Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2015, Article ID 491910, 11 pages
http://dx.doi.org/10.1155/2015/491910

Research Article

LKM: A LDA-Based *K*-Means Clustering Algorithm for Data Analysis of Intrusion Detection in Mobile Sensor Networks

Yuhua Zhang,[1] Kun Wang,[1,2] Min Gao,[2] Zhiyou Ouyang,[1] and Siguang Chen[1]

[1] *Key Lab of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Ministry of Education, Nanjing 210003, China*
[2] *Electrical Engineering Department, UCLA, Los Angeles, CA 90095, USA*

Correspondence should be addressed to Kun Wang; kwang@njupt.edu.cn

Received 8 July 2014; Accepted 9 September 2014

Academic Editor: Guangjie Han

Mobile sensor networks (MSNs), consisting of mobile nodes, are sensitive to network attacks. Intrusion detection system (IDS) is a kind of active network security technology to protect network from attacks. In the data gathering phase of IDS, due to the high-dimension data collected in multidimension space, great pressure has been put on the subsequent data analysis and response phase. Therefore, traditional methods for intrusion detection can no longer be applicable in MSNs. To improve the performance of data analysis, we apply *K*-means algorithm to high-dimension data clustering analysis. Thus, an improved *K*-means clustering algorithm based on linear discriminant analysis (LDA) is proposed, called LKM algorithm. In this algorithm, we firstly apply the dimension reduction of LDA to divide the high-dimension data set into 2-dimension data set; then we use *K*-means algorithm for clustering analysis of the dimension-reduced data. Simulation results show that LKM algorithm shortens the sample feature extraction time and improves the accuracy of *K*-means clustering algorithm, both of which prove that LKM algorithm enhances the performance of high-dimension data analysis and the abnormal detection rate of IDS in MSNs.

1. Introduction

FOR the special network application, mobile sensor networks (MSNs) are proposed as a new type of wireless sensor network with low node density and sparse network environment. Nodes in MSNs are no longer distributed statically to sample data but wore on the mobile carriers and move with them, which conduces to highly dynamic network topology, poor network connectivity, and intermittent communication between nodes [1]. Furthermore, nodal mobility will easily lead to invalid data transmission, retransmission storm, and easiness of being injected by virus, risking the security of network. Keung et al. studies the intrusion detection problem in a mobile sensor network, focusing on providing barrier coverage against moving intruders [2]. It becomes particularly challenging when the movement route of sensors and intruders needs to be captured. Abduvaliyev et al. propose a comprehensive classification of various IDS approaches according to their employed detection techniques such as anomaly detection, misuse detection, and specification-based detection protocols [3]. Because of wireless broadcast communication, wireless sensor networks (WSNs) are vulnerable to denial-of-service (DoS) attacks. Han et al. identifie malicious nodes through energy consumption of sensor nodes in WSNs, distinguishing the ongoing DoS attack species effectively. It is difficult to set a suitable threshold for evaluating the abnormal energy consumption [4]. To detect the sinkhole nodes in WSNs, Han et al. also make full use of neighbor information stored in the nodes, avoiding the occurrence of abnormal energy holes and severe malicious attacks, for example, the selective forwarding attack [5]. However, there is little related work about the security of MSNs. In the previous studies, IDS is the key technology to ensure network security and also a proactive technology to protect network from attacks. IDS can detect system or network resources in real-time to find out network intruders or prevent legitimate users from misusing resources.

IDS is divided into three parts, for example, data gathering, data analysis, and response. In order to find traces of network intrusion, IDS collects data from multiple points in the network system. Thus, data gathering covers multidimension space including system logs, network packets, important documents, and status or behavior of user activity. In traditional data analysis phase, the collected high-dimension data will be handled with pattern matching, anomaly detection, or integrity testing. Once intrusion behaviors are found, IDS will immediately enter the response phase, including logging, alarming, and security controlling. In the data gathering phase of IDS, therefore, the high-dimension data obtained from multidimension space bring great challenges to the subsequent analysis and response stages. This paper proposes the LKM algorithm that firstly applies linear discriminant analysis (LDA) to accomplish dimensionality reduction of original high-dimension data and then uses the K-means cluster analysis for the 2-dimension data. In this way, LKM solves the high-dimension data analysis problems in IDS.

Lee et al. are the first one to apply data mining techniques to intrusion detection, such as association rules algorithm, classification algorithm, and sequence mining algorithm [6]. Recently, it has been a new trend to use the data mining algorithm in IDS. Related algorithm research will help to improve the security of network. Clustering analysis in intrusion detection is the key to realizing intelligent IDS. Clustering analysis is an important method for data partitioning or packet processing of huge data. Clustering algorithm can be roughly classified into division-based method, hierarchy-based method, density-based method, grid-based method, model-based method, and the fuzzy clustering. K-means algorithm is a typical clustering algorithm based on distance partition. The distance is used as the similarity evaluation index. In the case of two-dimensional or three-dimensional data, K-means algorithm can ensure the quality of the clustering. However, *curse of dimensionality* is common in K-means clustering algorithm when dealing with n-dimensional ($n > 3$) data set. In this circumstance, the processing time of K-means algorithm will be too long and the efficiency will be rather low.

In this paper, an improved K-means clustering algorithm (namely, LKM algorithms) is proposed. Firstly, LDA is introduced to reduce the dimensionality of the original high-dimensional data, and then K-means clustering algorithm is adopted to make clustering analysis on the dimension reduced data. The so-called dimension reduction refers to the process in which samples from the high-dimensional space are mapped to the low-dimensional space for a meaningful representation of the high-dimensional data by linear or nonlinear method. Through data dimension reduction, the *curse of dimensionality* can be alleviated and other irrelevant properties in the high-dimensional space are eliminated. Therefore, we offset the defects and improve the performance of the K-means clustering algorithm when dealing with high-dimensional data set.

The rest of the paper is organized as follows. In Section 2, related works of the K-means algorithm are introduced, mainly about existing problems and corresponding improvements. In Section 3, LDA and K-means clustering algorithm are described respectively, and then LKM algorithm is described in detail. Simulation experiments on LKM algorithm are implemented, and experimental comparison results of existing PCA-Km algorithm and LKM algorithm are discussed in Sections 4 and 5. Section 6 concludes the paper.

2. Related Work

Lee et al. introduced an intrusion detection method based on data mining. The basic idea is to use the audit program to extract a large number of network connections and the host session features and apply data mining technology to export the rules that correctly distinguish between normal and intrusion behavior [6]. Elbasiony et al. proposed a hybrid detection framework which is based on data mining classification and clustering techniques. In the framework, K-means clustering algorithm is used to detect novel intrusions by clustering the network connections' data to collect most of the intrusions together in one or more clusters [7]. Muniyandi et al. proposed an anomaly detection method using K-means combined with C4.5, a method to cascade K-means clustering, and the C4.5 decision tree methods for classifying anomalous and normal activities in a computer network [8]. However, there are still some shortcomings of K-means algorithm. The main problems are in the following areas. (i) The result of the clustering mostly depends on the selection of the initial centers. Yedla et al. put forward searching for better initial centers and providing an effective method to allocate a suitable cluster of data points [9]; Neha and Kirti selected the data located at the center as the initial point [10]. (ii) The number of clusters k needs to be given in advance. Li suggested the optimal value of k and the conditions of its upper bound [11], which confirms the rationality of the thumb k-max theoretically. (iii) The clustering results are vulnerable to the noise data point. Momin and Yelmar utilized the possible members to reduce the noise points [12]; Wang and Su reduced the impact of noise points via preprocessing [13]. (iv) The algorithm is not applicable for a large amount of data clustering problems. Kanungo et al. used a simple data structure to store the information for each iteration [14]; Li et al. speeded up the computation rate by reducing the grid data [15]. (v) The algorithm cannot deal with the high-dimensional data effectively. Literatures [16, 17] failed to solve the fusion problem of K-means and dimension reduction; Napoleon and Pavalakodi proposed PCA-Km algorithm [18], which applied PCA on original data set and obtained a reduced data set containing possibly uncorrelated variables; Ding and He proposed a coherent framework to adaptively select the most discriminative subspace [19].

Different from other K-means, LKM achieves linear dimension reduction for the original high-dimension data and then generates 2D data for K-means clustering analysis. In this way, LDA and K-means are combined with each other. LKM improves the performance of K-means dealing with high-dimension data than other modified K-means solutions. Since the works of Napoleon et al. and Ding et al. are much similar to ours, the simulation analysis and the experimental comparison of LKM algorithm and PCA-Km algorithm are discussed in Section 4.

3. LKM Algorithm

In LKM algorithm, the linear dimension reduction method, LDA, is firstly adopted to reduce the dimension of the original n-dimension data set A. After the dimension reduction, the l-dimension data set Y can be obtained. Then K-means clustering algorithm is applied to clustering analysis, and the final result is output.

3.1. Algorithm Definition. LDA is the method that minimizes within-class distance while it maximizes the interclass distance as much as possible [20]. So we can obtain the optimal projection direction to obtain the best classification. That is, we choose characteristic description of samples that maximize the ratio of within-class dispersion and interclass dispersion. For a given matrix $A \in R^{d\times n}$ ($R^{d\times n}$ represents the n-dimension real linear space constituted by all the $d \times n$ real matrices), LDA is used to generate a transformation matrix $G \in R^{d\times l}$ ($R^{d\times l}$ represents the l-dimension real linear space constituted by all $d \times l$ real matrices). Each column vector a_i of the matrix A in the n-dimensional space is mapped to a column vector y_i in the l-dimensional space as follows:

$$y_i = G^T a_i \in R^l \quad (l < d), \ 1 \le i \le n. \tag{1}$$

The matrix A is divided into k classes; that is

$$A = [A_1, A_2, \ldots, A_k], \quad A_i \in R^{d\times n_i}, \qquad n = \sum_{i=1}^{k} n_i, \tag{2}$$

where n_i represents the number of data in A_i and R^l indicates the l-dimensional linear space.

The concepts of within-class scattering matrix, interclass scattering matrix, and the total scattering matrix of LDA are defined as follows.

Definition 1. Within-class scattering matrix S_w is as follows:

$$S_w = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in A_i}\left(x - c^{(i)}\right)\left(x - c^{(i)}\right)^T. \tag{3}$$

The within-class scattering matrix S_w reflects the mean square distance between various kinds of samples and the centers of various classes. S_w indicates the dispersion degree of samples in the same class.

Definition 2. Interclass scattering matrix S_b is as follows:

$$S_b = \frac{1}{n}\sum_{i=1}^{k} n_i\left(c^{(i)} - c\right)\left(c^{(i)} - c\right)^T. \tag{4}$$

The interclass scattering matrix S_b reflects the mean square distance between centers of various classes and the overall center. S_b indicates the dispersion degree of centers in different classes.

Definition 3. In the total scattering matrix S_t, obviously, S_t is equal to the sum of S_w and S_b; namely,

$$S_t = S_w + S_b. \tag{5}$$

According to formula (5), total scattering matrix can be deduced as follows:

$$S_t = \frac{1}{n}\sum_{j=1}^{n}\left(a_j - c\right)\left(a_j - c\right)^T. \tag{6}$$

The total scattering matrix S_t reflects overall dispersion degree of entire sample. Here, $c^{(i)}$ is the initial centroid of A_i. Calculating mean value for all the data objects in A_i, $c^{(i)}$ can be expressed as

$$c^{(i)} = \frac{1}{n_i}A_i e^{(i)}, \tag{7}$$

where $e^{(i)}$ represents n order column matrix and the values of all matrix elements are one; namely,

$$e^{(i)} = (1, 1, \ldots, 1)^T \in R^n. \tag{8}$$

On the basis of formulas (5) and (7), the expression of the overall centroid can be deduced as follows:

$$c = \frac{1}{n}Ae, \tag{9}$$

where $e = (1, 1, \ldots, 1)^T \in R^n$, $n = \sum_{i=1}^{k} n_i$.

In the low-dimensional space reduced by the linear transformation matrix G, S_w is turned into $G^T S_w G$. S_b is changed to $G^T S_b G$, and S_t becomes $G^T S_t G$. However, in practical applications of LDA, when sample dimension is greater than or close to the number of samples, the within-class scattering matrix is not reversible. Meanwhile, it is difficult to calculate the matrix directly, which is the so-called problem of *smallsamplesize*(*SSS*) [21].

Therefore, we take advantage of the best transformation matrix G^* to overcome the *SSS* problem, which is defined as follows.

Definition 4. Calculate optimal transformation matrix G^* by solving the optimization problems:

$$G^* = \arg\max_G \left\{\operatorname{trace}\left(\left(G^T S_w G\right)^{-1} G^T S_b G\right)\right\}. \tag{10}$$

Considering formulas (3), (4), and (5), we can get the equivalent form of (10):

$$G^* = \arg\max_G \left\{\operatorname{trace}\left(\left(G^T S_t G\right)^{-1} G^T S_b G\right)\right\}. \tag{11}$$

In this way, we are able to obtain the optimum conversion matrix G^*. The above optimization problem is equivalent to the equation when $\lambda \neq 0$:

$$S_b x = \lambda S_t x. \tag{12}$$

The value of x is solved out which satisfies the above conditions. When matrix S_t is nonsingular, by conducting singular value decomposition (SVD) on matrix $S_t^{-1}S_b$, we can obtain x that meets these conditions. Finally, each column vector a_i of matrix A in $R^{d\times n}$ is a one-to-one mapping to column vector y_i in l-dimension space $R^{d\times l}$; namely,

$$y_i = (G^*)^T a_i, \quad 1 \le i \le n. \tag{13}$$

In addition, in order to measure dissimilarity of data object, we use Euclidean distance ranging method in K-means algorithm.

Definition 5. In the high-dimension space, dimensional mapping process of data objects makes Euclidean distance in two-dimension space close to the shortest path from the high-dimensional space between objects. In two-dimension space, Euclidean distance between two points can be calculated based on the coordinates or vectors of them; namely,

$$D\left(y_i, Z_j(I)\right) = \sqrt{\left(y_i - Z_j(I)\right)^2}. \tag{14}$$

Definition 6. In order to obtain the best clustering results, error square and the guideline functions are adopted to obtain the optimal value of J_c, which is defined as

$$J_c(I) = \sum_{j=1}^{k}\sum_{k=1}^{n_j}\left\| y_k^{(j)} - Z_j(I)\right\|^2. \tag{15}$$

J_c represents sum of squared error of all data samples and their centers, when a data set containing n data objects is divided into k classes. The value of J_c is related to the size of cluster center. Apparently, the error of data objects within their class center enlarges as J_c increases. Thus, the degree of differences between various types of data within the object will be greater, and quality of clustering will be worse than before.

Definition 7. The iteration is implemented repeatedly to find k clustering centers; n sample points are assigned to the nearest cluster center, so that the clustering error sum of squares reaches a minimum. Clustering center Z_j is calculated as follows:

$$Z_j(I) = \frac{1}{n}\sum_{i=1}^{n_j} x_i^{(j)}, \quad j = 1, 2, 3, \dots, k. \tag{16}$$

3.2. Algorithm Steps. LKM algorithm flow chart is shown in Figure 1.

The specific process of LKM algorithm is illustrated as follows:

input: n-dimension data set of matrix $A = (a_1, a_2, a_3, \dots, a_n)$, the number of clusters k;

output: k clusters C_m that meets conditions of l-dimension data.

Step 1. Apply LDA to generate a transformation matrix G.

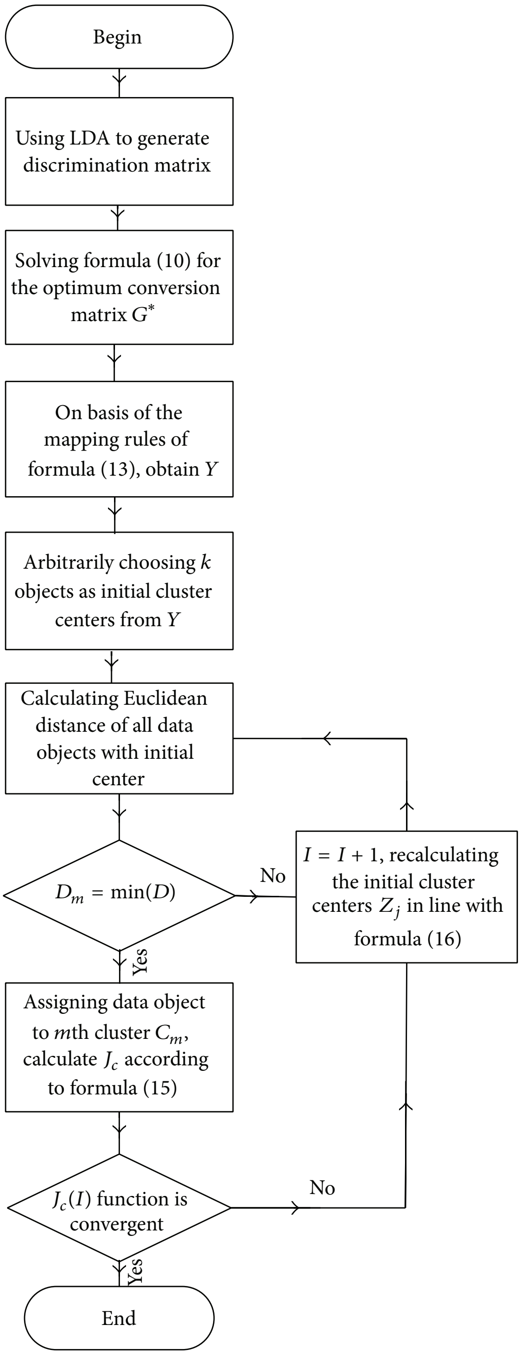

Begin
Using LDA to generate discrimination matrix
Solving formula (10) for the optimum conversion matrix G^*
On basis of the mapping rules of formula (13), obtain Y
Arbitrarily choosing k objects as initial cluster centers from Y
Calculating Euclidean distance of all data objects with initial center
$D_m = \min(D)$
No
$I = I + 1$, recalculating the initial cluster centers Z_j in line with formula (16)
Yes
Assigning data object to mth cluster C_m, calculate J_c according to formula (15)
$J_c(I)$ function is convergent
No
Yes
End

Figure 1: LKM algorithm flow chart.

Step 2. Generate within-class scattering matrix S_w, interclass scattering matrix S_b, and total scattering matrix S_t.

Step 3. Solve the optimization problems and then the best transformation matrix G^* is obtained.

Step 4. Using optimal transformation matrix G^*, each column vector a_i of matrix A in n-dimension space $R^{d\times n}$ is one-to-one mapped to column vector y_i in l-dimension space $R^{d\times l}$. The data set Y is obtained.

```
Input: $n$-dimensional dataset of matrix $A = (a_1, a_2, a_3, \dots, a_n)$
Output: $l$-dimensional dataset $Y = (y_1, y_2, y_3, \dots, y_n)$ after dimension reduction
Begin
    initial $n$ prototype $a_i$, $i \in [1, n]$
    repeat
    compute $S_w$, $S_b$, $S_t$, $G$, $G^*$
    if $G^* = \arg\max_G\{\mathrm{trace}((G^T S_w G)^{-1} G^T S_b G)\}$
        for $I = 1$ to $n$ do
            $y_i = (G^*)^T a_i$
        end for
End
```

Algorithm 1: Pseudocode of dimensionality reduced data set Y through LDA.

The pseudocode of the dimensionality reduced data set Y by LDA is shown in Algorithm 1.

Step 5. Randomly select k objects from Y as the initial cluster centers $Z_j(I)$, $j = 1, 2, 3, \dots, k$, $I = 1$.

Step 6. Calculate the Euclidean distanced $D(y_i, Z_j(I))$ of all the data objects and k initial center, $i = 1, 2, 3, \dots, n$, $j = 1, 2, 3, \dots, k$. If the condition that $D(y_i, Z_m(I)) = \min D(y_i, Z_j(I))$ is satisfied, $m \in 1, 2, 3, \dots, k$, y_i will be assigned to the mth cluster C_m.

Step 7. Calculate sum of squared error criterion function J_c.

Step 8. Judge: if function J_c is convergent, or $|J_c(I) - J_c(I-1)| < \varepsilon$, the algorithm is finished and outputs the results; otherwise, $I = I + 1$, recalculate new cluster center $Z_j(I)$, and return to Step 6 to recalculate distance.

The pseudocode of K-means clustering analysis of l-dimensional data sets Y is shown in Algorithm 2.

4. Simulation Analysis

We employ matlab7.0 programming for simulations. The PCA-Km algorithm [18] is similar to LKM algorithm described in this paper. To compare performance of these two cluster analysis algorithms for high-dimension data processing, we, respectively, utilize the LKM algorithm and PCA-Km algorithm to experiment with 40-dimension data set and 70-dimension data set. The results are shown from Figure 2 to Figure 9. The dimension of the initial data set is changed successively; the LKM algorithm and PCA-Km algorithm are applied to make clustering analysis for 2-dimension, 3-dimension, 4-dimension,. . ., 70-dimension initial data sets. Therefore, the feature extraction time of LDA and PCA is shown in Figure 10. The changes of the three algorithms (PCA-km, LKM, and K-means algorithm) are shown in Figure 11.

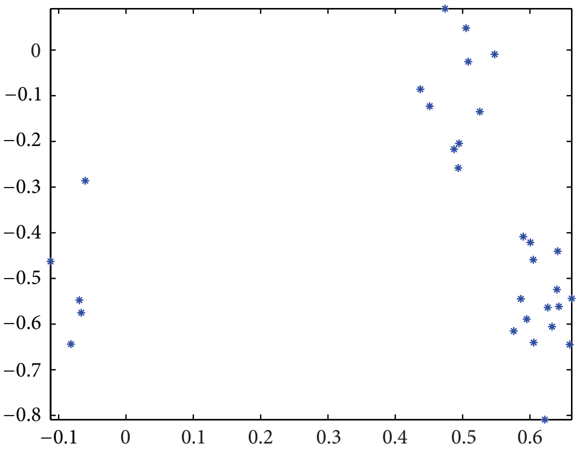


Figure 2: 30 × 2-dimension data set Y after dimension reduction of LKM.

(1) Use the rand() function to randomly generate a 40-dimension data set A with 30 rows and 40 columns. LKM algorithm is used for LDA linear dimension reduction to generate a 2-dimension data set Y with 30 rows and 2 columns. The result is shown in Figure 2. PCA-Km algorithm is used for the PCA linear dimension reduction to obtain a 2-dimension data set Y' with 30 rows and 2 columns. The result is shown in Figure 3.

Comparing Figures 2 and 3, although LDA and PCA both complete the process of the data dimension reduction, LDA is superior to PCA in classification.

As we keep running LKM algorithm to perform clustering analysis of 2-dimension data set Y, the final outputs are two clusters. The clustering result of LKM algorithm for the 40-dimension data is shown in Figure 4; similarly, we proceed to implement the PCA-Km algorithms cluster analysis of the data set Y'; two cluster classes finally are output. The output of PCA-Km algorithm is shown in Figure 5.

By comparison of Figures 4 and 5, the distribution of data objects remains relatively dispersed after PCA-Km clustering. On the contrary, after the clustering process of LKM algorithm, the data objects are divided into two obviously different clusters.

(2) Similarly, simulation is implemented with a 70-dimension data set A with 50 rows and 70 columns randomly generated by rand() function to verify the performance of experiment (1). The LKM algorithm is executed to make LDA

```
Input: $l$-dimensional data sets $Y = (y_1, y_2, y_3, \ldots, y_n)$, clustering number $k$
Output: $k$ of clusters $C_m$ consisting of $l$-dimensional data
Begin
  $I = 1$
  initial $k$ prototype $Z_j$, $j \in [1, k]$
  repeat
    for $I = 1$ to $n$ do
      compute $D(y_i, Z_j(I))$
      if $D(y_i, Z_m(I)) = \min D(y_i, Z_j(I))$ then $y_i \in C_m$
    end for
  if $I = 1$, then
      compute $J_c(I)$
      $I = I + 1$
    for $j = 1$ to $k$ do
      compute $Z_j(I), J_c(I)$
    Until $|J_c(I) - J_c(I-1)| < \varepsilon$
End
```

Algorithm 2: Pseudocode of K-means clustering analysis of l-dimensional data sets Y.

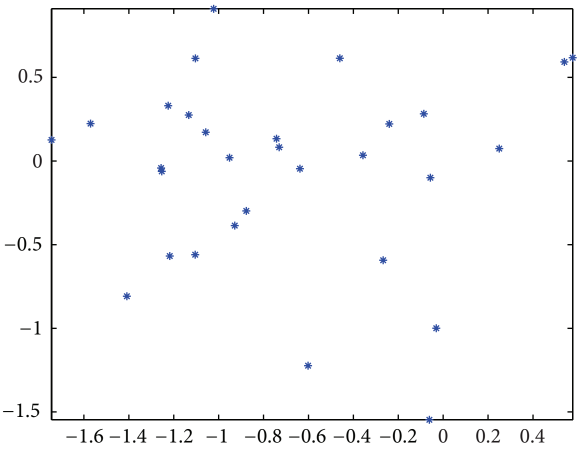


Figure 3: 30×2-dimension data set Y' after dimension reduction of PCA-Km.

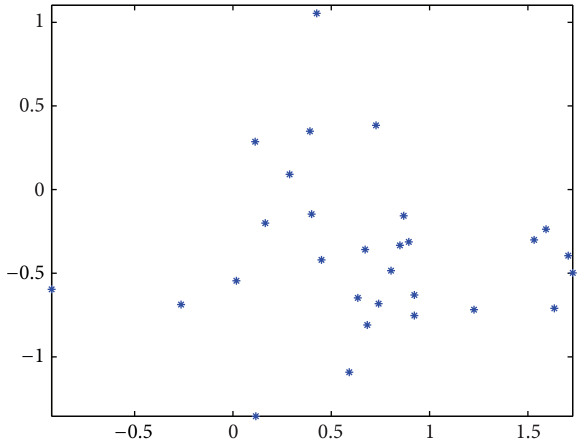


Figure 5: PCA-Km algorithm clustering output.

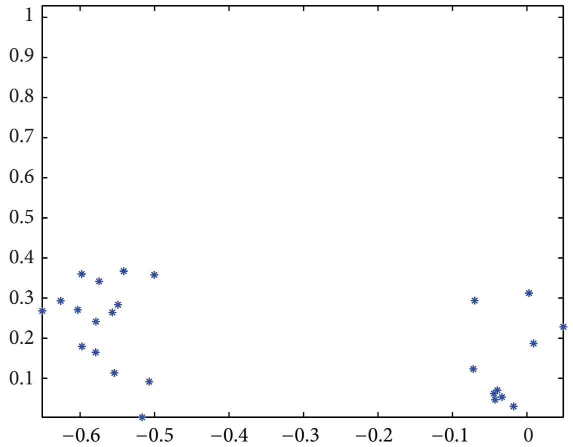


Figure 4: LKM algorithm clustering output.

linear dimension reduction on A. The result of 2-dimension data set Y in 50 rows and 2 columns is shown in Figure 6. After the PCA linear dimension reduction of PCA-Km, the result of the two-dimension data set Y' in 50 rows and 2 columns is shown in Figure 7.

Comparing the results of dimension reduction from Figures 6 and 7, LDA not only maintains the best projection identify information of the original data, but also improves the classification performance. Figures 6 and 7 further verified that the classification performance of LDA is superior to that of PCA.

If we go on with LKM algorithm, the final output will be two clusters. The results of LKM clustering algorithm for 70-dimensional data analysis are shown in Figure 8. And the results of PCA-Km algorithms are shown in Figure 9.

In Figures 8 and 9, the data is divided into two distinct clusters classes after the LKM algorithm clustering analysis, and the clustering effect is very ideal. However, the PCA-Km algorithm clustering analysis still fails to achieve the desired

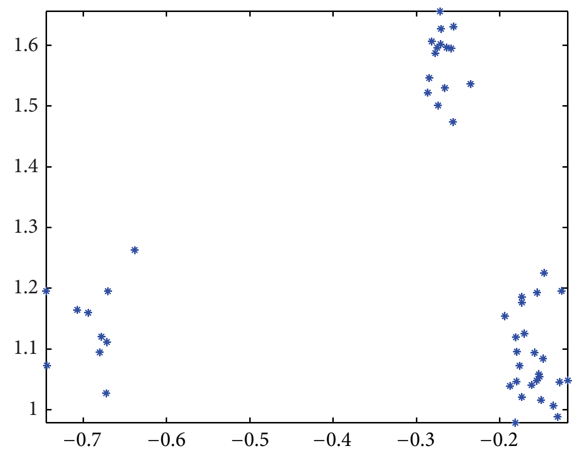


Figure 6: 50 × 2-dimension data set Y after dimension reduction of LKM.

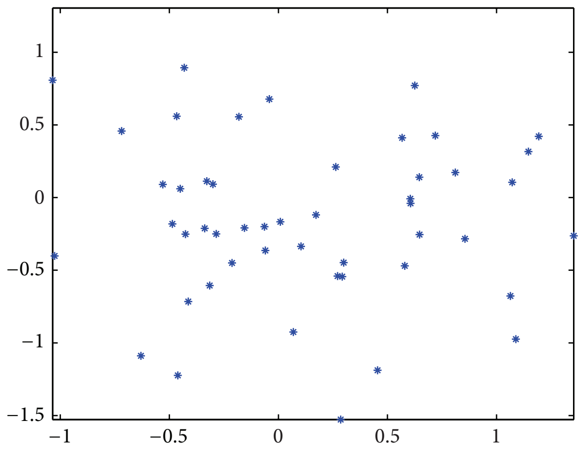


Figure 7: 50 × 2-dimension data set Y' after dimension reduction of PCA-Km.

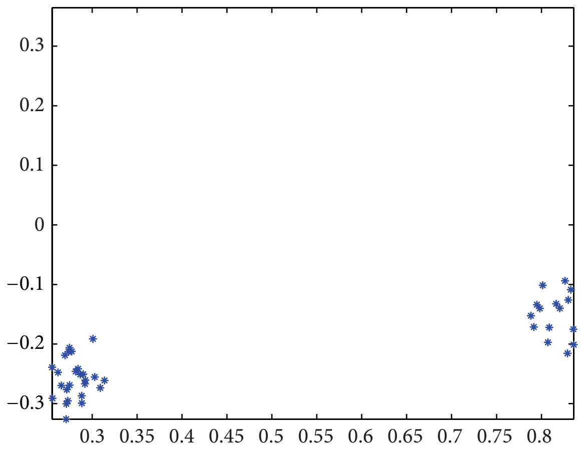


Figure 8: LKM algorithm clustering output.

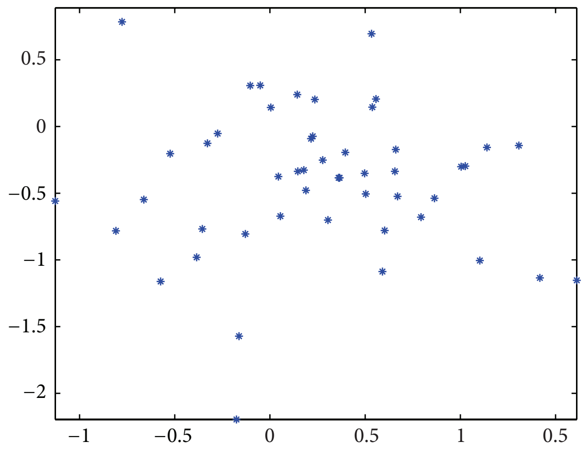


Figure 9: PCA-Km algorithm clustering output.

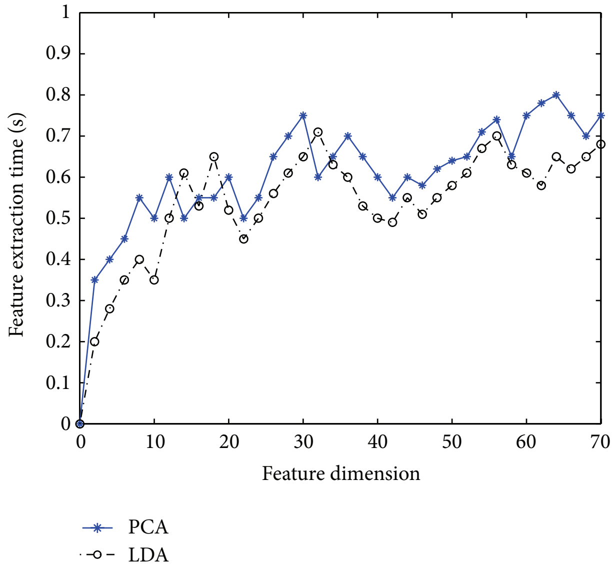


Figure 10: Feature extraction time of LDA and PCA.

clustering effect. We can conclude that LKM algorithm clustering outperforms PCA-Km algorithm.

(3) The rand() function is run to randomly generate 2-dimension, 3-dimension, 4-dimension,. . ., 70-dimension initial data sets A. Implement the above experiments, respectively, to make linear dimension reduction for different initial data sets. In order to effectively implement the classification, feature extraction is needed, which means that when all kinds of information contained in high-dimension space are analyzed and processed, the unique attributes are screened without the interference of external factors. And feature extraction time is the average time to complete the feature extraction process. Changes of feature extraction time for LDA and PCA linear dimension reduction technique are shown in Figure 10.

Experiments show that when dealing with the data sets of the same dimension, feature extraction time of LDA linear dimension reduction technique is shorter than that of PCA

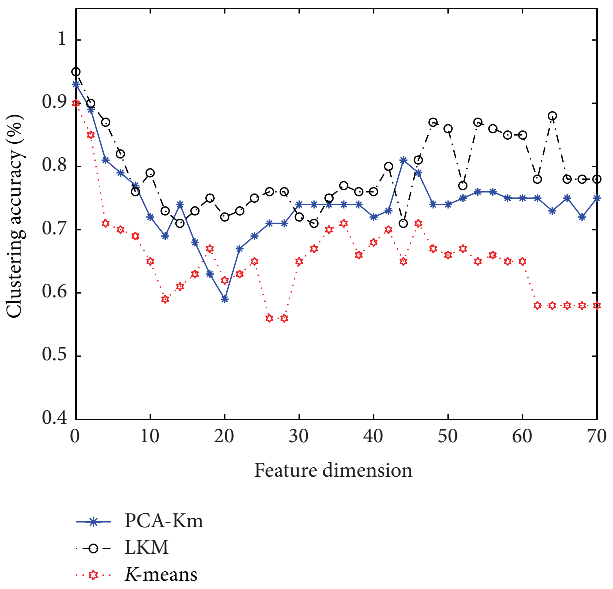


FIGURE 11: Clustering precision changes of the PCA-Km, LKM, and K-means.

linear dimension reduction techniques. Unlike PCA, LDA is a supervised feature extraction method, which can not only maintain the best projection identification information of the original data, but also improve the classification performance and the efficiency.

With the increasing dimension of initial data sets, the changes of the clustering precision for these three algorithms, PCA-Km, LKM, and K-means, are shown in Figure 11.

Figure 11 illustrates that when processing 1-dimension, 2-dimension, or 3-dimension data, K-means algorithm is still able to guarantee the quality of clustering. However, when processing the n-dimension ($n > 3$) data objects, the performance of K-means clustering algorithm is poor. Instead, the accuracy of improved K-means algorithm based on PCA or LDA (namely, PCA-Km or LKM algorithm) is significantly higher than K-means clustering algorithm. When characteristic dimension of initial data set is the same, it is shown in Figure 11 that the performance of LKM algorithm clustering is better than that of PCA-Km.

(4) To further validate the proposed algorithm's effectiveness in IDS, this paper applies the well-known KDD Cup 1999 (KDD'99) data sets to perform the following experiments. The attacks in KDD'99 data sets are categorized into five types: DoS (denial of rervice), Probe, U2R (user to root), date compromise (data), and R2L (remote to local) [22]. For testing, 2500 testing instances are randomly selected from the KDD'99 data sets. The experimental results are evaluated with two indicators that are DR (detection rate) and FPR (false positive rate):

(1) DR (detection rate): DR = the number of detected attacks/the total number of attacks,

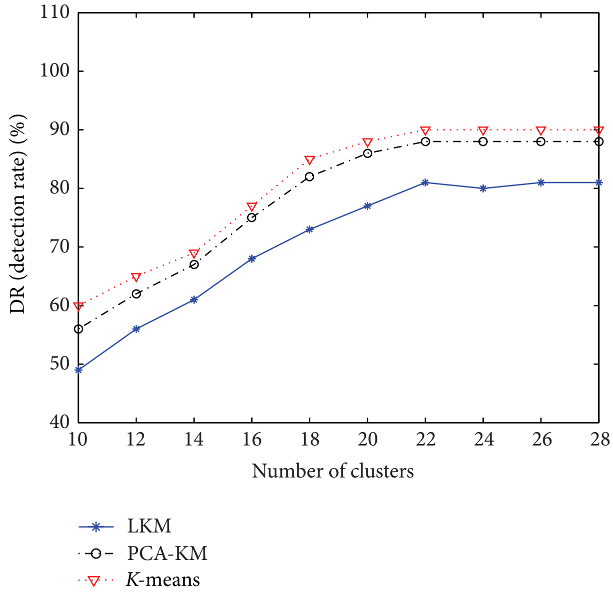


FIGURE 12: DR versus the number of clusters.

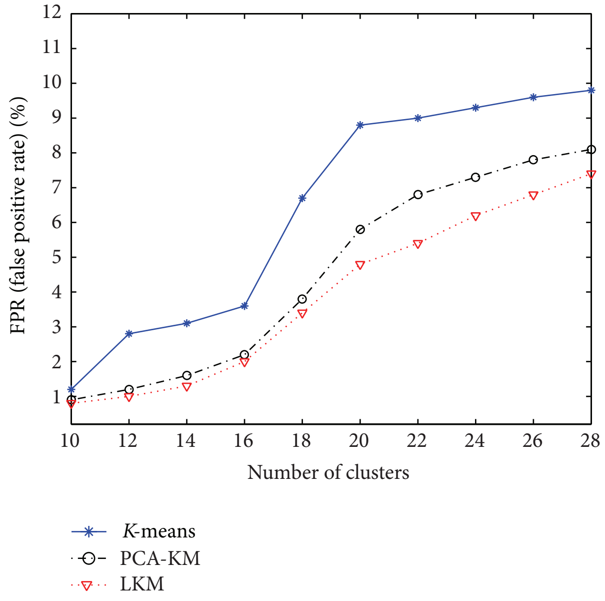


FIGURE 13: FPR versus the number of clusters.

(2) FPR (false positive rate): FPR = the number of normal connections that are misclassified as attacks/the total number of normal connections.

Since the K-means clustering algorithm requires to input the number of clusters in advance, different values of clusters have great influence on clustering outputs. Both LKM and PCA-KM incorporate the K-means method. Therefore, Figures 12 and 13 illustrate the performance of K-means, LKM, and PCA-KM when the number of clusters equals 10, 12, 14, 16, 18, 20, 22, 24, 26, and 28.

With the increment of the number of clusters, DR and FPR also increase. Figures 12 and 13 show that when the

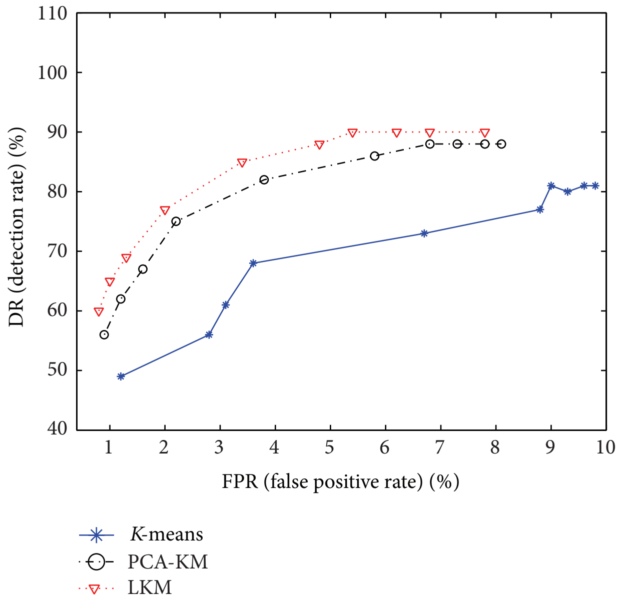


Figure 14: ROC curve of DR and FPR.

number of clusters is rare, abnormal data after training are rare. Many unusual attacks are recorded as normal and assigned to normal class while normal ones will not be assigned to abnormal class. But when the number of clusters increases, the abnormal class system after training also increases. Because of the supervised feature extraction (LDA), LKM detects more attacks and makes fewer mistakes than PCA-KM and K-means.

When Figures 12 and 13 are mapped into receiver operating characteristic (ROC) curve, relationship between DR and FPR can be reflected directly as Figure 14.

Seen from the ROC curve, FPR increases gradually with DR. At initial stage, the curve of LKM is relatively steep, indicating that DR can have greatly improvement at the expense of FPR. Then as FPR increases, DR begins to rise slowly. When the number of clusters is about 22, curve of LKM is relatively flat even FPR has increased significantly and DR increases significantly small. In general, curve inflection point is considered the best point of the system performance. Therefore, optimal number of LKM clusters is 22 or so.

5. Discussions

Since K-means is a classical algorithm in data mining, many literatures have done a lot of researches to improve the performance of K-means algorithm. Here we will make comparisons between some typical algorithms and our work in detail. Napoleon and Pavalakodi proposed PCA-Km algorithm [18], which incorporated principal component analysis (PCA) and K-means for original high-dimension data set analysis.

The main difference between LKM and PCA-KM is described as follows. LDA, the key technology used in LKM, is a supervised feature extraction method to screen low-dimension features with the strongest discriminating power from the high-dimension space. LDA not only maintains optimal projector identification information of original data, but also improves the classification performance and efficiency. PCA, the key technology used in PCA-KM, is an unsupervised feature extraction method that selects the original data projection with the maximum feature covariance. PCA and linear transformation are used for dimension reduction; then dimension-reduced data set is clustered by K-means clustering algorithm; Ding and He proved that continuous solutions of discrete K-means clustering membership indicators are data projections on principal directions (principal eigenvectors of the covariance matrix) [19]. New lower bounds for K-means objective function are derived, and it is directly related to eigenvalues of covariance matrix. Ding and Li combined linear discriminant analysis (LDA) and K-means clustering into a coherent framework. The most discriminative subspace is selected adaptively [23]. Relations among PCA, LDA, and K-means were clarified. However, further experiments and comparison are not given in detail. It is shown in Section 4 that characteristics extraction time of PCA is higher than LDA. Unlike PCA, LDA is a supervised feature extraction method. LDA maintains the best projection identification information of original data as well as classification performance.

6. Conclusions

To achieve IDS in MSNs, data gathered by mobile nodes covers multidimension space, including system logs, network packets, important documents, and status or behavior of user activity. Applying clustering analysis in intrusion detection is the key to realizing intelligent IDS. By K-means clustering analysis for high-dimension data, the performance of data analysis in intrusion detection can be enhanced. Through linear dimension reduction method of LDA, the LKM algorithm proposed in this paper reduces the curse of dimensionality. Other irrelevant attributes in high-dimension space are eliminated, and characteristics of sample extraction time have been shortened. The integration of K-means clustering into dimension-reduced data set improves clustering accuracy, which is rather meaningful in enhancing data analysis of high-dimension data set. As a result, the anomaly detection rate increases, which improves the performance of IDS. However, the initial centers of K-means algorithm used in the LKM algorithm are chosen arbitrarily, which have a great impact on results of clustering analysis. Optimization of initial centers of the LKM algorithm will be discussed in our future work.

Glossary

k-max:	the upper bound of k value optimization problem
A:	original high-dimension data sets before dimension reduction
Y:	low-dimension data sets after linear dimension reduction
$R^{d\times n}$:	n-dimension real linear space constituted by the $d \times n$ real matrix

$R^{d\times l}$:	l-dimension real linear space constituted by the $d \times l$ real matrix
a_i, a_j:	ith, jth column vector of A
y_i:	ith column vector of Y
G:	transformation matrix in LDA
G^*:	optimization transformation matrix in LDA
n_i:	the number of data in A_i
A_i:	matrix A is divided into ith class
S_w:	within-class scattering matrix
S_b:	between-classes scattering matrix
S_t:	total scattering matrix
$c^{(i)}, c$:	the initial center of mass in A_i and the overall center of mass in A
$e^{(i)}$:	all matrix elements are one; namely, $e^{(i)} = (1, 1, 1)^T$
Z_j:	cluster centers in K-means clustering algorithm
J_c:	data sample of its class where the center of the squared error criterion function
$D(y_i, Z_j)$:	the distance between y_i and Z_j
I:	the number of iterations counter
rand():	randomly generatee the initial function of n-dimension data set A
DR:	detection rate
FPR:	false positive rate
ROC:	receiver operating characteristic (ROC) curve.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by NSFC (61100199, 61100213, 61170276, 61201160, 61201222); Specialized Fund for the Doctoral Program of Higher Education (20113223120007); Natural Science Foundation of Jiangsu Province (BK20141427); China Scholarship Council Project (201208320212); Program of Natural Science for Universities of Jiangsu Province (12KJD510007); Open Lab Fund of NUPT (2014XSG03); Scientific Innovation Project for Postgraduates of Universities in Jiangsu Province (CXLX13_463).

References

[1] B. Liu, O. Dousse, P. Nain, and D. Towsley, "Dynamic coverage of mobile sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 2, pp. 301–311, 2013.

[2] G. Y. Keung, B. Li, and Q. Zhang, "The intrusion detection in mobile sensor network," *IEEE/ACM Transactions on Networking*, vol. 20, no. 4, pp. 1152–1161, 2012.

[3] A. Abduvaliyev, A.-S. K. Pathan, J. Zhou, R. Roman, and W.-C. Wong, "On the vital areas of intrusion detection systems in wireless sensor networks," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1223–1237, 2013.

[4] G. Han, J. Jiang, W. Shen, L. Shu, and J. Rodrigues, "IDSEP: a novel intrusion detection scheme based on energy prediction in cluster-based wireless sensor networks," *IET Information Security*, vol. 7, no. 2, pp. 97–105, 2013.

[5] G. J. Han, X. Li, J. F. Jiang, L. Shu, and J. Lloret, "Intrusion detection algorithm based on neighbor information against sinkhole attacks in wireless sensor networks," *The Computer Journal*, 2014.

[6] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in *Proceedings of IEEE Symposium on Security and Privacy*, pp. 120–132, Oakland, Calif, USA, 2009.

[7] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted K-means," *Ain Shams Engineering Journal*, vol. 4, no. 4, pp. 753–762, 2013.

[8] A. P. Muniyandi, R. Rajeswari, and R. Rajaram, "Network anomaly detection by cascading K-Means clustering and C4.5 decision tree algorithm," in *Proceedings of the International Conference on Communication Technology and System Design (ICCTSD '11)*, pp. 174–182, Coimbatore, India, December 2011.

[9] M. Yedla, S. R. Pathakota, and T. M. Srinivasa, "Enhancing K-means clustering algorithm with improved initial center," *International Journal of Computer Science and Information Technologies*, vol. 1, no. 2, pp. 121–125, 2010.

[10] A. Neha and A. Kirti, "A mid-point based k-mean clustering algorithm for data mining," *Computer Science and Engineering*, vol. 4, no. 6, pp. 1174–1180, 2012.

[11] X. W. Li, "Research on text clustering algorithm based on improved K-means," in *Proceedings of the International Conference on Computer Design and Appliations*, vol. 4, pp. 573–576, Qinhuangdao, China, June 2010.

[12] B. F. Momin and P. M. Yelmar, "Modifications in K-means clustering algorithm," *Soft Computing and Engineering*, vol. 2, no. 3, pp. 349–354, 2012.

[13] J. T. Wang and X. L. Su, "An improved K-Means clustering algorithm," in *Proceedings of the IEEE 3rd International Conference on Communication Software and Networks (ICCSN '11)*, pp. 44–46, Xi'an, China, May 2011.

[14] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithms: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[15] D. Li, J. Shen, and H. Chen, "A fast K-means clustering algorithm based on Grid data reduction," in *Proceedings of the IEEE Aerospace Conference (AC '08)*, pp. 1–6, Big Sky, Mich, USA, March 2008.

[16] B. B. Baridam, "More work on K-means clustering algorithm: the Dimen-sionality problem," *Computer Applications*, vol. 44, no. 2, pp. 23–30, 2012.

[17] J. M. Li and Y. T. Qian, "Dimention reduction of hyperrspectral images with sparse near discriminant analysis," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pp. 2927–2930, July 2011.

[18] D. Napoleon and S. Pavalakodi, "A new method of dimensionally reduction using K-means clustering algorithm for high dimensional data set," *International Journal of Computer Applications*, vol. 13, no. 7, pp. 41–46, 2011.

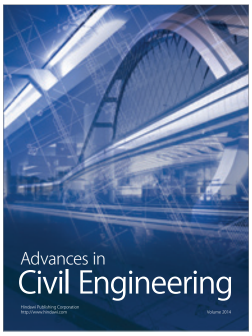
[19] C. Ding and X. He, “Principal component analysis and effective K-means clustering,” in *Proceedings of the International Conference on Data Mining (SIAM ’04)*, pp. 497–501, Orlando, Fla, USA.

[20] W.-K. Ching, D. Chu, L.-Z. Liao, and X. Wang, “Regularized orthogonal linear discriminant analysis,” *Pattern Recognition*, vol. 45, no. 7, pp. 2719–2732, 2012.

[21] M. I. Razzak, M. K. Khan, and K. Alghathbar, “Face recognition using layered linear discriminant analysis and small subspace,” in *Proceedings of IEEE 10th International Conference on Computer and Information Technology (CIT ’10)*, pp. 1407–1412, Bradford, UK, July 2010.

[22] Z. Y. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, “A system for denial-of-service attack detection based on multivariate correlation analysis,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 2, pp. 447–456, 2014.

[23] C. Ding and T. Li, “Adaptive dimension reduction using discriminant analysis and K-means clustering,” in *Proceedings of the 24th International Conference on Machine Learning (ICML ’07)*, pp. 521–528, Corvallis, Ore, USA, June 2007.



Hindawi

Submit your manuscripts at
http://www.hindawi.com