value for each unique nominal attribute defined. While the nominal to numeric conversion is done, a comparison is done for each current attribute from the connection record and if a match is found from the dictionary, the nominal attribute is mapped to its corresponding value from the dictionary, for example if the nominal attribute is tcp with its dictionary value as 17, then tcp will be replaced by the dictionary value 17. The target class for the classification method is also assigned, for a normal class label, a 0 and for an anomaly class label, a 1.

4. DIMENSION REDUCTION USING PRINCIPAL COMPONENT ANAYSIS

PCA is one of the most fundamental tools of dimensionality reduction for extracting effective features from high-dimensional vectors of input data [7, 8]. In this section, we will see the application of PCA for dimensionality reduction of network connection data consisting of forty two features, for making the classification problem more efficient.

4.1 Algorithm

- Consider the network data corresponding to each connection record after mapping. Thus each column represents a dimension of the input data.
- Compute the mean for each dimension, and subtract it from each data value.
- Compute the covariance matrix C of the input data matrix.
- Calculate the Eigen values and the corresponding eigenvectors for this covariance matrix, and the principal components are computed by solving the eigenvalues problem of covariance matrix C.
- To find the principal components, choose the eigenvectors corresponding to K largest eigenvalues, where K<<N.

Dimensionality reduction step keep only the terms corresponding to the K largest eigenvalues. Hence obtain a new feature vector consisting of eigenvectors of principal components. The final data computed using this feature vector and the mean adjusted original input data using the given equation

$$Final\ Data = RowFeatureVector \times RowDataAdjust$$

RowFeatureVector is the matrix in which eigenvectors in the columns transposed and RowDataAdjust is the mean adjusted input data. The obtained subspace is spanned by the orthogonal set of eigenvectors which reveal the maximum variance in the data space.

Using PCA mapping high-dimensional data into low dimensional data reduces the calculation cost of NIDS and improves the efficiency of the analysis. Here principal component analysis has been used for dimensionality reduction of the forty two dimensions and the output of PCA method provides a set of features that are the linear combination of the original set of features. It accomplishes this by projecting data from a higher dimensional space to lower dimensional space such that error incurred by reconstructing the data in higher dimension is minimized. Thus the input for the SVM becomes more efficient as it represents the principal components that are with maximum variance and that are orthogonal, thereby making the new sub space consisting of features somewhat clustered according to variance and hence the classification by discriminating plane which considers minimum variance becomes more accurate. When viewed from an informative view point, PCA provides SVM with the features that provide efficient classification.

5. CLASSIFICATION USING SUPPORT VECTOR MACHINE

In this section, we will apply multi-class Support Vector Machines (SVMs) [9] for classifier construction in IDSs and evaluate the performance of SVMs on the KDD99 dataset. Support vector machines are based on the idea of constructing optimal hyper-planes to improve generalization abilities [10]. It is a supervised learning method, where we provide training. The idea of multi-class is to construct multiple two-class SVM classifiers and combine their classification results using the voting method. The linear SVM is considered. Here we are having a positively labelled data set which can be our normal class and a negatively labelled data set which represents our anomaly class. The purpose of SVM is to maximize the separating margin of the two classes in the feature space and to minimize the training error. The generalization ability of the SVM depends upon the value of the margin. Then find the support vectors, that is the data

points that are on the margin. It helps us to correctly classify the classes. Alpha coefficient for the kernal function can be computed using the below equations

$$\alpha 1\Phi(s1).\Phi(s1) + \alpha 2\Phi(s2).\Phi(s1) + \alpha 3\Phi(s3).\Phi(s1) = -1$$

$$\alpha 1\Phi(s1).\Phi(s2) + \alpha 2\Phi(s2).\Phi(s2) + \alpha 3\Phi(s3).\Phi(s2) = +1$$

$$\alpha 1\Phi(s1).\Phi(s3) + \alpha 2\Phi(s2).\Phi(s3) + \alpha 3\Phi(s3).\Phi(s3) = +1$$

Here $\Phi(\ ) = I$. The alpha values relate to the discriminating hyper plane which discriminates the positive from the negative examples given by the equation

$$\hat{w} = \sum_i \alpha_{i\hat{s}_i}$$

Hence we get separating hyper plane as weight supported with offset c as

$$y = wx + c$$

There are two phases in the SVM algorithm, the training phase and the testing phase as it is a supervised learning algorithm.

5.1 Training Phase

We provide the SVM algorithm the input that includes the target class, and then the above steps are executed for the training dataset. It calculates the margin, the support vectors, the alpha values and then the weights. For our connection records, class labels as 0 for normal and 1 for anomaly class is assigned. This phase generates a training model for the data.

5.2 Testing Phase

In the testing phase, we provide the test dataset without the target class. This phase considers the model generated by training for classification problem. For classification, the voting method is used, where for each input set, the class having maximum votes is considered. Then the input data belongs to that class. Here vote represents the decision of each binary classification. For our connection records, the classification will be as whether each record is normal or an anomaly.

6. EXPERIMENTAL RESULTS

We perform the evaluation of our machine learning methods using KDD99 dataset. We are considering a set of connection records from the training and testing data of KDD99 dataset for evaluating our algorithms. We are comparing the performance based on SVM and SVM with PCA algorithms.

When we consider the PCA for dimensionality reduction, the entire set of forty two features are considered as the input. The PCA finds the principal components among the set of features that are having the largest eigenvalues. The PCA provides a linear combination of the original features as selected features which are uncorrelated with one another. The results show that the feature set is reduced to twenty eight retaining the class label feature. Even though the output is a linear combination of features, the class label feature doesn't change. The low dimensional data corresponding to the maximum variance principal components are used for classification. As the components are orthogonal in the new subspace where the features are grouped according to maximum variance, it enhances the classification algorithm to find a linear discriminating plane by classifying the features according to minimum variance, thereby increasing accuracy by decreasing the number of misclassification.

The classification algorithm identifies any anomaly in the test data according to the training model. SVM is evaluated using the original data and the reduced data for performance analysis. Using Principal Component Analysis for the data after pre-processing increases the classification accuracy of the network data as it finds the principal components which improves the linear separability of the data.

Evaluation based on the SVM with PCA approach gives less misclassification compared to SVM method. The SVM method uses the original set of forty two features and SVM with PCA uses the mapped set of twenty eight features with class label retained. Performances of both algorithms for anomaly detection are evaluated based on the precision and recall parameters. But the execution speed of second method is more as it takes less time with reduced feature set in the orthogonal space where the classification of the features is enhanced by the maximum variance components that are grouped thereby increasing classification speed, and the result is shown in Table 1. Comparison of both the methods shows that using PCA for reducing the high dimensional network data improves the speed of the detection system by enhancing the feature classification which is very important for an intrusion detection system.

Table 1. Performance based on execution time

Classifiers	Execution time(s)
SVM (42 features)	0.293
SVM with PCA (28 features)	0.009

Precision and recall values for each category, that is here normal and anomaly class, are calculated. Precision and recall can be calculated for a particular category as given below

$$Precision = \frac{Samples\ correctly\ classified\ into\ a\ category}{Total\ samples\ classified\ into\ this\ category}$$

$$Recall = \frac{Samples\ correctly\ classified\ into\ a\ category}{Total\ correct\ samples\ in\ this\ category}$$

Table 2 shows the performance of the classification methods based on precision and recall for anomaly and normal classes. Figure 2 shows the comparison results for both categories and we can observe that latter one shows higher precision and recall value which explains that it causes only one or two misclassifications as the values depends on the correctly classified samples and hence more accurate. The parameter values increases according to the number of correctly classified samples which shows its accuracy.

Table 2. Performance based on precision and recall

Classifiers	Precision	Recall
SVM	0.7708	0.8125
SVM with PCA	0.9375	0.875

From the figure 2 it is evident that the detection method using both PCA and SVM algorithm shows more classification accuracy than the other method according to the parameter values. This higher accuracy is because of the PCA method which finds the maximum variance components which can be classified using a linear SVM more efficiently. The existing supervised methods can be improved using an unsupervised dimensionality reduction method.

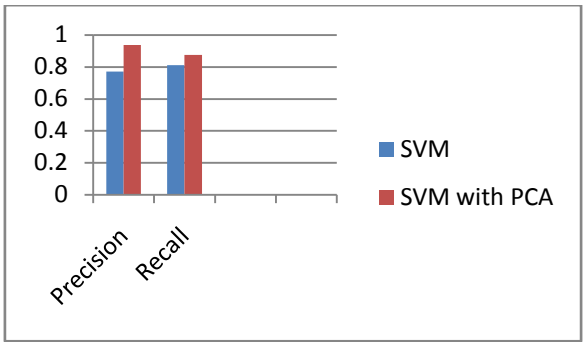

Figure 2. Comparison of classification methods using Precision and Recall

7. CONCLUSION

The work examines an anomaly detection system using machine learning algorithms such as Principal Component Analysis and Support Vector Machine. The KDD99 connection record has been converted into the required format for the machine learning algorithms using a mapping technique that is important in case of any machine learning algorithm. Dimensionality reduction with PCA helps to reduce high dimensional network data to provide the more informative features from the data thereby decreasing the execution time for classification and also increasing the classification accuracy.

Support vector machine (SVM) helps to classify our reduced network data to detect it as a normal or an anomaly connection. The generalization concept can help to obtain better classification result. Using PCA with SVM provides higher accuracy as the output of PCA is the maximum variance components which can be efficiently separated using the hyper plane. We can also consider multiple classes in the anomaly class as we use multi-class SVM for classification. The experimental results have been analyzed based on precision and recall values for each class and it shows that classification using dimensionality reduction is more accurate depending on the new subspace where the features are combined together according to maximum variance which enhances classification using discriminating plane. Further work can be done for more types of anomalies that are emerging at present, which helps the NIDS to be more efficient. The algorithms can be used for anomaly detection in other application areas also as here it is for network data.

8. ACKNOWLEDGEMENT

We would like to express our gratitude to Dr. M. Ramachandra Kaimal of Computer Science Department, Amrita School of Engineering for his motivation and direction towards preparation of this paper. We would also like to express our gratitude to Amrita School of Engineering, Computer Science Department for providing us with facilities to complete the project.

9. REFERENCES

[1] M. M. Sebring, E. Shellhouse, M. E. Hanna, and R. Alan Whitehurst, “Expert systems in intrusion detection: A case study”, In Proceedings of the 11th National Comput Security Conference, Baltimore, Maryland.

[2] W.K. Lee, S.J.Stolfo, “A data mining framework for building intrusion detection model”, In: Gong L., Reiter M.K. (eds.): Proceedings of the IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, 1999.

[3] W.K. Lee, et al., "Mining audit data to build intrusion detection models", In Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pp.66-72, 1998.

[4] H. Güneş Kayacık, A.NurZincir-Heywood, Malcolm I. Heywood, “Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets”, Dalhousie University, Faculty of Computer Science, 6050 University Avenue, Halifax, Nova Scotia.

[5] Mahbod Tavallaee,Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set “, Proceedings of the IEEE Symposium on Computational Intelligence in Security and Security Applications, 2009.

[6] Fengxi Song, Zhongwei Guo, Dayong Mei, “Feature selection using principal component analysis”, Department of Automation and Simulation New Star Research Inst. Of Applied Tech. in Hefei City Hefei, China, International Conference on System Science, Engineering Design and Manufacturing Informatization, 2010.

[7] Lindsay I Smith, “A tutorial on Principal Components Analysis”.

[8] CHEN Bo, Ma Wu, “Research of Intrusion Detection based on Principal Components Analysis”, Information Engineering Institute, Dalian University, China, Second International Conference on Information and Computing Science, 2009.

[9] Chin-Jen Lin, “Formulations of Support Vector Machines: A Note from an Optimization Point of View”, Department of Computer Science and Information Engineering, National Taiwan University, Neural Computation 13, 2001.

[10] Zhangxue-qin, Gu chun-hua and Linjia-jun, “Intrusion detection system based on feature selection and support vector machine”, East China University of Science and Technology, Proceedings of IEEE, 2006.