

A PCA-based Method for IoT Network Traffic Anomaly Detection

Dang Hai Hoang*, Ha Duong Nguyen**

*Posts and Telecommunication Institute of Technology, Hanoi, Vietnam

**Faculty of Information Technology, National University of Civil Engineering, Hanoi, Vietnam

hdhai.hn@gmail.com, nghaduong@gmail.com

Abstract— Anomaly detection in IoT (Internet of Things) networks is becoming a challenge task due to limited network resources and performance. Comprehensive detection methods are no longer effective for IoT networks, calling for developing lightweight solutions. Methods using Principal Component Analysis (PCA) is an attractive approach due to complexity reduction. However, there are remaining issues by applying PCA such as the choice of principal components for complexity reduction. This paper investigates PCA techniques used in previous typical research works and proposes a new general formula for distance calculation and a new detection method based on PCA for IoT networks. The paper investigates formula parameters for the new method and presented the quick detection of network traffic anomalies with lower complexity.

Keywords— IoT Network Traffic Anomaly, Anomaly Detection, Principal Component Analysis, Information Security, Network Security

I. INTRODUCTION

Internet of Things (IoT) is the world of interconnected things, where the humans are interacting with the machines and machines are talking with each other. As indicated in [1,2], there is a number of challenges relating to IoT security. Network traffic anomaly detection (NTAD) can be used for identifying network attacks. Anomalies are unusual patterns in traffic data that do not conform to expected normal behaviour. A formal definition was given by Hawkins as indicated in [3,4]. Anomalies may be performance related (due to network failures, changes in link traffic, flash crowd, etc...) or security related (due to attacks such as denial of service attacks, network scans, etc...). In both cases, anomaly detection (AD) is a very critical task of network operators. Especially, network operators need an effective tool to quickly detect exceptional non-conforming patterns in traffic data in order to identify abnormal traffic flows or the causes of anomalies for further handling.

Many AD techniques have been proposed in the past. There are several good surveys in the literature [3-9]. Traditional outlier detection concept can apply for anomaly detection as indicated in [3-7, 10, 11], where anomalies could be seen as outlier. Typically, these methods use statistical properties of observed data to construct a normal profile based on normal activities of the system. The current activity pattern will be compared to this profile for checking any deviation to find

anomalies. Generally speaking the main principle of NTAD is as follows: identifying features of network traffic in normal activity condition, building the baseline (the normal region) and determining the deviation (anomalies) by comparing live traffic data to this baseline. In practice, it is difficult to determine a precise boundary between normal and abnormal behaviour. Other difficulties are: multivariate features of traffic data, correlation of various data features, complexity due to high dimension of traffic dataset, detection accuracy. Anomaly detection, especially quick online anomaly detection remains a challenging research task.

In contrast to traditional Internet, AD in IoT networks is becoming a challenge task due to constrained resources and limited performance. In IoT networks, we can not use complex methods. In contrast, we can only use lightweight solutions. Among existing AD techniques [3-10], multivariate statistical approaches for AD received a lot of attention. Principal Component Analysis (PCA) is one of the best-known techniques. PCA is a dimension reduction technique, which transforms a set of correlated original variables into a set of few uncorrelated variables, called Principal Components (PC). These PCs are linear combinations of the original variables. The number of PCs is less than or equal to the number of original variables. Thus, PCA allows lower complexity. PCA is considered as a simple but effective method for NTAD [4, 8-18]. Thus, this paper focuses on NTAD using PCA (PCA-NTAD) for IoT networks with limited resources.

Pioneer research works for NTAD using PCA were given by Shyu [5], Lakhina [12], Kwitt [19], Ringberg [14]. A variety number of further works has been proposed with several enhancements such as [8-10,15-28]. However, there are remaining issues of PCA as described in [3,4,6,9,15] such as sensitivity, feature selection, effectiveness, dimension-depending computation complexity. Two main issues for the goodness of PCA-NTAD are: how many PCs are to select and how to calculate the distance (the deviation measure).

The choice of PCs was addressed in many works, e.g. the selection of major PCs and minor PCs [11,19,22,17,10], the selection of PC subspace [13-16,18,27-29]. However, heuristic is the common way of choosing PCs. It is not clear, how to select many PCs and which PCs. On the other hand, the choice of distance formula is not clear in the past works. Distance is a quantitative metric for deviation of traffic

pattern. Most of the past works proposed either T^2 distribution formula [3,4,5,9-12], or Euclidean distance [6,10,24,25], or Mahalanobis distance [6,9-11,13-17,19-21,23-28].

This paper addresses PCA-NTAD in IoT networks. We propose a new distance formula for NTAD and indicate that our formula can be used to derive formulas in most of previous works. To our best knowledge, the proposed formula is the first one for interpreting the parameters in different distance formulas. We propose a PCA-NTAD scheme based on this new formula. We investigate issues of NTAD with various parameters and show that our scheme has simple formulas while retaining acceptable detection rate. Only few PCs are needed for quick detection.

II. NETWORK TRAFFIC ANOMALY DETECTION USING PCA

A. The Basic Concept of PCA

AD often requires a high dimension data including many features (attributes). The analysis has high complexity, needs much time, and is not suitable for quick detection. PCA is the most common technique to reduce high dimension [3-7]. It converts the original data into new set of axes called PCs by keeping the most essential features of the original data.

Let X be the original observed dataset with n rows and p columns $\{X_1, X_2, \dots, X_p\}$. The dataset is a $n \times p$ matrix. Each column represents an attribute of data, i.e. a p -dimensional vector of p correlated variables. Let R be a $p \times p$ sample correlation matrix of $\{X_1, X_2, \dots, X_p\}$. Let $(\lambda_1, e_1), (\lambda_2, e_2), (\lambda_3, e_3), \dots, (\lambda_p, e_p)$ be p eigenvalue and eigenvector pairs of the matrix R . PCA converts X into a new dataset Y using transformation matrix R as follows:

$$Y = RX \quad (1)$$

We have the i^{th} PC as follows:

$$y_i = e_i^T (x - \bar{x}), \text{ where } i = 1, 2, \dots, p. \quad (2)$$

$e_i = (e_{i1}, e_{i2}, \dots, e_{ip})^T$ is the i^{th} eigenvector.

$\bar{x} = (\sum_{i=1}^n x_i) / n$ is the sample mean of the observation x .

The eigenvalues (λ) are the roots of equation $|R - \lambda I| = 0$.

Each eigenvalue has a corresponding non-zero eigen-vector e , which satisfies: $Re = \lambda e$

Let $z = (z_1, z_2, \dots, z_p)^T$ be the vector of standardized observations, i.e. $z = x - \bar{x}$, we rewrite (2) as:

$$y_i = e_i^T z \quad (3)$$

Important properties of PCA are: PCs are uncorrelated; the first PC has the highest variance; the second PC has the next highest variance, and so on. The total variance in all PCs combined is equal to the total variance of original variables $X_1, X_2, \dots, X_p$. The PCs are sorted in descending order of the eigenvalues, $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_p > 0$. The quotient $\lambda_i / \sum_{j=1}^p \lambda_j$ describes the contribution of the i th PC on the variance of the data.

B. Anomaly Detection Using PCA

The common principle for PCA-NTAD is to calculate the statistical distance from each observed data to the normal dataset (i.e. centroid or statistical average). The most popular distance formulas are Euclidean or Mahalanobis distance or statistic T^2 distribution. The distance calculation is performed

in the PC space. Mathematically, distance is a quantitative degree of how far two data instances apart from each other. Observed instances which are at far distance from the new axes represented by PCs are considered as abnormal behaviour. A threshold value is established for the comparison. If the calculated distance of the observed data is larger than the threshold, this data instance is considered as an anomaly. The threshold value is usually determined by the statistical distribution function of the distance [3-5, 29].

1) Euclidean Distance: Let $x = (x_1, x_2, \dots, x_p)$ and $\mu = (\mu_1, \mu_2, \dots, \mu_p)$ be the observed data, where p is the number of original attributes, i.e. the number of input variables. The Euclidean distance between x and μ is [6, 10]:

$$d(x, \mu) = \sqrt{(x - \mu)^T (x - \mu)} \quad (4)$$

If μ is the vector of mean values of each input variable, the Euclidean distance between x and μ in the PCA space is:

$$\Delta_{Euclid} = \sqrt{\sum_{i=1}^p y_i^2} \quad (5)$$

The square distance is usually used, thus

$$d_E = \Delta_{Euclid}^2 = \sum_{i=1}^p y_i^2 \quad (6)$$

2) Mahalanobis Distance: The Euclidean distance does not take into account the variance of each variable. Thus, the Mahalanobis distance was proposed to consider average value, variance and covariance of variables. Instead of simply computing the distance from mean value, it weights each variable by its standard deviation and covariance [24]. The Mahalanobis distance between two observed data x and μ is computed as [6, 9-11, 13-17, 19-21, 24-28]:

$$d(x, \mu) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)} \quad (7)$$

Where S^{-1} is the transposition of the sample covariance matrix between x and μ . In PCA space, the Mahalanobis distance is computed as:

$$\Delta_{Mahalanobis} = \sqrt{\sum_{i=1}^p y_i^2 / \lambda_i} \quad (8)$$

The square distance is usually used, thus:

$$d_m = \Delta_{Mahalanobis}^2 = \sum_{i=1}^p y_i^2 / \lambda_i \quad (9)$$

3) Hotelling's T^2 Statistic: This statistic gives the deviation from each vector x to the centroid μ of the statistical distribution of the data [9-12, 29].

$$T^2 = (x - \bar{x})^T S^{-1} (x - \bar{x}) \quad (10)$$

The statistic T^2 is similar to the Mahalanobis distance.

$$T^2 = \sum_{i=1}^p y_i^2 / \lambda_i \quad (11)$$

T^2 is distributed as an F -distribution with p and $n-p$ degrees of freedom. A large value of T^2 indicates a deviation of the observation X from the centroid of the normal dataset and the F -statistic can be used to test for an anomaly [11, 29].

C. Issues of Previous PCA Based Methods

In principle, tests for anomaly can be applied on individual PCs. However, standard PCA based techniques are typically proportional quadratic in the number of variables regarding the distance calculation. Many works tried using a subset of PCs in order to reduce the computation complexity [5-10, 22, 25-28].

PCA transformation is performed in such a way that the first PCs account for the most of the variance in the original data, and the last PCs represent linear functions of the original variables with very small variance. The first PCs represent the large cumulative proportion of the total variance. Thus, these PCs tend to be strongly related to the anomalies on one or more original variables. Intuitively, it is reasonable to look at few first PCs [17]. If most of the variance of an n -dimensional data set is accounted by $k < p$ principal components, it is possible to select only first k PCs. The dimension of the data is then reduced to k [25]. It is still not clear how many PCs needed for an effective detection.

On the other hand, last PCs are sensitive to the observation that are inconsistent with the correlation of data [11-15]. Large values on few last PCs reflect multivariate anomalies that are not detectable using the criterion based on large values of the original variables [9]. It is useful to check the last PCs for a significant variation of an observation. That is why many works suggested using few first PCs and few last PCs.

Typically, two thresholds are used: one for major PCs (q first PCs) and one for minor PCs (last r PCs). As indicated in [11,16, 19,20], a data instance x is anomalous if

$$\sum_{i=1}^q y_i^2 / \lambda_i > c_1 \text{ or } \sum_{i=p-r+1}^p y_i^2 / \lambda_i > c_2 \quad (12)$$

and is normal if

$$\sum_{i=1}^q y_i^2 / \lambda_i \leq c_1 \text{ and } \sum_{i=p-r+1}^p y_i^2 / \lambda_i \leq c_2 \quad (13)$$

c_1 and c_2 are the thresholds, which are calculated with chi-square distribution. They are very sensitive to detection results.

Several works proposed the subspace approach based on decomposition of the data into normal and anomalous subspaces using projection of data on the first PCs and the last PCs respectively [12,23,25]. The subspace methods are also called residual analysis methods. One issue with this approach is that PCA based anomaly detection techniques are sensitive to the number of PCs in each of the two subspaces [15,4,6].

Another issue of PCA-NTAD is the distance formula. As presented in the above section, most of the proposed methods are using either Euclidean distance [6,10,16,23-25] or Mahalanobis distance [3-5,9,10,15,19,20,25,24,29]. The computation of the distance is quadratic proportional to of the variables. With high dimensional dataset, the distance calculation will increase the computational cost.

In summary, there are still several issues by applying PCA for traffic anomaly detection, especially: 1) selection of PCs, 2) distance measure and 3) quick detection of anomalies. These issues are the topics of this paper.

III. RELATED WORKS

The pioneering works by Shyu [11] and Lakhina [12] introduced the use of PCA for NTAD. The authors in [11] proposed a principal component classifier (PCC) consisting of two functions of PC scores, one for major PCs and one for minor PCs. The major PCs are used to detect extreme observations with large values on some original variables. The minor PCs are for observations that do not conform to the normal correlation structure. The Mahalanobis distance uses two thresholds. Further works such as [19, 20] suggested

improving PCC with multivariate trimming for robustness, threshold calculation. Lakhina [12] divided network traffic data into normal and anomalous subspace. Similar approaches [16,25] proposed traffic matrix consisting of original/destination flows (OD flows). The Euclidean distance is employed to check the dissimilarity between data instances. AD is using the first PCs and the mean value of the distance. The authors in [15] investigated issues of [12] and proposed a control approach for pre-processing the network data. [17] improved the work of [12]. They decomposed traffic variations into normal and anomalous components for identifying anomalous flows inside the aggregates.

A variety number of research works followed the approach of [11,12]. The authors in [13] presented limitations of PCA such as temporal correlation. They proposed an expansion based on a predictive filter. [14] indicated the sensitivity of PCA-NTAD. PCA is very sensitive to its parameters. Current methods for tuning PCA are inadequate. [8] presented the application of PCA subspace method for AD in backbone networks with a lower approximation. [21] discussed category-based selection of the features by classifying network traffic into six typical groups in order to reduce the amount of data. [25] presented three major approaches: PCA-based, sketch-based and signal-analysis-based. [22] proposed an online over-sampling PCA algorithm. [16] proposed a sketch-based algorithm, in which each local monitor maintains a series of sketches to reduce the raw data. [9] suggested an integration of PCA, Hotelling's T^2 and Q statistics. The authors in [5] provided a survey on outlier detection techniques and applications. [10] discussed three statistical techniques: PCA, Chi-square and Gaussian mixture distribution. [18] presented a multi-step outlier-based approach.

The choice of a reasonable number of PCs is important for alleviate the autocorrelation of the residual PCs. The authors [6, 9,24-28] concluded that the goodness of PCA model depends on a good choice of how many PCs are retained. The paper [9, 27] indicated techniques such as Scree plot and cumulative percentage variance (CPV) for determining PCs. Nevertheless, a threshold value of cumulative variance is needed to determine the number of PCs. The paper [7] provided a broad survey of the existing research on NTAD.

Anomalies are usually detected by the help of a distance measure [18]. A good survey was given in [6]. Two common distance measures including Euclidean [6,7,10,21,23,24,25] and Mahalanobis distance [9-11, 13-17,19,20,24-28] with some modifications. In our best knowledge, there is still no any work investigating PCA-NTAD for IoT networks.

IV. A NEW METHOD FOR NETWORK TRAFFIC ANOMALY DETECTION BASED ON PCA

A. The New Distance Formula

For quick online AD, we suggest to use only few PCs and a simple distance formula as possible. Intuitively, the price to be paid is the accuracy. However, AD can be implemented into two steps. At the first step, network operators only require to know whether there are abnormal traffic flows or not. A quick

detection with acceptable accuracy is desirable. At the next step, a detailed analysis is applied for higher accuracy.

By keeping this in mind, we develop a new calculation formula using the well-known Minkowski formula. This formula helps to investigate various formulas used in past works and to compare their accuracy. The Minkowski distance between two observed data $\mathbf{x} = (x_1, x_2, \dots, x_p)$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)$ is as follows [30]:

$$\Delta_{\text{Minkowski}} = \left(\sum_{i=1}^p |x_i - \mu_i|^c \right)^{1/c}, c \geq 1; \quad (14)$$

$$d = \Delta_{\text{Minkowski}}^c = \sum_{i=1}^p |x_i - \mu_i|^c, c \geq 1; \quad (15)$$

In PCA subspace, Minkowski distance from each observation to the centre (the origin of PC axes) is:

$$d = \sum_{i=1}^p |y_i|^c, c \geq 1; \quad (16)$$

Since each variable has different variance, we can follow the approach of weighted Euclidean distance as in [29, 15] to give a weight to each PC in the equation (16) as follows:

$$d = \sum_{i=r}^q w_i |y_i|^c, c \geq 1; \quad (17)$$

Where w_i is the weight for the i^{th} PC ($w_i \neq 0$) and $1 \leq r \leq q \leq p$.

The formula (17) is our new distance formula which has the following properties:

- By choosing w_i and c , we can flexible set up the parameters for the distance formula (17).
- The contribution of each PC can be controlled using suitable choice of parameters.
- Other formulas such as Euclidean, Mahalanobis can be derived from (17) with corresponding parameters.

If $c=2$, from (17) we get the weighted Euclidean distance:

$$d = \sum_{i=r}^q w_i y_i^2 \quad (18)$$

If $c=2$, $w_i=1$, we get the squared Euclidean distance:

$$d = \sum_{i=r}^q y_i^2 = y_r^2 + y_{r+1}^2 + \dots + y_q^2 \quad (19)$$

If $c=2$, $w_i=1/\lambda_i$, we get the squared Mahalanobis distance (equal to the squared Euclidean distance with weights):

$$d = \sum_{i=r}^q \frac{y_i^2}{\lambda_i} = \frac{y_r^2}{\lambda_r} + \frac{y_{r+1}^2}{\lambda_{r+1}} + \dots + \frac{y_q^2}{\lambda_q} \quad (20)$$

It is the same distance formula used in past works [6, 9-11, 13-17, 19-21, 24-28]. If $c=1$, from (17) we get:

$$d = \sum_{i=r}^q w_i |y_i| \quad (21)$$

If $c=1$, $w_i=1$, we get the Manhattan distance:

$$d = \sum_{i=r}^q |y_i| = |y_r| + |y_{r+1}| + \dots + |y_q| \quad (22)$$

B. A New Method for Network Traffic Anomaly Detection

The figure 1 depicts the general model of our anomaly detection scheme using the new distance formula.



Figure 1. The Model of the Anomaly Detection Scheme

The captured data is standardized and transformed into PCA. After this transformation, we get eigenvector and eigenvalue pairs together with p PCs of each dataset.

By transforming into PCA space, we classify PCs into m major PCs and $p-m$ minor PCs similar to the approach in [11, 16-21, 23- 26]. As suggested in [11, 25], the major PCs include the PCs representing 50% of the variance of total eigenvalues. The minor PCs have eigenvalues smaller than or

equal to 0.2.

For distance calculation, we apply the formula (17) with $k = p-m$ minor PCs where $1 < m < p$. An observed dataset is considered as anomaly if $d > d_N$ where d_N is the corresponding thresholds of d . For determining the thresholds, we use the empirical cumulative distribution function (ECDF) that is defined as follows:

$$F_{\text{ECDF}} = P(d \leq d_N) \quad (23)$$

With the assumption of statistical distribution and an estimation rate α for false estimation, we can look at the table for F_{ECDF} -distribution function and seek for the values of d_N corresponding to $(1-\alpha)$ of the ECDF. For instance, if $\alpha = 5\%$, the thresholds d_N is determined corresponding to 95% of ECDF.

V. EXPERIMENTS

A. Dataset and Metrics

The Kyoto 2006+ dataset [31] is a real dataset collected from networks. The dataset includes most of the anomalies originated by the Internet. Thus, this dataset reflects objectively the anomalous events of collected network traffic.

We use in our experiments 14 essential features of the Kyoto 2006+ dataset including the most important features of the network layer and transport layer traffic data. These features are selected similar to the previous works [8,11-20,22-26] for comparison purpose.

For the experiments, we select three random datasets of the Kyoto 2006+ datasets collected from networks as indicated in the Table 1. Following parameters are used for the accuracy:

$$TPR = TP / (TP + FN) \quad (24)$$

$$FPR = FP / (TN + FP) \quad (25)$$

TPR is True Positive Rate, FPR is False Positive Rate, TP is true positive alerts, FN is false negative alerts, FP is false positive alerts, and TN is true negative alerts. The total number of anomalous events: $TP+FN$; normal events: $TN+FP$.

TABLE 1. THREE RANDOM DATASETS FOR EXPERIMENTS

Dataset	Number of flows	Number of anomalous flows	Number of normal flows
Dataset 1	125643	84476	41167
Dataset 2	114148	38790	75358
Dataset 3	120857	57337	63520

B. Experiment Results

To build the sample dataset, we use traffic data from 5000 connections (similar to [11]). We transform this dataset into PCA space, calculate the centroid of data using the distance formula. We determine the threshold using ECDF as described. In the next step, we use the collected Kyoto 2006+ datasets for AD applying to our model (Fig.1). The procedure for AD includes following steps: standardizing data, PCA transforming, distance calculation using c and w_i , comparing the distance to the thresholds. The thresholds are 95% of the ECDF function.

Table 2 shows the experiment result with dataset 1 using the threshold 95% for different combination of k (number of PCs), c and w_i . The table presents the results for three cases:

- Case 1: $c = 2$ and $w_i = 1$
- Case 2: $c = 2$ and $w_i = 1/\lambda_i$
- Case 3: $c = 1$ and $w_i = 1$

If $k = 3$, TPR and FPR for case 3 are better than case 1, approximate to case 2. If $k = 5$, three cases have similar TPR (91.8, 91.9 and 91.3), FPR of case 3 ($FPR = 5.6$) is a little bit better than case 1 and case 2 ($FPR = 5.2$). If $k = 14$, case 1 has better TPR ($TPR = 94.4$) than case 2 ($TPR = 94.3$) and case 3 ($TPR = 92.1$). But FPR of case 3 ($FPR = 5.4$) is better than case 1 ($FPR = 5.3$) and case 2 ($FPR = 5.0$).

Table 3 shows the number of true anomalous flows for case 1, 2 and 3 (dataset 1), respectively. In fact, the total number of anomalous flows in dataset 1 is 84476.

TABLE 2. RESULTS WITH DATASET 1, THRESHOLD 95%

Case	k	c	w_i	TPR (%)	FPR (%)
1	3	2	1	92.4	4.7
1	5	2	1	91.8	5.2
1	14	2	1	94.4	5.3
2	3	2	$1/\lambda_i$	94.6	4.9
2	5	2	$1/\lambda_i$	91.9	5.6
2	14	2	$1/\lambda_i$	93.8	5.0
3	3	1	1	94.3	4.8
3	5	1	1	91.3	5.2
3	14	1	1	92.1	5.4

TABLE 3. NUMBER OF TRUE DETECTED ANOMALY FLOWS (TP)

Case	Parameters	TP			
		$k=3$	$k=4$	$k=5$	$k=14$
1	$c=2, w_i=1$	78106	77430	77611	80082
2	$c=2, w_i=1/\lambda_i$	79951	78626	77633	79302
3	$c=1, w_i=1$	79783	77678	78465	78431

Fig. 2 shows the number of true detected anomalous flows in dataset 1 for three cases. If $k=3$, TP of case 2 and 3 are higher than TP of case 1. If $k=5$, TP of case 3 is higher than TP of case 1 and case 2. The works [6,10,16,23-25] use case 1, the works [6,9-21,24-28] use case 2. Case 3 is for our method.

Table 2 and Fig. 2 showed that the case 3 ($c=1, w_i=1$) provided acceptable TPR and FPR in comparison to the case 1 and case 2.

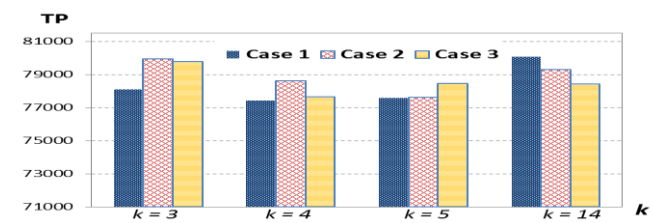


Figure 2. TP for Anomalous Flows in Dataset 1

Table 4 and Table 5 show the experiment results with dataset 2 and dataset 3, respectively. Threshold is 95%. The results are showed for different combination of the parameters k (number of PCs), c and w_i .

TABLE 4. RESULTS WITH DATASET 2, THRESHOLD 95%

Case	k	c	w_i	TPR (%)	FPR (%)
1	3	2	1	95.9	5.4
1	5	2	1	99.9	5.1

1	14	2	1	93.1	4.6
2	3	2	$1/\lambda_i$	93.8	4.7
2	5	2	$1/\lambda_i$	96.9	4.8
2	14	2	$1/\lambda_i$	99.0	4.9
3	3	1	1	95.9	5.0
3	5	1	1	92.3	4.7
3	14	1	1	93.8	5.1

Table 6 shows the number of normal flows that are true detected in dataset 2 for case 1, case 2 and case 3, respectively. The total number of actual normal flows in dataset 2 is 75358.

TABLE 5. RESULTS WITH DATASET 3, THRESHOLD 95%

Case	k	c	w_i	TPR (%)	FPR (%)
1	3	2	1	99.4	5.6
1	5	2	1	99.9	5.3
1	14	2	1	99.9	5.4
2	3	2	$1/\lambda_i$	99.7	4.9
2	5	2	$1/\lambda_i$	99.8	5.0
2	14	2	$1/\lambda_i$	99.9	5.3
3	3	1	1	98.6	4.9
3	5	1	1	97.6	5.3
3	14	1	1	100	4.9

TABLE 6. NUMBER OF TRUE DETECTED ANOMALY FLOWS (TP)

Case	Parameters	TN			
		$k=3$	$k=4$	$k=5$	$k=14$
1	$c=2, w_i=1$	71275	71385	71500	71886
2	$c=2, w_i=1/\lambda_i$	71820	71417	71668	71675
3	$c=1, w_i=1$	71586	71699	71767	71504

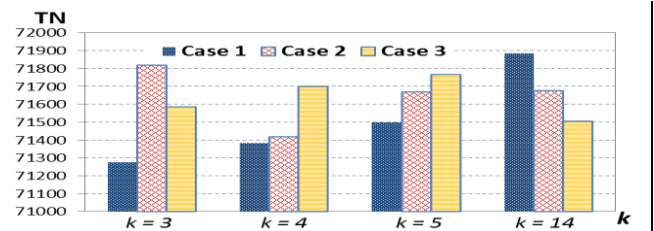


Figure 3. TN for Nomalous Flows in Dataset 2

Figure 3 shows the number of true detected normal flows in dataset 2 for three cases. As showed in the figure, if $k=3$, TN of case 2 is higher than TN of case 1 and case 3. If $k=4$ or $k=5$, TN of case 3 is higher than TN of case 1 and case 2. If $k=14$, TN of case 1 is higher than TN of case 2 and case 3. The reason is that the last PCs are responsible for the variance.

As showed in Table 4, although FPR of case 1 for $k=14$ ($FPR=5.4\%$) is higher than FPR of case 3 for $k=14$ ($FPR=4.9\%$), the true positive rate TPR of case 3 for $k=14$ ($TPR=100\%$) is higher than TPR of case 1 for $k=14$ ($TPR=99.7\%$).

Table 5 and Fig.3 showed that case 3 has comparable accuracy to those of case 1 and case 2. The experiments showed the use of the new method for quick traffic anomaly detection with the new formula and few minor PCs.

VI. CONCLUSIONS

NTAD is one of the challenging tasks of network operators.

Among many approaches, PCA-NTAD received a lot of attention. It is suitable for IoT networks with constrained network resources and limited performance. This paper discussed such problems of PCA-NTAD for IoT networks. By investigating various parameters, we proposed a new distance formula and showed that it can be used to derive formulas in some previous works. We then proposed a new method for NTAD in IoT networks. The paper presented various experiments to show the feasibility of the proposed method. We showed an acceptable detection rate using a simple distance formula to reduce the computation overhead.

ACKNOWLEDGMENT

This work is supported by the ASEAN IVO project "A Hybrid Security Framework for IoT Networks".

REFERENCES

- [1] A.V. Vijayalakshmi, L. Arockiam, "A Study on Security Issues and Challenges in IoT". Engineering Sciences & Management Research, Vol 3, No 11, Nov. 2016, pp. 34-43
- [2] Q.Jing, et al, "Security of the IoT: Perspectives and Challenges". Wireless Networks, Vol 20, Issue 8, Nov. 2014, pp.2481-2501
- [3] V. Chandola, A. Banerjee, V. Kumar, "Anomaly detection: a survey", ACM Computing Surveys, 2009, 41, (3), pp. 1-58.
- [4] M. Ahmed, A. Mahmood, J. Hu, "A survey of network anomaly detection techniques", Netw& CompApp, 2016, 60 (2016), pp. 19-31.
- [5] R. Bansal, et al, "Outlier detection: applications and techniques in data mining", Cloud Syst./Big Data Eng. 2016, pp. 373-377.
- [6] D. Weller-Fahy, B. Borgehtti, A. Sodemann, "A survey of distance and similarity measures used within network intrusion anomaly detection", IEEE Communication Survey & Tutorials, 2015, 17, (1), pp. 70-91.
- [7] M. Bhuyan, D. Bhattacharyya, J. Kalita, "Network anomaly detection: methods, systems and tools", IEEE Com. 2014, 16, 303-336.
- [8] M. Molina, et al, "Operational experiences with anomaly detection in backbone networks", Computer & Security, 2012, 31, (3), pp. 273-285.
- [9] F. Harrou, et al, "Amalgamation of anomaly detection indices for enhanced process monitoring", Loss Prevention in the Process Industries, 2016, 40, (3), pp. 365-377.
- [10] H. Om, T. Hazra, "Statistical techniques in anomaly intrusion detection system", Advanced in Eng. & Technology, 2012, 5, (1), pp.387-398.
- [11] M. Shyu, et al, "A novel anomaly detection scheme based on principle component classifier", Workshop, Florida, 2003, pp. 172-179.
- [12] A. Lakhina, M. Crowella, C. Diot, "Diagnosing network-wide traffic anomalies", SIGCOMM '04, 2004, pp. 219-230.
- [13] D. Brauckhoff, et al, "Applying PCA for traffic anomaly detection: problems and solutions", INFOCOM'09, pp. 2866-2870.
- [14] H. Ringberg, A. Soule, J. Rexford, C. Diot, "Sensitivity of PCA for traffic anomaly detection", Proc. ACM SIGMETRICS '07, San Diego, USA, Jun.2007, pp. 109-120.
- [15] J. Camacho, A. Perez-Villegas, P. Garcia-Teodoro, G. Macia-Fernandez, "PCA-based multivariate statistical network monitoring for anomaly detection", Computer & Security, 2016, 59 (2016), pp. 118-137.
- [16] Y. Liu, L. Zhang, Y. Guan, "Sketch-based streaming PCA algorithm for network-wide traffic anomaly detection", Proc. IEEE 30th Int. Conf. on Dist. Computing Systems, June 2010, Genoa, Italy, pp. 807-816.
- [17] C. Callegari, L. Gazzarrini, S. Giordano, M. Pagano, T. Pepe, "A novel PCA-based network anomaly detection", Prof. IEEE Int. Conf. on Communications, Kyoto, Japan, Jun. 2011, pp. 1-5.
- [18] M. Bhuyan, D. Bhattacharyya, J. Kalita, "A multi-step outlier-based anomaly detection approach to network-wide traffic", Information Sciences 2016, 348 (2016), pp. 243-271.
- [19] U. Kwitt, Hofmann, "Robust methods for unsupervised PCA-based anomaly detection", Proc. IEEE/IST Workshop, Germany, 2006, 1-3.
- [20] A. Das, S. Misra, et al, "An efficient FPGA implementation of principle component analysis based network intrusion detection system", Germany, Mar. 2008, pp. 1160-1165.
- [21] G. Zargar, T. Baghaie, "Category-based intrusion detection using PCA", Journal of Information Security, 2012, (3), pp. 259-271.
- [22] Y. Lee, Y. Yeh, Y. Wang, "Anomaly detection via online oversampling principal component analysis", Know. Data Eng. 2013, 1460-1470.
- [23] M. Elrawy, T. Abdelhamid, A. Mohamed, "IDS in telecommunication network using PCA", Comp. Nw & Com, 2013, 5, (4), pp.147-157.
- [24] D. Bayarjargal, G. Cho, "Detecting an anomalous traffic attack area based on entropy distribution and Mahalanobis distance", Int.Journal of Security and Its Applications, 2014, 8, (2), pp. 87-94.
- [25] H. Huang, H. Al-Azzawi, H. Brani, "Network Traffic Anomaly Detection", arXiv:1402.0856 preprint, 2014.
- [26] Z. Fan, et al, "Modified principal component analysis: an integration of multiple similarity subspace models", Neur.Nw 2014, pp. 1538-1552.
- [27] H.D. Nguyen, D.H. Hoang, "A model for network traffic anomaly detection", ICACT, 2015, 4, (4), pp.644-650.
- [28] E.G. Nascimento, O. Tavares, A.F. Souza, "A cluster-based algorithm for anomaly detection in time series using Mahalanobis distance", Proc. Int. Conf. Artificial Intelligence, July 2015, Nevada, USA, pp. 622-628.
- [29] S. Franklin, M. Brodeur, "A practical application of a robust multivariate outlier detection method", <http://www.amstat.org/sections/srms/>.
- [30] J.P. Geer, "Some aspects of Minkowski distance", research report, University of Leiden (1995), pp. 1-31.
- [31] http://www.takakuara.com/Kyoto_data/



Dang Hai Hoang, A/Prof. Dr. DSc., PhD (1999), DSc (2002) at TU Ilmenau, Germany. Current institution: Posts and Telecommunication Institute of Technology. Research interests: Communication network, IoT networks, information security, network security.



Ha Duong Nguyen, BSc (2001), MSc (2003) at TU Hanoi, PhD (2017) at PTIT, Vietnam. Current institution: Faculty of Information Technology. Research interests: Communication network, telecommunication networks, information security, network security.