NRL Release Number 03-1221.1-2312

A Novel Anomaly Detection Scheme Based on Principal Component Classifier

Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL, USA
shyu@miami.edu

Shu-Ching Chen
School of Computer Science
Florida International University
Miami, FL, USA
chens@cs.fiu.edu

Kanoksri Sarinnapakorn
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL, USA
ksarin@miami.edu

LiWu Chang
Center for High Assurance Computer Systems
Naval Research Laboratory
Washington, DC, USA
lchang@itd.nrl.navy.mil

Abstract

This paper proposes a novel scheme that uses robust principal component classifier in intrusion detection problems where the training data may be unsupervised. Assuming that anomalies can be treated as outliers, an intrusion predictive model is constructed from the major and minor principal components of the normal instances. A measure of the difference of an anomaly from the normal instance is the distance in the principal component space. The distance based on the major components that account for 50% of the total variation and the minor components whose eigenvalues less than 0.20 is shown to work well. The experiments with KDD Cup 1999 data demonstrate that the proposed method achieves 98.94% in recall and 97.89% in precision with the false alarm rate 0.92% and outperforms the nearest neighbor method, density-based local outliers (LOF) approach, and the outlier detection algorithm based on Canberra metric.

***Keywords**: Anomaly detection, data mining, intrusion detection, outliers, principal component analysis.*

1. Introduction

Communication networks make physical distances meaningless. People can communicate with each other through the networks without any restriction of the real distance. While we treasure the ease of being connected, it is also recognized that an intrusion of malicious or unauthorized users from one place can cause severe damages to wide areas. Heady et al. [8] defined an intrusion as "any set of actions that attempt to compromise the integrity, confidentiality or availability of information resources." The identification of such a set of malicious actions is called intrusion detection problem that has received great interest from the researchers.

The existing intrusion detection methods fall in two major categories: *signature recognition* and *anomaly detection* [10][18]. For signature recognition techniques, signatures of the known attacks are stored and monitored events are matched against the signatures. The techniques signal an intrusion when there is a match. An obvious limitation of these techniques is that they cannot detect new attacks whose signatures are unknown. In contrast, anomaly detection builds a model from normal training data and detects deviation from the normal model in the new piece of test data. A large departure from the normal model is likely to be anomalous. Anomaly detection algorithms have the advantage that they can detect new types of intrusions [3] with the trade-off of a high false alarm rate. This is because the previously unseen, yet legitimate, system behaviors may also be recognized as anomalies [4][16].

There are various intrusion detection techniques in anomaly detection category including machine learning techniques (e.g., robust support vector machines [9]) and statistical-based methods. An extensive review of a number of approaches to novelty detection was given in [19][20]. Statistical-based anomaly detection techniques use statistical properties of the normal activities to build a norm profile and employ statistical tests to determine whether the observed activities deviate significantly from the norm profile. A multivariate normal distribution is usually assumed, which can be a drawback. A technique based on a chi-square statistic that has a low false alarm and a high detection rate was presented in [25]. Emran and Ye [5] developed a multivariate statistical based technique called Canberra technique. Though this method does not suffer from the normality assumption of the data, however, their experiments showed that the technique performed very well only in the case where all the attacks were placed together. Ye et al. [26] proposed a multivariate quality control technique based on Hotelling's T test that detects both counterrelationship anomalies and mean-shift anomalies. When testing with a small set of data, all intrusions were detected with no false alarms; while for a large data set, 92% of intrusions were detected.

Many anomaly detection techniques employ the outlier detection concept. A detection technique that finds outliers by studying the behavior of the projections from the data set was discussed [1]. In [2], a degree of being an outlier called the local outlier factor (LOF) was assigned to each object. The degree depends on how isolated the object is with respect to the surrounding neighborhood. Lazarevic et al. [16] proposed several detection schemes for detecting network intrusions. A comparative study of these schemes on DARPA 1998 data set indicated that the most promising technique was the LOF approach [18].

In this paper, we propose a novel anomaly detection scheme based on principal components and outlier detection. The underlined assumption of the proposed method is that the attacks appear as outliers to the normal data. The principal component based approach has some advantages. First, it does not have any distributional assumption. Many statistical based intrusion detection methods assume a normal distribution or resort to the use of central limit theorem by requiring the number of features to be greater than 30 [25][26]. Secondly, it is typical for the data of this type of problem to be high dimensional. Hence, in our scheme, robust principal component analysis (PCA) is applied to reduce the dimensionality to arrive at a simple classifier which is the functions of some principal components. Since only a few parameters of the principal components need to be retained for future detection, the benefit is that the statistics can be computed in little time during the detection stage, which makes it possible to use the method in real time. Being an outlier detection method, the principal component classifier can find itself in many applications other than intrusion detection, e.g., fault detection, sensor detection, statistical process control, distributed sensor network, etc. Our experimental results show that the method has a good detection rate with a low false alarm, and outperforms the k-nearest neighbor method, the LOF approach, and the Canberra metric.

This paper is organized as follows. Section 2 provides the background on the concept of distance, PCA, and outlier detection. The proposed scheme is described in Section 3. Section 4 gives the details of the experiments followed by the results and the discussions in Section 5. We conclude our study in Section 6.

2. Multivariate Statistical Analysis

2.1. Distance

Many multivariate techniques applicable to anomaly detection problems are based upon the concept of distance. The most familiar distance metric is the Euclidean distance. It is frequently used as a measure of similarity in the nearest neighbor method. Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ and $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$ be two p-dimensional observations. The Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ is

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})} \tag{1}$$

Since each feature contributes equally to the calculation of the Euclidean distance, this distance is undesirable in many applications. When the features have very different variability or different features are measured on different scales, the effect of the features with large scales of measurement or high variability would dominate others that have smaller scales or less variability.

As an alternative, a measure of variability can be incorporated into the distance metric directly. One of this metric is the well-known Mahalanobis distance

$$d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x}-\mathbf{y})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y}) \tag{2}$$

where $\mathbf{S}$ is the sample covariance matrix.

Another distance measure that has been used in the anomaly detection problem is the Canberra metric. It is defined for nonnegative variables only.

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)} \tag{3}$$

2.2. Principal Component Analysis (PCA)

PCA is often used to reduce the dimension of data for easy exploration and further analysis. It is concerned with explaining the variance-covariance structure of a set of variables through a few new variables which are functions of the original variables. Principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$ with three important properties: (1) the principal components are uncorrelated, (2) the first principal component has the highest variance, the second principal component has the second highest variance, and so on, and (3) the total variation in all the principal components combined is equal to the total variation in the original variables $X_1, X_2, \ldots, X_p$. They are easily obtained from an eigenanalysis of the covariance matrix or the correlation matrix of $X_1, X_2, \ldots, X_p$ [13].

Principal components from the covariance matrix and the correlation matrix are usually not the same. In addition, they are not simple functions of the others. When some variables are in a much bigger magnitude than others, they will receive heavy weights in the leading principal components. For this reason, if the variables are measured on scales with widely different ranges or if the units of measurement are not commensurate, it is better to perform PCA on the correlation matrix.

Let $\mathbf{R}$ be a p x p sample correlation matrix computed from n observations on each of p random variables $X_1, X_2, \ldots, X_p$. If $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ are the p

eigenvalue-eigenvector pairs of $\mathbf{R}$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, then the i^{th} sample principal component of an observation vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ is

$$y_i = \mathbf{e}_i'\mathbf{z} = e_{i1}z_1 + e_{i2}z_2 + \ldots + e_{ip}z_p, \quad i = 1,2,\ldots,p \qquad (4)$$

where

$\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{ip})'$ is the i^{th} eigenvector

and

$\mathbf{z} = (z_1, z_2, \ldots, z_p)'$ is the vector of standardized observations defined as

$$z_k = \frac{x_k - \bar{x}_k}{\sqrt{s_{kk}}}, \quad k = 1,2,\ldots,p$$

where $\bar{x}_k$ and s_{kk} are the sample mean and the sample variance of the variable X_k.

The i^{th} principal component has sample variance λ_i and the sample covariance of any pair of principal components is 0. In addition, the total sample variance in all the principal components is the total sample variance in all standardized variables $Z_1, Z_2, \ldots, Z_p$, i.e.,

$$\lambda_1 + \lambda_2 + \ldots + \lambda_p = p \qquad (5)$$

This means that all of the variation in the original data is accounted for by the principal components.

2.3. Outlier Detection

Most data sets contain one or a few unusual observations. When an observation is different from the majority of the data or is sufficiently unlikely under the assumed probability model of the data, it is considered an outlier. With data on a single feature, unusual observations are those that are either very large or very small relative to the others. If the normal distribution is assumed, any observation whose standardized value is large in an absolute value is often identified as an outlier. With many features, the situation becomes complicated, however. In high dimensions, there can be outliers that do not appear as outlying observations when considering each dimension separately and therefore will not be detected from the univariate criterion. Thus, all features need to be considered together using a multivariate approach.

Let $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_n$ be a random sample from a multivariate distribution.

$$\mathbf{X}_j = (X_{j1}, X_{j2}, \ldots, X_{jp})', \quad j = 1,2,\ldots,n$$

The procedure commonly used to detect multivariate outliers is to measure the distance of each observation from the center of the data. If the distribution of $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_n$ is multivariate normal, then for a future observation $\mathbf{X}$ from the same distribution, the statistic T^2 based on the Mahalanobis distance

$$T^2 = \frac{n}{n+1}(\mathbf{X} - \overline{\mathbf{X}})'\mathbf{S}^{-1}(\mathbf{X} - \overline{\mathbf{X}}) \qquad (6)$$

is distributed as $\frac{(n-1)p}{n-p}F_{p,n-p}$, where

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j, \quad \mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})' \qquad (7)$$

and $F_{p,n-p}$ denotes a random variable with an F-distribution with p and n-p degrees of freedom [12]. A large value of T^2 indicates a large deviation of the observation $\mathbf{X}$ from the center of the population and the F-statistic can be used to test for an outlier.

Instead of the Mahalanobis distance, we can use other distance measures such as Euclidean distance and Canberra metric. Any observation that has the distance larger than a threshold value is considered an outlier. The threshold is typically determined from the empirical distribution of the distance. This is because the distributions of these distances are hard to derive even under the normality assumption.

PCA has long been used for multivariate outlier detection. Consider the sample principal components, y_1, y_2, …, y_p, of an observation $\mathbf{x}$. The sum of the squares of the standardized principal component scores,

$$\sum_{i=1}^{p}\frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \ldots + \frac{y_p^2}{\lambda_p} \qquad (8)$$

is equivalent to the Mahalanobis distance of the observation $\mathbf{x}$ from the mean of the sample [11].

It is customary to examine individual principal components or some functions of the principal components for outliers. Graphical exploratory methods such as bivariate plotting of a pair of principal components were recommended in [6]. There are also several formal tests, e.g., the tests based on the first few components [7]. Since the sample principal components are uncorrelated, under the normal assumption and assuming the sample size is large, it follows that

$$\sum_{i=1}^{q}\frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \ldots + \frac{y_q^2}{\lambda_q}, \quad q \leq p \qquad (9)$$

has a chi-square distribution with the degrees of freedom q. For this to be true, it must also be assumed that all the eigenvalues are distinct and positive, i.e., $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$. Given a significance level α, the outlier detection criterion is then

Observation $\mathbf{x}$ is an outlier if $\sum_{i=1}^{q}\frac{y_i^2}{\lambda_i} > \chi_q^2(\alpha)$

where $\chi_q^2(\alpha)$ is the upper α percentage point of the chi-square distribution with the degrees of freedom q. The value of α indicates the error or false alarm probability in classifying a normal observation as an outlier.

The first few principal components have large variances and explain the largest cumulative proportion of the total sample variance. These major components tend to be strongly related to the features that have relatively large variances and covariances. Consequently, the observations that are outliers with respect to the first few components usually correspond to outliers on one or more of the original variables. On the other hand, the last few principal components represent linear functions of the original variables with the minimal variance. These components are sensitive to the observations that are inconsistent with the correlation structure of the data but are not outliers with respect to the original variables [11]. The large values of the observations on the minor components will reflect multivariate outliers that are not detectable using the criterion based on the large values of the original variables. In addition, the values of some functions of the last r components, e.g., $\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$ and $\max_{p-r+1 \le i \le p} \left| \frac{y_i}{\sqrt{\lambda_i}} \right|$, can also be examined. They are useful in determining how much of the variation in the observation $\mathbf{x}$ is distributed over these latter components. When the last few components contain most of the variation in an observation, it is an indication that this observation is an outlier with respect to the correlation structure.

3. The Proposed Anomaly Detection Scheme

PCA has been applied to the intrusion detection problem as a data reduction technique, not an outlier detection tool. It is our interest to use PCA to identify attacks or outliers in the anomaly detection problem. Though graphical methods are effective in identifying multivariate outliers, particularly when working on principal components, they may not be practical for real time detection applications. Applying an existing formal test also presents a difficulty since the data need to follow some assumptions in order for the tests to be valid, e.g., the data have a multivariate normal distribution. Thus, we develop a novel anomaly detection scheme based on the principal components that can be applied in real time and does not impose too many restrictions on the data.

Following the anomaly detection approach, we assume that the anomalies are qualitatively different from the normal instances. That is, a large deviation from the established normal patterns can be flagged as attacks. No attempt is made to distinguish different types of attacks. To establish a detection algorithm, we perform PCA on the correlation matrix of the normal group. The correlation matrix is used because each feature is measured in different scales. It is important that the training data are free of outliers before they are used to determine the detection criterion because outliers can bring large increases in variances, covariances and correlations. The relative magnitude of these measures of variation and covariation has a significant impact on the principal component solution, particularly for the first few components. Therefore, it is of value to begin a PCA with a robust estimator of the correlation matrix. One simple method to obtain a robust estimator is multivariate trimming. First, we use the Mahalanobis metric to identify the 100γ% extreme observations that are to be trimmed. Beginning with the conventional estimators $\bar{\mathbf{x}}$ and $\mathbf{S}$, the distance $d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$ for each observation $\mathbf{x}_i$ (i=1,2,…,n) is computed. For a given γ (0.005 in our experiments), the observations corresponding to the $\gamma * n$ largest values of $\{d_i^2, i = 1,2,\ldots,n\}$ are removed. New trimmed estimators $\bar{\mathbf{x}}$ and $\mathbf{S}$ of the mean and the covariance matrix are computed from the remaining observations. A robust estimator of the correlation matrix is obtained using the elements of $\mathbf{S}$. The trimming process can be repeated to ensure that the estimators $\bar{\mathbf{x}}$ and $\mathbf{S}$ are resistant to outliers. As long as the number of observations remaining after trimming exceeds p (the dimension of the vector $\bar{\mathbf{x}}$), the estimator $\mathbf{S}$ determined by the multivariate trimming will be positive definite [11].

This robust procedure incidentally makes our method well suited for unsupervised anomaly detection. We cannot expect that the training data will always consist of only normal instances. Some suspicious data or intrusions may be buried in the data set. However, in order for the anomaly detection to work, we assume that the number of normal instances has to be much larger than the number of anomalies. Therefore, with the trimming procedure as described above, anomalies would be captured and removed from the training data set.

In our proposed scheme, the principal component classifier (PCC) consists of two functions of principal component scores, one from the major components $\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i}$ and one from the minor components $\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$. The first function that has been used in the literature is to detect extreme observations with large values on some original features. Different from other existing approaches, we propose the use of the second function in addition to the first one to help detect the observations that do not conform to the normal correlation structure. A clear advantage of this scheme over others is that it provides the information concerning the nature of the outliers whether they are extreme values or they do not have the same correlation structure as the normal instances.

The number of major components is determined from the amount of the variation in the training data that is accounted for by these components. Based on our experiments, we suggest using q major components that can explain about 50 percents of the total variation in the standardized features. When the original features are uncorrelated, each principal component from the correlation matrix has an eigenvalue equal to 1. So the r

minor components used in PCC are those components whose variances or eigenvalues are less than 0.20 which would indicate some relationships among the features.

The classification scheme using PCC goes as follows. Compute the principal component scores of the observation **x** for which the class is to be determined.

Classify **x** as an attack if

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > c_1 \quad \text{or} \quad \sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i} > c_2$$

Classify **x** as a normal instance if

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} \le c_1 \quad \text{and} \quad \sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i} \le c_2$$

where c_1 and c_2 are outlier thresholds such that the classifier would produce a specified false alarm rate.

Define $\alpha_1 = P\left(\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > c_1 \middle| \ \mathbf{x} \text{ is normal instance} \right)$

and $\alpha_2 = P\left(\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i} > c_2 \middle| \ \mathbf{x} \text{ is normal instance} \right)$.

Assuming the data are distributed as multivariate normal, the false alarm rate of this classifier is

$$\alpha = \alpha_1 + \alpha_2 - \alpha_1 \alpha_2 . \tag{10}$$

Under other circumstances, Cauchy-Schwartz inequality and Bonferroni inequality provide a lower bound and an upper bound for the false alarm rate α [15].

$$\alpha_1 + \alpha_2 - \sqrt{\alpha_1 \alpha_2} \le \alpha \le \alpha_1 + \alpha_2 \tag{11}$$

The values of α_1 and α_2 are chosen to reflect the relative importance of the types of outliers to detect. In our experiments, $\alpha_1 = \alpha_2$ is used. For example, to achieve 2% false alarm rate, Equation (10) gives $\alpha_1 = \alpha_2 = 0.0101$. Since the normality assumption is likely to be violated, we opt to set the outlier thresholds based on the empirical distributions of $\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i}$ and $\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$ in the training data rather than the chi-square distribution. That is, c_1 and c_2 are the 0.9899 quantile of the empirical distributions of $\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i}$ and $\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$, respectively.

4. Experiments

We study the performance of the PCC method by comparing it to the density-based local outliers (LOF) approach [2] and two other distance based intrusion detection methods: Canberra metric and Euclidean distance. The method based on the Euclidean distance is, in fact, the k-nearest neighbor method. We choose k=1 and 5 for the comparative study. The experiments are conducted under the following framework:

1) All the outlier thresholds are determined from the training data. We vary the false alarm rate from 1% to 10%. For the PCC method, the thresholds are chosen such that $\alpha_1 = \alpha_2$.
2) Both the training and testing data are from KDD'99 training data set.
3) Each training data set consists of 5,000 normal connections randomly selected by systematic sampling from all normal connections in the KDD'99 data.
4) To assess the accuracy of the classifiers, we carry out five independent experiments with five different training samples. In each experiment, the classifiers are tested with a test set of 92,279 normal connections and 39,674 attack connections randomly selected from the KDD'99 data.

4.1. The KDD'99 Data

KDD CUP 1999 data set [14] was used for the Third International Knowledge Discovery and Data Mining Tools Competition that was held in conjunction with The Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99). The contest task was to build a network intrusion detector from the data set, which is capable of distinguishing between "bad" connections (called attacks) and "good" normal connections. Three winning entries in this contest were [17][22][23]. The training data set contains 494,021 connection records, and the test data set contains 311,029 records that were not from the same probability distribution as the training data. Since the probability distributions were not the same, in our experiments, we sample data only from the training data set and use in both the training and testing stages.

A connection is a sequence of TCP packets containing values of 41 features and labeled as either normal or an attack, with exactly one specific attack type. There are 22 attack types in the training data. However, for the purpose of this study, we treat them the same as one attack group. The 41 features can be divided into three groups; the first group is the basic features of individual TCP connections, the second group is the content features within a connection suggested by domain knowledge, and the third group is the traffic features computed using a two-second time window. Among the 41 features, 34 are numeric and 7 are symbolic. Only the 34 numeric features are used in our experiments. A complete listing of features and details are in KDD CUP 1999 data [14].

4.2. Performance Measures

The result of classification is typically presented in a confusion matrix as shown in Table 1 [4]. The accuracy of a classifier is measured by its misclassification rate, or alternatively, the percentage of correct classification. Two other performance measures, precision and recall are also of interest [24].

Precision = TP/(TP+FP), Recall = TP/(TP+FN).

Another valuable tool for evaluating an anomaly detection scheme is the receiver operating characteristic (ROC) curve, which is the plot of the detection rate against the false alarm rate. The nearer the ROC curve of a scheme is to the upper-left corner, the better the performance of the scheme is. If the ROCs of different schemes are superimposed upon one another, then those schemes have the same performance [21].

Table 1. Confusion metrics for evaluations of attacks

		Predicted Connection Attack	Predicted Connection Normal
Actual Connection	Attack	Correctly detected (TP)	False negative (FN)
	Normal	False alarm (FP)	True negative (TN)

5. Experimental Results and Discussion

In an attempt to determine the appropriate number of major components to use in the PCC, we conduct a preliminary study by varying the percentage of total variation that is explained by the major components. A classifier of only the major components (r=0) is used.

Classify **x** as an attack if $\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > c$

Classify **x** as normal if $\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} \leq c$

where c is the outlier threshold corresponding to the desired false alarm rate.

Table 2 shows the detection rates from five classifiers with different numbers of the major components. The components account for 30% up to 70% of the total variation. We observe that as the percentage of the variation explained increases, which means more major components are used, the detection rate tends to be higher except for the false alarm rates of 1-2%. The PCC based on the major components that can explain 50% of the total variation is the best for a low false alarm rate, and it is adequate for a high false alarm rate as well. This suggests the use of q = 5 major components that can account for about 50% of the total variation in the PCC method.

Table 2. Detection rates of five PCCs at different false alarm rates

False Alarm	PC 30%	PC 40%	PC 50%	PC 60%	PC 70%
1%	67.12%	93.68%	97.25%	94.79%	93.90%
2%	68.97%	94.48%	99.05%	98.76%	96.07%
4%	71.07%	94.83%	99.23%	99.24%	99.24%
6%	71.79%	94.91%	99.33%	99.45%	99.44%
8%	75.23%	98.85%	99.34%	99.49%	99.58%
10%	78.19%	99.26%	99.35%	99.53%	99.65%

We now compare the performance of the PCC with both the major and minor components to other methods. The detection rates of five detection methods at different false alarm levels are presented in Table 3. The results are the average of five independent experiments. The standard deviation indicates how much the detection rate can vary from one experiment to another. As seen from the table, the results of some methods vary wildly, e.g., when the false alarm is 6%, the NN method (k=1) has 9.68% standard deviation, and the detection rate from the 5 experiments ranges from 70.48% to 94.58%.

Table 3. Average detection rates of five anomaly detection methods (Standard deviation of the detection rate is shown in the parenthesis)

False Alarm	PCC	Canberra	NN	KNN k=5	LOF
1%	98.94% (±0.20%)	4.12% (±1.30%)	58.25% (±0.19%)	0.60% (±0.00%)	0.03% (±0.03%)
2%	99.14% (±0.02%)	5.17% (±1.21%)	64.05% (±3.58%)	61.59% (±4.82%)	20.96% (±10.90%)
4%	99.22% (±0.02%)	6.13% (±1.14%)	81.30% (±8.60%)	73.74% (±3.31%)	98.70% (±0.42%)
6%	99.27% (±0.02%)	11.67% (±2.67%)	87.70% (±9.86%)	83.03% (±3.06%)	98.86% (±0.38%)
8%	99.41% (±0.02%)	26.20% (±0.59%)	92.78% (±9.55%)	87.12% (±1.06%)	99.04% (±0.43%)
10%	99.54% (±0.04%)	28.11% (±0.04%)	93.96% (±8.87%)	88.99% (±2.56%)	99.13% (±0.44%)

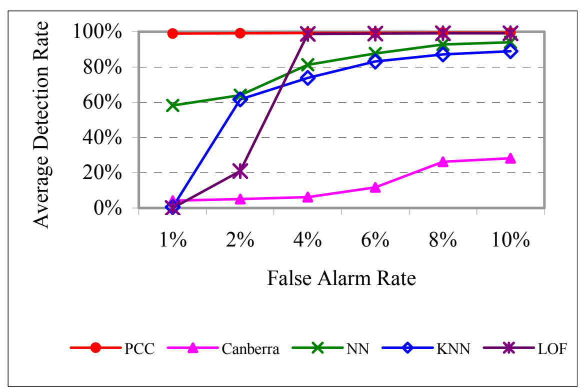

Figure 1. ROC curves of five detection methods

In general, the Canberra metric performs poorly. This result is consistent to Emran and Ye [5] that it does not perform at an acceptable level. The PCC has a detection rate about 99% with a very small standard deviation at all false alarm levels. It outperforms all other methods as easily seen from the ROC curves in Figure 1. It is the only method that works well at low false alarm rates.

Since the detection rate depends on the outlier threshold which is determined by the specified false alarm level, it is interesting to see what false alarm rate is actually attained when PCC is applied. As Table 4 shows, PCC has false alarm rates lower than the specified value, while the detection rate reaches almost perfection. Table 5 presents the average precision and recall values of PCC from 5 experiments when the false alarm is fixed at 1%. PCC clearly has high precision and recall values. It achieves 98.94% in recall and 97.89% in precision, while maintaining the false alarm rate at 0.92%. It also has a good balance of these two measures.

Table 4. Observed false alarm rate of PCC from 92,279 normal connections

Specified False Alarm	Observed False Alarm
1%	0.92%
2%	1.92%
4%	3.92%
6%	5.78%
8%	7.06%
10%	8.49%

Table 5. Average precision and recall of PCC (Fixed 1% false alarm)

Actual	Predicted		Recall
	Attack	Normal	
Attack	39,254	420	98.94%
Normal	848	91,431	99.08%
Precision	97.89%	99.54%	

In KDD'99 training data, there are 24 attack types that fall into 4 big categories: DOS – denial-of-service, Probe – surveillance and other probing, u2r – unauthorized access to local superuser (root) privileges, and r2l – unauthorized access from a remote machine. A detailed analysis of the detection results indicates that a large number of attacks can be detected by both major and minor components, some can only be detected by either one of them, and a few are not detectable at all since those attacks are not qualitatively different from the normal instances. An example is some attack types in category Probe. The detection rate in this category is not high, but it does not hurt the overall detection rate due to a very small proportion of this class in the whole data set, 414 out of 39,674 connections. We use the Probe group to illustrate advantages of incorporating minor components in our detection scheme.

Figure 2 gives detailed results of how the major components and minor components alone perform as compared to the combination of these two in PCC. In general, for this attack category, the minor component function gives a better detection rate than the major component function does. Many more attacks are detected by the minor components but would otherwise be ignored by using the major components alone. Hence, the use of the minor function improves the overall detection rate for this group. These experimental results show that our anomaly detection scheme based on the principal components works effectively in identifying the attacks. The only comparable competitor in our study is the LOF approach, but only when the false alarm rate is 4% or higher. Our proposed scheme has not only good precision and recall, but also the ability to maintain the false alarm at the desired level.

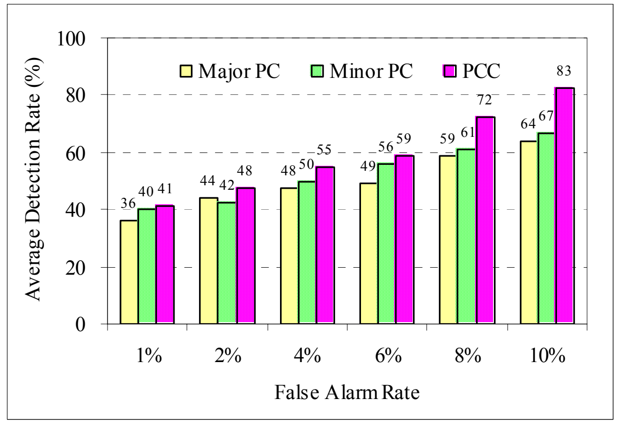

Figure 2. Average detection rates in Probe attack type by PCC and its major and minor components

As noted earlier, the sum of the squares of all standardized principal components $\sum_{i=1}^{p} \frac{y_i^2}{\lambda_i}$ is basically the Mahalanobis distance. By using some of the principal components, the detection statistic would have less power. However, in the experiments with the KDD'99 data, PCC has sufficient sensitivity to detect the attacks. Also, unlike the Mahalanobis distance, PCC offers more information on the nature of attacks from the use of two different principal component functions. One more benefit of PCC is that during the detection stage, the statistics can be computed in less amount of time, which makes it possible to use the method in real time. This is because only one third of the principal components are used in PCC, 5 major principal components which explain 50% of the total variation in 34 features and 6-7 minor components that have eigenvalues less than 0.20.

6. Conclusions

In this paper, we study the use of robust PCA in outlier detection and apply it to the anomaly detection problem. The predictive model is developed from two functions of the principal components of normal connections, which include the major principal components that explain about 50% of the total variation and the minor components whose eigenvalues are less than 0.20. A benefit of this approach is its ability to distinguish the nature of the anomalies whether they are different from the normal instances in terms of extreme values or different correlation structures. The experiments with the KDD'99 data indicate that the proposed anomaly detection scheme performs better than other techniques. The performance is consistently good regardless of the specified false alarm rates. It actualizes the detection rate close to 99% for the false alarm rate as low as 1%. With its robustness feature, our proposed scheme will also work with unsupervised training data.

7. Acknowledgements

For Mei-Ling Shyu, this research was supported in part by NSF ITR (Medium) IIS-0325260. For Shu-Ching Chen, this research was supported in part by NSF EIA-0220562, NSF HRD-0317692, and the office of the Provost/FIU Foundation.

8. References

[1] C. C. Aggarwal and P. S. Yu, "Outlier Detection for High Dimensional Data" *Proceedings of the ACM SIGMOD Conference,* Santa Barbara, CA, May 21-24, 2001.

[2] M. M. Breunig, H-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," *Proceedings of the ACM SIGMOD Conference,* Dallas, TX, May 16-18, 2000.

[3] D. E. Denning, "An Intrusion Detection Model," *IEEE Transactions on Software Engineering*, SE-13, pp. 222-232, 1987.

[4] P. Dokas, L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava, and P-N. Tan, "Data Mining for Network Intrusion Detection," *Proceedings of National Science Foundation Workshop on Next Generation Data Mining* , November 1–3, 2002.

[5] S. M. Emran and N. Ye, "Robustness of Canberra Metric in Computer Intrusion Detection," *Proceedings of the 2001 IEEE Workshop on Information Assurance and Security, United States Military Academy, West Point, NY,* June 5-6, 2001.

[6] R. Gnanadesikan and J. R. Kettenring, "Robust Estimates, Residuals and Outlier Detection with Multiresponse Data," *Biometrics*, 28, 81-124, 1972.

[7] D. M. Hawkins, "The Detection of Errors in Multivariate Data Using Principal Components," *Journal of the American Statistical Association*, Vol. 69, No. 346, 340-344, 1974.

[8] R. Heady, G. Luger, A. Maccabe, and M. Servilla, "The architecture of a network level intrusion detection system," *Technical report, Computer Science Department, University of New Mexico*, August 1990.

[9] W. Hu, Y. Liao, and V. R. Vemuri, "Robust Support Vector Machines for Anomaly Detection in Computer Security," *Proceedings of the 2003 International Conference on Machine Learning and Applications (ICMLA),* Los Angeles, CA, June 23-24, 2003.

[10] H. S. Javitz and A. Valdes, "The SRI Statistical Anomaly Detector," *Proceedings of the 1991 IEEE Symposium on Research in Security and Privacy*, May 1991.

[11] J. D. Jobson, "Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods," *Springer-Verlag, NY, 1992.*

[12] R. A. Johnson and D. W. Wichern, "Applied Multivariate Statistical Analysis," 4th Ed., *Prentice-Hall, NJ, 1998.*

[13] I. T. Jolliffe*, "*Principal Component Analysis*,"* 2nd Ed., *Springer-Verlag, NY, 2002.*

[14] KDD Cup 1999 Data, Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, August 2003.

[15] M. G. Kendall, A. Stuart, and J. K. Ord, "Kendall's Advanced Theory of Statistics V. 1 Distribution Theory," 5th Ed., *Oxford University Press, NY, 1987.*

[16] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection," *Proceedings of the Third SIAM Conference on Data Mining*, May 2003.

[17] I. Levin, "KDD-99 Classifier Learning Contest: LLSoft's Results Overview", *ACM SIGKDD Explorations2000*, pp. 67-75, January 2000.

[18] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. P. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and M. A. Zissman, Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation, *Proceedings DARPA Information Survivability Conference and Exposition (DISCEX) 2000*, Vol. 2, pp. 12-26, IEEE Computer Society Press, CA, 2000.

[19] M. Markou and S. Singh, "Novelty Detection: A Review Part1: Statistical Approaches," *Signal Processing,* (under submission, 2003), Available on http://www.dcs.ex.ac.uk/research/pann/pdf/pann_SS_086.PDF, August, 2003.

[20] M. Markou and S. Singh, "Novelty Detection: A Review Part2: Neural Network-based Approaches," *Signal Processing,* (under submission, 2003), Available on http://www.dcs.ex.ac.uk/research/pann/pdf/pann_SS_087.PDF, August, 2003.

[21] R. A. Maxion and K. M. C. Tan, "Benchmarking Anomaly-Based Detection Systems," *1st International Conference on Dependable Systems & Networks,* pp. 623-630, June 2000.

[22] B. Pfahringer, "Winning the KDD99 Classification Cup: Bagged Boosting," *ACM SIGKDD Explorations 2000*, pp. 65-66, January 2000.

[23] M. Vladimir, V. Alexei, and S. Ivan, "The MP13 Approach to the KDD'99 Classifier Learning Contest," *ACM SIGKDD Explorations 2000*, pp. 76-77, January 2000.

[24] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization," *Kluwer Academic Publishers, Netherlands, 1999.*

[25] N. Ye and Q. Chen, "An Anomaly Detection Technique Based on a Chi-Square Statistic for Detecting Intrusions into Information Systems," *Quality and Reliability Eng. Int'l*, Vol. 17, No. 2, pp. 105-112, 2001.

[26] N. Ye, S. M. Emran, Q. Chen, and S. Vilbert, "Multivariate Statistical Analysis of Audit Trails for Host-Based Intrusion Detection," *IEEE Transactions on Computers,* Vol. 51, No. 7, July 2002.