ICIC Express Letters ICIC International ©2016 ISSN 1881-803X
Volume **10**, Number **4**, April **2016** pp. 971–975

IMPROVED KERNEL PRINCIPAL COMPONENT ANALYSIS ALGORITHM FOR NETWORK INTRUSION DETECTION

Dahui Li, Xin He and Xuefeng Dai

School of Computer and Control Engineering
Qiqihar University
No. 42, Culture Street, Qiqihar 161000, P. R. China
1125073174@qq.com

Received September 2015; accepted December 2015

Abstract. *The feature extraction is one of the key technologies of network intrusion detection. However, there is the insufficiency of extracted feature portfolio for the pattern classification in classical kernel principal component analysis (KPCA) algorithm. This paper designs an improved KPCA algorithm based on information measure. The information degree is defined with the class aggregation degree and the discrete degree, which replaces the cumulative contribution rate of KPCA. The algorithm introduces the concept of similarity to select the eigenvector combination set that is helpful to detect abnormalities. The simulation results on KDDCUP99 show that the improved algorithm has the lower dimension and realizes a clear classification effect.*

Keywords: Intrusion detection, KPCA, Feature extraction, Classification

1. **Introduction.** The feature extraction is widely used in network intrusion detection. The current approaches for feature extraction include the principal component analysis (PCA) [1-3], kernel principal component analysis (KPCA) [4-7] and nonlinear component analysis [7], etc. The basic idea of PCA method retains the largest original data covariance. The high-dimensional data will be mapped to the low dimensional space. PCA is only a linear mapping and cannot deal with the nonlinear relationship of some data. KPCA utilizes the nonlinear mapping to get the high-dimensional feature space of the raw data, and then uses PCA to analyze the principal component in the feature space [8,9]. KPCA retains the maximum information of eigenvector in the feature space and does not consider the classing efficiency. However, a shortage of the KPCA is the abnormal data classification ability [11-13]. In this paper, the information degree is defined and utilized to select the eigenvector in the improved KPCA algorithm. The dimension of eigenvector space of the sample set is reduced. The proposed approach uses multifractal to find the network anomaly, and the lack of the research is anomaly classification [14]. The simulation shows the obvious classification effect of the improved KPCA algorithm for the network.

This paper is organized as follows. Section 1 analyzes the advantage and insufficiency of KPCA. Section 2 introduces the basic principle of KPCA and the solving process of eigenvector with the whole information. Section 3 designs the process of the improved KPCA algorithm. Section 4 gives the simulation of the improved KPCA algorithm. Finally, Section 5 concludes the paper.

2. **The Basic Principle of KPCA.** Assume that $X = [x_1, \cdots, x_N] \in R^N$ is a group of training samples, where N is the numbers of the training samples. $\phi : X \to F$, where F is a Hilbert functional space. $\phi(X) = [\phi(x_1), \cdots, \phi(x_N)]$. The covariance matrix C of ϕ is calculated by

$$C = \frac{1}{N}\phi(X)\phi(X)^T \tag{1}$$

where $\sum_{i=1}^{N} \phi(x_i) = 0$, and normally, the above equation is difficult to be solved. In KPCA, the covariance matrix $K = [K_{ij}]$ is defined with the kernel function

$$K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle \tag{2}$$

where $\langle , \rangle$ is a kind of inner product, $\langle \phi(x_i), \phi(x_j) \rangle = \phi(x_i)^T \phi(x_j)$. The characteristic equation of K is defined as

$$K\alpha = \lambda\alpha \tag{3}$$

the solution of the above equation is $\Lambda = \{\lambda_i : i = 1, \cdots, N\}$ with $\lambda_1 \geq \cdots \geq \lambda_N$.

KPCA can construct the vector space with the whole features of sample information and cannot consider the sample class information meanwhile.

3. The Design of Improved KPCA Algorithm.

The processing program of improved KPCA algorithm is given as follows.

Firstly the kernel matrix K can be calculated and centralized as

$$K(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\sigma^2}\right) \tag{4}$$

$$K^C = K - IK - KI + IKI \tag{5}$$

where $K^C = [K_{ij}^C]$, $K_{ij}^C = \left\langle \phi_i^C \phi_j^C \right\rangle$, $\phi_i^C = \phi_i - \frac{1}{N} \sum_{K} \phi_K$ and $I = [1/N]_{N \times N}$.

Secondly the information degree is defined with the class aggregation degree of eigenvector inside each class and the discrete degree between the different classes.

Definition 3.1. *The class aggregation degree of each eigenvector is defined as*

$$\sigma_k^2 = \sum_{i=1}^{N} S_{ki} \tag{6}$$

where S_{ki} is the variance of a set of eigenvectors of the ith class.

Definition 3.2. *The class discrete degree is defined as*

$$D_k^2 = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \|m_{kj} - m_{ki}\|^2 \tag{7}$$

where $m_{ij} = \sum_{j=1}^{l_i} x_{k_{ij}}$, l_i is the number of samples of the ith class and $x_{k_{ij}}$ is the kth eigenvector of the ith sample.

Definition 3.3. *The information degree is defined as*

$$J_k = \frac{\sigma_k^2}{D_k^2}, \quad k = 1, \cdots, N \tag{8}$$

If J_k is smaller, then the class effect is more obvious. J_k is used to replace the cumulative contribution rate in KPCA. According to J_k, $\Lambda' = \{\lambda_i' : i = 1, \cdots, d]$ is selected with the most information and $|\lambda'| < |\lambda|$. $\lambda_k \in \Lambda$ is one of the mainly components, and $v^k = (v_1^k, \cdots, v_N^k)$ is its eigenvector. So the eigenvector space may be reconstructed with reduced dimension.

Finally, the correlation coefficient of X and v is computed. If the correlation coefficient is zero or tending to zero, then the eigenvector is removed from the original eigenvector set; otherwise, assuming y_t is the projection of X on v, it is calculated by

$$y_t = \sum_{t=1}^{N} v_t^k K(X_t, X) \tag{9}$$

4. **The Simulation of Improved KPCA Algorithm.** The paper utilizes a data subset of KDDCUP99 [15] to simulate detecting effect. This data subset is widely used in network security and retains the four network attacks such as DoS, the Probe, U2R and R2L. The simulation process is described as follows.

Firstly, the data subset is preprocessed, such as numerical value, normalization, remove the duplicate and eliminate the influence of different dimension.

Secondly, 2000 random samples are selected for feature extraction. Gaussian radial basis function (RBF) is instead of the kernel function ($\sigma^2 = 10$). The classifier is C4.5.

Finally, the classification accuracy of the whole training set is shown in Figure 1. There is not classification effect for KPCA, because KPCA can show the classification effect only on a high dimensional space.

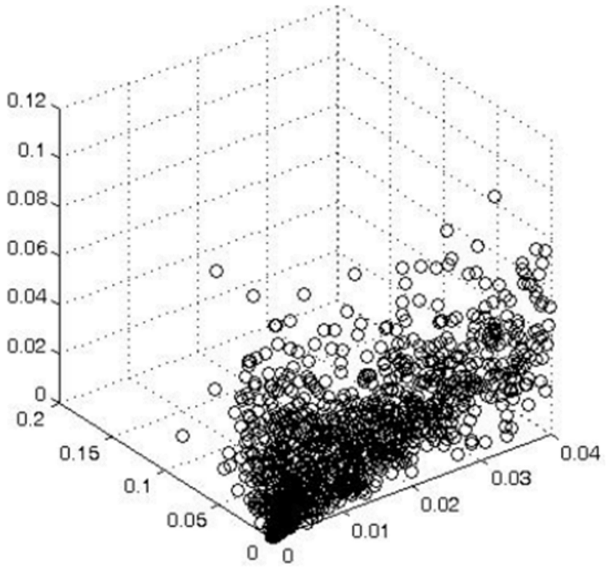

Figure 1. The class distribution of KPCA algorithm

In order to test the effect of the improved KPCA, the dimension of the random samples is reduced to three. Three eigenvectors with the most similar degree will be observed, and its effect of the classification is distinct in Figure 2. In Figure 3, the accuracy curves show that the accuracy of the improved KPCA algorithm is higher than that of KPCA algorithm on the 13th eigenvector combination. The accuracy of the improved KPCA algorithm is the highest on the 16th iteration. So compared with KPCA algorithm, the improved algorithm has obvious advantages, such as the better accuracy and the less number of feature combination.

The improved KPCA algorithm is compared with dataset without feature extraction, PCA and KPCA in Table 1. The result shows that the improved KPCA algorithm has the obvious effectiveness on accuracy rate, false-alarm rate and missing-detection rate.

Table 1. The effect contrast of feature extraction

	Accuracy rate (%)	*False-alarm rate (%)*	*Missing-detection rate (%)*
Dataset without feature extraction	*92.89*	*2.17*	*7.06*
PCA	*93.65*	*1.91*	*6.66*
KPCA	*94.82*	*1.52*	*4.59*
Improved KPCA	*96.45*	*1.41*	*3.76*

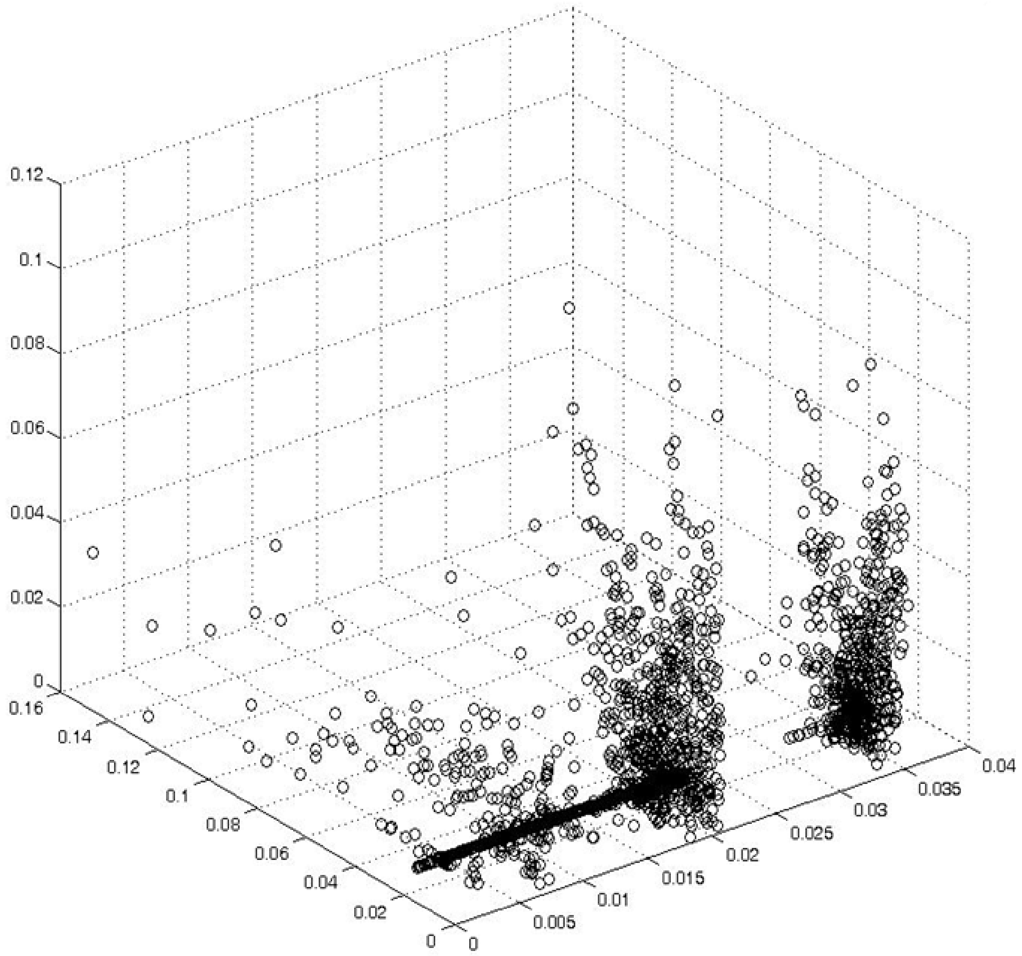

FIGURE 2. The class distribution of the improved KPCA algorithm

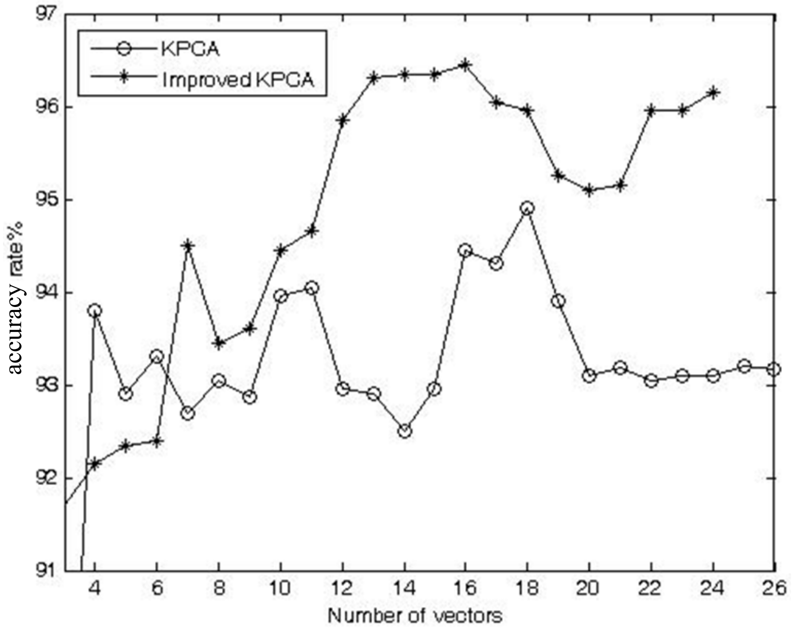

FIGURE 3. The accuracy curves of two algorithms

5. **Conclusions.** On the basis principle of the KPCA, this paper introduced the information degree and similarity to improve KPCA algorithm. The simulation results have shown that the proposed approach retains the reduced dimension ability of the KPCA and the classification ability. And the improved KPCA algorithm is more stable than KPCA algorithm. The purpose of network anomaly detection is to detect anomalies, and classify the abnormality. The future work is the proposed approach to video stream with dynamic backgrounds.

Acknowledgment. This work is supported by the Natural Science Foundation of Heilongjiang Province, China (Grant Nos. F201218 and F201331). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] X. Senkuo, *Principal Component Analysis Based Feature Extraction Methods Applied to Biomedical and Communication Network Data*, Ph.D. Thesis, University of Guelph, 2010.

[2] F. Theljani, K. Laabidi, S. Zidi and M. Ksouri, A new kernel based classification algorithm for systems monitoring: Comparison with statistical process control methods, *Arabian Journal for Science and Engineering*, vol.40, no.2, pp.645-658, 2015.

[3] L. Khaled and V. Rao, An application of principal component analysis to the detection and visualization of computer network attacks, *Annals of Telecommunications*, vol.61, no.1, pp.218-234, 2006.

[4] Y. Li, B. X. Fang, L. Guo and K. Chang, A novel data mining method for network anomaly detection based on transductive scheme, *Proc. of the 4th International Symposium on Neural Networks*, Nanjing, China, pp.1286-1292, 2007.

[5] M. Li, *Sparse Kernel Principal Component Analysis*, MIT Press, Cambridge, 2001.

[6] Y. Xu, D. Zhang, F. X. Song, J. Yang, Z. Jing and M. Li, A method for speeding up feature extraction based on KPCA, *Neurocomputing*, vol.70, no.4-6, pp.1056-1061, 2007.

[7] Y. W. Zhang, Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM, *Chemical Engineering Science*, vol.64, no.4-6, pp.801-811, 2009.

[8] J. Li, X. L. Li and D. C. Tao, KPCA for semantic object extraction in images, *Pattern Recognition*, vol.41, no.5, pp.3244-3250, 2008.

[9] W. D. Chang, W. F. Liu and X. A. Yan, The application of two kinds of feature extraction technology in intrusion detection, *Journal of Changchun University of Technology (Natural Science Edition)*, vol.11, no.3, pp.42-47, 2007.

[10] Z. Z. Wei, Feature extraction based on kernel principal component analysis, *Journal of Guangxi University of Technology*, vol.17, no.4, pp.27-31, 2006.

[11] W. D. Chang, Intelligent feature selection and classification techniques for intrusion detection in networks: A survey, *EURASIP Journal on Wireless Communications and Networking*, vol.271, no.1, pp.271-287, 2013.

[12] H. H. Gao, H. H. Yang and X. Y. Wang, PCA/KPCA feature extraction approach to SVM for anomaly detection, *Journal of East China University of Science and Technology (Natural Science Edition)*, vol.32, no.3, pp.321-325, 2006.

[13] J. Xu and X. M. Tao, One class intrusion detection system based on KPCA space similarity, *Journal of Computer Applications*, vol.29, no.9, pp.2459-2463, 2009.

[14] D. H. Li, *Study on Multi-fractal Modeling and Multistep Predicting of Network Video Traffic*, Ph.D. Thesis, Harbin Engineering University, 2011.

[15] *http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html*, 2006.