Journal of the Korean Society of Marine Engineering, Vol. 40, No.8 pp. 726~732, 2016
J. Korean Soc. of Marine Engineering (JKOSME)
http://dx.doi.org/10.5916/jkosme.2016.40.8.726

ISSN 2234-7925 (Print)
ISSN 2234-8352 (Online)
Original Paper

Performance evaluation of principal component analysis for clustering problems

Jae-Hwan Kim† · Tae-Min Yang[1] · Jung-Tae Kim[2]

(Received September 26, 2016 ; Revised October 10, 2016 ; Accepted October 17, 2016)

Abstract: Clustering analysis is widely used in data mining to classify data into categories on the basis of their similarity. Through the decades, many clustering techniques have been developed, including hierarchical and non-hierarchical algorithms. In gene profiling problems, because of the large number of genes and the complexity of biological networks, dimensionality reduction techniques are critical exploratory tools for clustering analysis of gene expression data. Recently, clustering analysis of applying dimensionality reduction techniques was also proposed. PCA (principal component analysis) is a popular methd of dimensionality reduction techniques for clustering problems. However, previous studies analyzed the performance of PCA for only full data sets. In this paper, to specifically and robustly evaluate the performance of PCA for clustering analysis, we exploit an improved FCBF (fast correlation-based filter) of feature selection methods for supervised clustering data sets, and employ two well-known clustering algorithms: *k*-means and *k*-medoids. Computational results from supervised data sets show that the performance of PCA is very poor for large-scale features.

Keywords: Clustering algorithm, Dimensionality reduction, PCA, Feature selection

1. Introduction

Clustering analysis is widely used in data mining to classify data into categories on the basis of their similarity. It is the formal study of methods and algorithms for the natural grouping, or clustering, of objects according to measured or perceived intrinsic characteristics or similarities. Clustering methods are especially useful for exploring interrelationships among neighbors. Applications range broadly from pattern recognition to microarrays, multimedia, bibliometrics, bioinfomatics, and gene profiling.

Clustering algorithms can be classified into two general categories: hierarchical and non-hierarchical methods. Hierarchical methods classify the original data into similar categories without predetermining the number of clusters. Several hierarchical methods have been developed : linkage methods **[1]**, Ward's method **[2]**, DIANA algorithm **[3]**, and DPC (density peaks clustering) **[4]** algorithm, etc. Linkage methods include single, complete, and average or centroid linkage. For example, the single linkage method is to connect the nearest neighbor by Euclidean distance. In contrast to the single linkage, complete linkage method is to cluster the farthest neighbor. Rodriguez *et al.* **[4]** proposed an efficient DPC algorithm that cluster centers are characterized by a higher density than their neighbors.

Non-hierarchical methods are clustering algorithms in which the number of clusters is predetermined. These include *k*-means **[5]**, *k*-medoids (PAM(patitioning around medoids) **[3]**), CLARA (clustering large applications) **[3]**, and so on.

In gene profiling, however, because of the large number of genes and the complexity of biological networks, dimensionality reduction techniques are critical exploratory tools for clustering analyses of gene expression data. Recently, clustering methods themselves have been proposed as ways of reducing dimension. Dimensionality reduction techniques are based on feature extraction methods: SVD (singular value decopositon) **[6]**, PCA **[7]**, LSA (latent semantic analysis) **[8]**, RP (random projection) **[9]**. Feature extraction is generally a procedure of unsupervised learning to apply the clustering analysis. In particular, PCA is a useful method of feature extraction for clustering problems. Yeung *et al.* **[10]** showed that PCA was efficient to apply clustering analysis in gene profiling problems. Song *et al.* **[11]** performed a comparative study of dimensionality reduction techniques to enhance trace clustering performances, and showed that

† Corresponding Author (ORCID: http://orcid.org/0000-0002-8248-6325): Department of Data Information, Korea Maritime and Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: jhkim@kmou.ac.kr, Tel: 051-410-4374
1 Department of Data Information, Korea Maritime and Ocean University, E-mail: canada0704@naver.com, Tel: 051-410-4775
2 Department of Data Information, Korea Maritime and Ocean University, E-mail: jt0998@gmail.com, Tel: 051-410-4884

dimensionality reduction could improve trace clustering performance in relation to the computation time and average fitness of the mined local-process models.

However, Yeung *et al.* **[10]** and Song *et al.* **[11]** compared the classification accuracies for the full data and the reduced data by PCA. In this paper, to specifically and robustly evaluate the performance of PCA for clustering analysis, we exploit an improved FCBF (fast correlation-based filter) of feature selection methods for supervised clustering data sets. We employ two well-known clustering algorithms: *k*-means and *k*-medoids. Computational results from supervised learnig show that the performance of PCA is very poor for large-scale features.

The remainder of this paper is organized as follows. Diensionality reduction techniques and clustering methods are introduced in Sections 2 and 3, respectively. In Section 2, we focus on PCA of feature extraction methods, and briefly mention the feature selection of dimensionality reduction techniques. In Section 3, we deal with two well-known clustering algorithms: *k*-means and *k*-medoids clustering algorithm. In Section 4, the computational results for performance are presented. Finally conclusions are drawn in Section 5.

2. Dimensionality Reduction Techniques

The dimensionality of data refers to the number of features or attributes that describe each record in real-world data. Most data mining applications are composed of high-dimensional data, in which not all of the features are relevant. High-dmensional data can contain many irrelevant, redundant or noisy features that may greatly decrease the performance of data mining process. In addition, many algorithms become computationally intractable because of the so-called "curse of dimensionality".

In data mining field, reduction-dimensionality is an important step in the preprocessing data. Dimensionality reduction techniques not only profilerate the data mining algorithms faster, but also provide higher accuracy of a data mining process so that the model can be represented better from the data.

Dimensionality reduction techniques usually involve two steps: feature extraction and feature selection. Feature extraction is generally an unsupervised learning procedure for applying the clustering analysis. Feature extraction methods include SVD, PCA, LSA, and RP. Typically, PCA is a popular method of feature extraction for clustering problems. Yeung *et al.* **[10]** showed the efficiency of PCA to apply clustering analysis in gene profiling problems. Song *et al.* **[11]** performed a comparative study of dimensionality reduction techniques to enhance trace clustering performances.

However, Yeung *et al.* **[10]** and Song *et al.* **[11]** analyzed the performance of PCA for only full data sets. To evaluate PCA performance for clustering problems, we exploited feature selection for the supervised clustering cases. We selected an I-FCBF as the feature selection method, known as the best **[12]**.

2.1 PCA

The most popular feature extraction method for reducing the dimensionality of a large data set is a form of PCA known as the Karhunen-Loeve method. PCA is to reduce the dimensionality of data by transforming the full original attribute space into a smaller space by using an important property of eigenvalue decomposition. The basic idea of PCA is to derive new variables that are combinations of the original ones, and that are uncorrelated and arranged in order of decreasing variance.

The full procedure of this algorithm is given in **[13]**.

(i) Normalization: make each attributes have the same mean (zero) and variance.
(ii) Calculate the covariance matrix Σ.
(iii) Calculate the eigenvectors u_i and eigenvalues λ_i of Σ.
(iv) Sort these eigenvalues in decreasing order and stack the eigenvectors λ_i corresponding to the eigenvalues λ_i in columns to form the matrix $\boldsymbol{U}$.

2.2 I-FCBF (Improved FCBF)

Feature selection is usually a supervised dimensionality reduction procedure that involves removing irrelevant and redundant features of high-dimensional data. The filter methods of feature selection can be broadly divided into two classes: univariate and multivariate approaches. Univaiate filter methods are computationally very efficient due to the ignorance of the dependency between features. Thus, with univariate approaches, computing time is extremely fast, but they produce less accurate solutions. To overcome this flaw of the univariate filter, in which the dependency between features is ignored, multivariate approaches have been proposed in the literature. The FCBF **[14]**, I-FCBF **[12]**, and mRMR (minimum

Redundancy Maximum Relevance) **[15]** are well-known efficient multivariate approaches. Among them, we exploite the I-FCBF to evaluate performance of PCA.

The pseudocode of I-FCBF is as follows.

Algorithm. I- FCBF
Input: $X(x_1, x_2, \dots, x_n)$, Y // a training data set
δ // a predefined threshold
Output: S // the selected I-FCBF set

for i **in** 1: n **do**
  if SU(x_i, Y) < δ **then**
    remove(X, x_i)
  end if
end for
$x_p \leftarrow x_i$ with the largest SU(x_i, Y) in X
append(S, x_p)
remove(X, x_p)
while $x_p \neq$ null **do**
  for x_q **in** X **do**
    if SU(x_p, x_q) ≥ SU(x_q, Y) **then**
      remove(X, x)
    end if
  end for
  $x_p \leftarrow \max_{x_j \in X-S}$ [SU(x_j, c) - $\frac{1}{|S|}\sum_{x_i \in S} SU(x_i, x_j)$]
end while

3. Clustering Methods

Many clustering methods including hierarchical and non-hierarchical algorithms have been developed. Non-hierarchical methods are clustering algorithms in which the number of clusters is predetermined. These include *k*-means, *k*-medoids (PAM), CLARA, and so on. Among them, we employed two well-known methods for clustering algorithms (*k*-means and *k*-medoids) in order to compare the performance of PCA. In this section, we briefly mention the two methods.

3.1 *k*-means method

The *k*-means algorithm is one of the simplest unsupervised learning algorithms for cluster analysis. The algorithm aims to partition observations into *k* clusters that each observation belongs to the cluster of minimizing the following objective function,

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \quad (1)$$

where $\left\|x_i^{(j)} - c_j\right\|$is the Euclidean distance between observation $\left\|x_i^{(j)} - c_j\right\|$ and cluster center c_j.

The algorithm is composed of the following steps:

(i) Select k observations randomly. Set these observations to be the initial centroids.
(ii) Assign each observation to the cluster that has the nearest centroid.
(iii) Once all observations have been assigned, recalculate the positions of the *k* centroids.
(iv) Repeat steps (ii) and (iii) until the centroids no longer move.

3.2 *k*-medoids method

The *k*-medoids algorithm is similar to the *k*-means clustering algorithm. However, *k*-menas algorithm is sensitive to outliers. For this reason, the *k*-mediods algorithm considers representative observations called medoids instead of centroids. Because it chooses the most centrally located observation in a cluster, it is less sensitive to outliers than the *k*-maens algorithm. The most common realization of K-medoids algorithm, PAM, is selected in this study. The PAM procedure is as follows:

(i) Select k observations randomly. These observations are the initial medoids.
(ii) Assign each observation to the cluster that has the nearest medoid.
(iii) Select a non-medoid observation randomly.
(iv) Compute the total cost of swapping the old medoid with the currently selected non-medoid observation.
(v) If the total cost of swapping is less than zero, then perform the swap operation to generate a new set of *k* medoids.
(vi) Repeat steps (ii), (iii), (iv) and (v) until the medoid locations stabilize.

4. Computational Results

To evaluate the performance of PCA, five data sets were selected from the literatures. These data sets are shown in **Table 1**. For example, the ORL dataset contained 1024 features, 400 samples, and 40 classes. The class sizes of ISOLET, COIL, and ORL are 26, 20, and 40, respectively. Their sizes are larger than the data set of Lung and Carcinom. Features of Lung and Carcinom have a large number of 3312 and 9182, respectively.

Table 1: Five data sets

Data set	Features	Samples	Classes
ORL	1024	400	40
COIL	1024	1440	20
ISOLET	617	1560	26
Lung	3312	203	5
Carcinom	9182	174	11

To evaluate the performances of clustering methods, the following measure of popular Rand index **[16]** is usually used,

$$RI = (a + b) / (a + b + c + d) \quad (2)$$

a: the number of pairs in the original data set S that are in the cluster X and Y

b: the number of pairs in S that are not in the cluster X and not in the cluster Y

c: the number of pairs in S that are in cluster X and not in the cluster Y

d: the number of pairs in S that are not in cluster X and in the cluster Y.

In our experiments, we employed the following measure of adjusted Rand index **[17]** for more accurate analysis.

$$ARI = 2(ad - bc) / ((a + b)(b + d) + (a + c)(c + d)) \quad (3)$$

In general, the higher the value of the adjusted Rand index, the better the performance of the clustering algorithms.

In this study, computational experiments were conducted on an Intel i7 PC with 3.4 GHz CPU and 8 GB RAM. All source codes were implemented with the R language.

PCA is an unsupervised procedure of dimensionality reduction techniques. To robustly evaluate the performance of PCA, we employed five supervised data sets and I-FCBF of supervised feature selection for reduction-dimensionality.

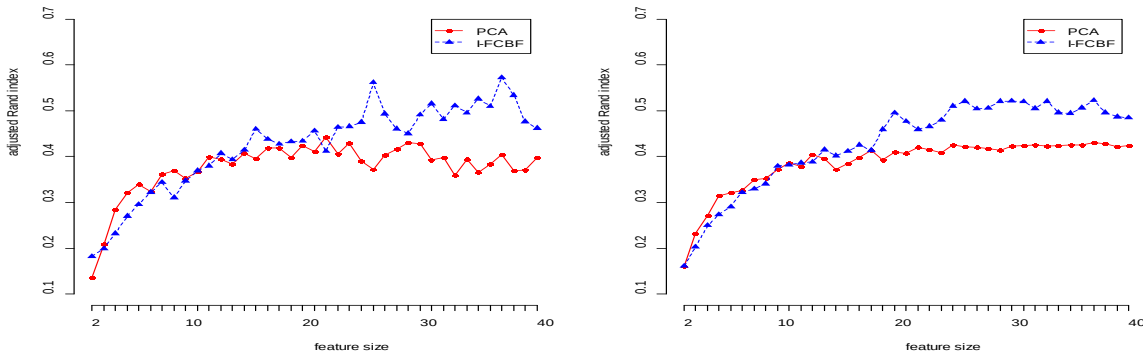
The results for adjusted Rand index are shown in **Tables 2 - 6**. For the ORL data set (**Table 2**), we note that the performance of PCA is very close to that of I-FCBF. However, as the reduction dimension grows, the performance of PCA was more or less worse than that of I-FCBF for *k*-means and *k*-medoids.

The results for the COIL data set are given in **Table 3** and **Figure 2**. In this case, the performance of PCA is very good for most cases. However, as the reduction dimension grows, the performance of PCA becomes worse than that of I-FCBF for *k*-means method. For the ISOLET data set, we note that the performance of PCA is very close to that of I-FCBF of supervised feature selection. That is, this means that PCA obtains good results with ISOLET which has relatively few features (617).

We also tested data sets with large-scale features. From **Table 1**, the Lung and Carcinom data sets have a large number of 3312 and 9182, respectively. For these data sets, however, we observed that the PCA performance was very poor for *k*-medoids. Specifically, from **Table 5**, the value of the adjusted Rand index of PCA is 0.4248 for *k*-medoids when the number of dimensions is 40, whereas I-FCBF obtained the value of 0.7647. This is an appreciably gap (>0.3) in adjusted Rand index. We also obtained very similar results for the Carcinom data set. From **Table 6**, as the number of dimensions grows, we note that the PCA performance is very poor for both methods.

Table 2: Adjusted Rand index of ORL

Clustering	Dimension	2	10	20	30	40
***k*-means**	PCA	0.1355	0.3520	0.4237	0.4276	0.3983
	I-FCBF	0.1820	0.3470	0.4339	0.4917	0.4619
k-medoids (PAM)	PCA	0.1608	0.3713	0.4097	0.4231	0.4241
	I-FCBF	0.1612	0.3793	0.4956	0.5212	0.4845

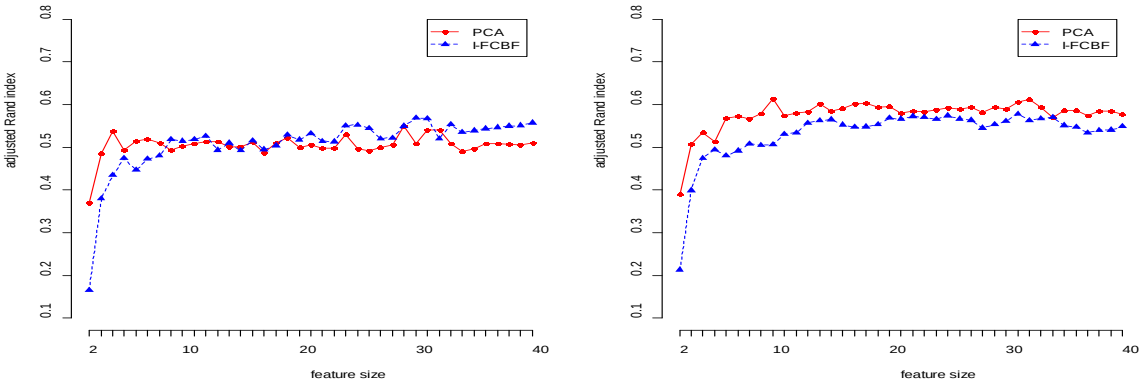


(a) *k*-means (b) *k*-medoids (PAM)

Figure1: Adjusted Rand index of ORL

Table 3: Adjusted Rand index of COIL

Clustering	Dimension	2	10	20	30	40
k-means	PCA	0.3702	0.5017	0.4997	0.5086	0.5101
	I-FCBF	0.1652	0.5147	0.5178	0.5686	0.5577
k-medoids (PAM)	PCA	0.3895	0.6135	0.5949	0.5895	0.5775
	I-FCBF	0.2133	0.5063	0.5684	0.5610	0.5488

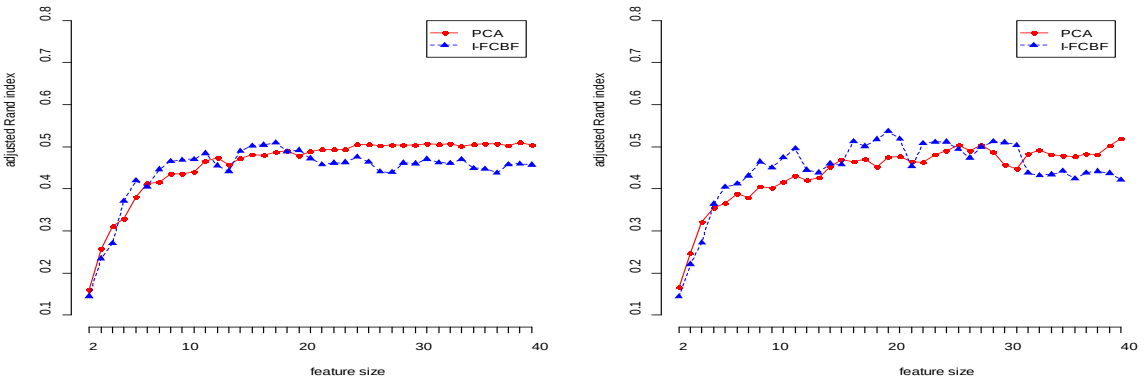


(a) *k*-means (b) *k*-medoids (PAM)

Figure 2: Adjusted Rand index of COIL

Table 4: Adjusted Rand index of ISOLET

Clustering	Dimension	2	10	20	30	40
k-means	PCA	0.1600	0.4359	0.4784	0.5033	0.504
	I-FCBF	0.1447	0.4683	0.4916	0.4600	0.4571
k-medoids (PAM)	PCA	0.1651	0.4022	0.4755	0.4564	0.5190
	I-FCBF	0.1439	0.4504	0.5365	0.5099	0.4214

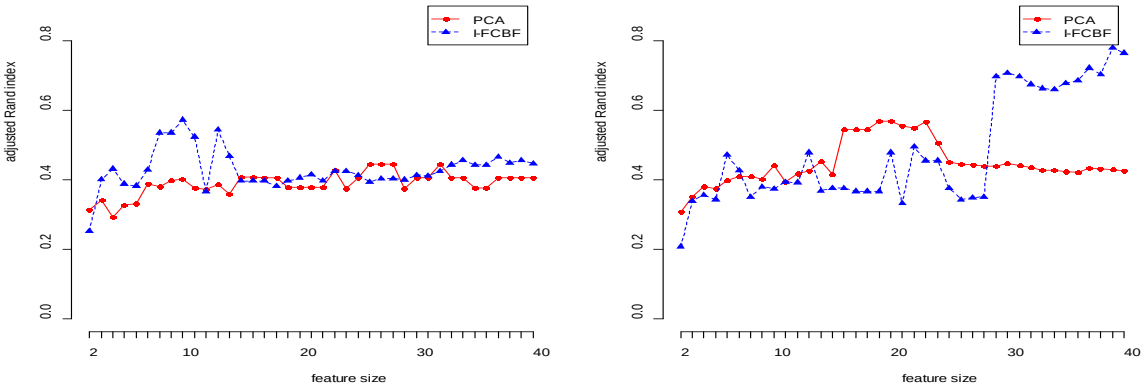


(a) *k*-means (b) *k*-medoids (PAM)

Figure 3: Adjusted Rand index of ISOLET

Table 5: Adjusted Rand index of Lung

Clustering	Dimension	2	10	20	30	40
***k*-means**	PCA	0.3138	0.4017	0.3784	0.4052	0.4065
	I-FCBF	0.2523	0.5727	0.4057	0.4128	0.4461
k-medoids (PAM)	PCA	0.3076	0.4404	0.5684	0.4468	0.4248
	I-FCBF	0.2080	0.3737	0.4786	0.7068	0.7647

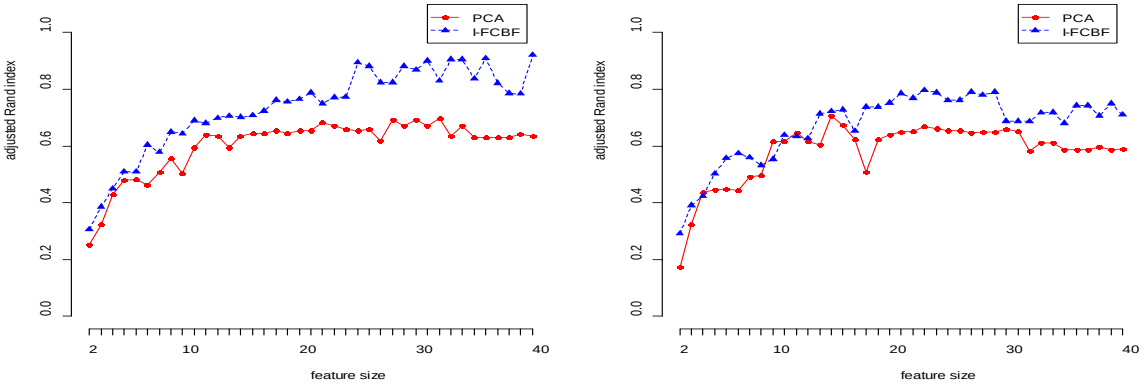


(a) *k*-means (b) *k*-medoids (PAM)

Figure 4: Adjusted Rand index of Lung

Table 6: Adjusted Rand index of Carcinom

Clustering	Dimension	2	10	20	30	40
k-means	PCA	0.2715	0.5038	0.6549	0.6934	0.6337
	I-FCBF	0.3058	0.6433	0.7656	0.8684	0.9204
k-medoids (PAM)	PCA	0.1736	0.6149	0.6399	0.6589	0.5901
	I-FCBF	0.2920	0.5539	0.7519	0.6875	0.7099



(a) *k*-means (b) *k*-medoids (PAM)

Figure 5: Adjusted Rand index of Carcinom

5. Conclusions

Through the decades, many clustering techniques have been developed, including hierarchical and non-hierarchical algorithms. Recently, dimensionality reduction using cluster analysis has also been proposed. A popular method of feature extraction in clustering problems is PCA. Yeung *et al.* **[10]** showed the efficiency of PCA for clustering analysis in gene profiling problems. Song *et al.* **[11]** performed a comparative study of dimensionality reduction techniques for enhancing trace clustering performances.

To evaluate the performance of PCA robustly, we employed five supervised data sets and I-FCBF of supervised feature selection for reduction-dimensionality, using two well-known clustering algorithms: *k*-means and k-medoids.

From our computational results, we observed that PCA obtained good results for the relatively few features of ORL, COIL, and ISOLET. For large-scale features, however, we noted that the performance of PCA was very poor for both methods as the number of dimensions grew.

In future work, we intend to develop a compensated and efficient PCA for large-scale features.

References

[1] T. Sorensen, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons,” Biologiske Skrifter, vol. 5, pp. 1-34, 1948.

[2] J. R. Ward and H. Joe, “Hierarchical grouping to optimize an objective function,” Journal of the American statistical association, vol. 58, no. 301, pp. 236-244, 1963.

[3] L. Kaufman, and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, John Wiley & Sons, 2009.

[4] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” Science, vol. 344, no. 6191, pp. 1492-1496, 2014.

[5] J. B. MacQueen, “Some Methods for classification and Analysis of Multivariate Observations,” Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, no. 14, pp. 281-297, 1967.

[6] K. Barker, “Singular value decomposition tutorial,” The Ohio State University, vol. 24, 2005.

[7] I. Jolliffe, Principal Component Analysis, John Wiley & Sons, 2002.

[8] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” Discourse processes, vol. 25 no. 2-3, pp. 259-284, 1998.

[9] E. Bingham and H. Mannila, “Random projection in dimensionality reduction: applications to image and text data,” Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, pp. 245-250, 2001.

[10] K. Y. Yeung and W. L. Ruzzo, “Principal component analysis for clustering gene expression data,” Bioinformatics, vol. 17, no. 9, pp. 763-774, 2001.

[11] M. Song, H. Yang, S. H. Siadat, and M. Pechenizkiy, “A comparative study of dimensionality reduction techniques to enhance trace clustering performances,” Expert Systems with Applications, vol. 40, no. 9, pp. 3722-3737, 2013.

[12] J. T. Kim, H. Y. Kum, and J. H. Kim, “A comparative study of filter methods based on information entropy,” Journal of the Korean Society of Marine Engineering, vol. 40, no. 5 pp. 437-446, 2016.

[13] M. Du, S. Ding and H. Jia, “Study on density peaks clustering based on k-nearest neighbors and principal component analysis”, Knowledge-Based Systems, vol. 99, pp. 135-145, 2016.

[14] L. Yu and H. Liu, “Feature selection for high- dimensional data: A fast correlation-based filter solution,” International Conference Machine Learning, vol. 3, pp. 856-863, 2003.

[15] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency max-relevance and min-redundancy,” IEEE Transactions on pattern analysis and machine intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.

[16] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” Journal of the American Statistical Association, vol. 66, no. 336, pp. 846-850, 1971.

[17] L. Hubert and P. Arabie, “Comparing partitions,” Journal of Classification, vol. 2, no. 1, pp. 193-218, 1985.