Regular paper

Journal of the KIECS. pp. 703-708, vol. 17, no. 4, Aug. 31. 2022, t. 114, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2022.17.4.703

통합 측도를 사용한 주성분해석 부공간에서의 k-평균 군집화 방법

류재홍[*]

K-Means Clustering in the PCA Subspace using an Unified Measure

Jae-Hung Yoo[*]

요 약

k-평균 군집화는 대표적인 클러스터링 기법이다. 하지만 성능 평가 척도와 최소 개수의 군집을 정하는 방법에 대하여 통합하지 못한 한계가 있다. 본 논문에서는 수치적으로 최소 개수의 군집을 정하는 방법을 도입한다. 설명된 분산을 통합측도로 제시한다. 최소 개수의 군집과 설명된 분산 달성을 동시에 만족하려면 주성분 해석의 부공간에서 k-평균 군집화 방법을 수행해야한다는 것을 제시하고자 한다. 패턴인식과 기계학습에서 왜 주성분 분석과 k-평균 군집화를 순차적으로 수행하는가에 대한 설명을 원론적으로 제시한다.

ABSTRACT

K-means clustering is a representative clustering technique. However, there is a limitation in not being able to integrate the performance evaluation scale and the method of determining the minimum number of clusters. In this paper, a method for numerically determining the minimum number of clusters is introduced. The explained variance is presented as an integrated measure. We propose that the k-means clustering method should be performed in the subspace of the PCA in order to simultaneously satisfy the minimum number of clusters and the threshold of the explained variance. It aims to present an explanation in principle why principal component analysis and k-means clustering are sequentially performed in pattern recognition and machine learning.

키워드

Explained Variance, K-Means Clustering, Numerical Elbow Method, PCA, Unified Measure
설명된 분산, K-평균 군집화, 수치화된 팔꿈치 방법, 주성분해석, 통합 측도.

Ⅰ. Introduction

In the analysis of the clustering method we established the following measures[1]. Total Sum of Square(TSS) is defined as following.

$$TSS = \sum_{i=1}^{n} \| \boldsymbol{x}_i - \boldsymbol{m} \|^2 \tag{1}$$

Here, n is the number of data, $\boldsymbol{x_i}$ is data vector, and $\boldsymbol{m}$ is mean vector.

[*] 교신저자: 전남대학교 컴퓨터공학과

• 접 수 일 : 2022. 06. 30
• 수정완료일 : 2022. 07. 21
• 게재확정일 : 2022. 08. 17

• Received : Jun. 30, 2022, Revised : Jul. 21, 2022, Accepted : Aug. 17, 2022

• Corresponding Author : Jae-Hung Yoo
Dept. of Computer Engineering, Chonnam Nat. Univ.
Email : jhy@jnu.ac.kr

Within-Cluster Sum of Square(WCSS) is defined as following.

$$WCSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \| \boldsymbol{x}_{ji} - \boldsymbol{m}_j \|^2 \tag{2}$$

Here, k is the number of clusters and n_j is the number of data in the cluster j. $\boldsymbol{x_{ji}}$ is data vector, and $\boldsymbol{m_j}$ is mean vector of cluster j.

Between-Cluster Sum of Square(BCSS) is defined as following.

$$BCSS = \sum_{j=1}^{k} n_j \| \boldsymbol{m}_j - \boldsymbol{m} \|^2 \tag{3}$$

BCSS becomes TSS as k increases to n, n_j decreases to one and $\boldsymbol{m_j}$ becomes $\boldsymbol{x_j}$.

Explained Variance(EV) is the performance measure to be increased. It is defined as the ratio of BCSS to TSS.

$$Var_E = \frac{BCSS}{TSS} \tag{4}$$

EV is the normalized version of BCSS.

Residual Variance(RV) is the error measure to be decreased. It is defined as the ratio of WCSS to TSS.

$$Var_R = \frac{WCSS}{TSS} = 1 - Var_E \tag{5}$$

RV is the normalized version of WCSS.

The minimum number of clusters k is defined as the clustering algorithm has the EV over the user specified threshold value such as 0.8, 0.85, or 0.9.

$$Var_E \geqq Th \tag{6}$$

Selecting the number of clusters k by elbow method[2] is meaningless if the EV of the clustering system is less than the user specified threshold value. In that case, we need the principal component analysis(PCA) feature reduction[3-6] to achieve the minimum number of clusters k determined by the elbow method.

The elbow method is traditional graphical tool showing dominant elbow point in the BCSS(RV) versus number of clusters k as in the Figure 1.

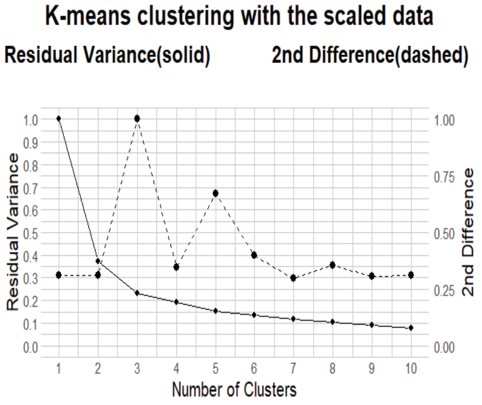

Fig. 1 Numerical Elbow method

Here, we introduce the second order difference that is calculated as the numerical measure of curvature for each k. It is normalized having the maximum value one. The elbow point is coincided by the value one.

In this paper, the k-means clustering algorithm[7-11] in the PCA subspace is formulated for the feature reduction and finding the minimum number of clusters that satisfying the user specified threshold on the EV.

In section Ⅱ, EV of PCA is introduced from SVD formulation. In section Ⅲ, k-means clustering algorithm in the PCA subspace is developed. In section IV, experimental results show the effectiveness of the proposed method followed by the conclusion and reference sections[12-15].

Ⅱ. Explained Variance of PCA

In the PCA space, the covariance matrix is diagonalize. The EV is the ratio cumulative sum of diagonal term to trace of the matrix in the R project language as following.

$$svd1 \ = \ svd(scale(df)) \qquad (7)$$

$$trace2 \ = \ sum(svd1\$d\hat{\ }2)$$

$$EV_i \ = \ svd1\$d\hat{\ }2/trace2$$

$$Var_{PCA} = \ cumsum(EV_{i)}$$

It is the indicator of feature selection or reduction that filter the noise from the signal. Here it can be the criterion to select principal components of PCA to search for the optimal number of clusters for the user specified threshold value.

$$Var_{PCA}(j) \geqq Th \qquad (8)$$

It can be checked for j-1 to j+1 dimensional PCA components or subspaces.

Equation (8) is the local criteria for the clustering performance. The global measure can be the one factored by the $Var_E(j)$.

$$Var_{PCA}(j)\, Var_E(j) \geqq Th \qquad (9)$$

This is due to the fact that the total sum of square in the scaled data is the same amount of trace2 in the equation (7).

Ⅲ. K-means Clustering Algorithm in the PCA Subspace

We can start with the scaled version of original data and check the elbow point as in the previous Figure 1. If the elbow point is identified but not satisfied the user specified threshold for the explained variance, It can be searched on the PCA subspaces. The search of PCA subspace can begin from the first PCA component but may be not necessary for the high dimensional input data The search space can be limited using Equation (8). An appropriated termination condition is satisfied for the elbow point above the threshold value. We can designate the method as K-means clustering algorithm in the PCA subspace as in the Table 1.

Table 1. K-means clustering algorithm in the PCA subspace

1. Check for the elbow point in the scaled data.
2. Get j using equation (8).
3. Perform elbow point search using PCA for each subspace j-1, j, j+1.
4. Decide the optimal space for the threshold value and elbow point.

Ⅳ. Experimental Results

We report the experiments with the algorithm proposed in the previous section. We use the Fisher's Iris data set from UCI machine learning Repository[12,13]. It has 4 input dimension and 3 class output for the classification problem. It has 50 examples for each class and total is 150 data examples. Clustering problem is to use the input data only for the unsupervised learning. The algorithm has applied to the scaled version of Iris data. Figure 1 shows that the elbow point is clearly found at the k = 3. But explained variance is 0.76696 is less than the any threshold value over 80. Thus we seek the PCA sub space that satisfies the elbow point getting over the prespecified threshold.

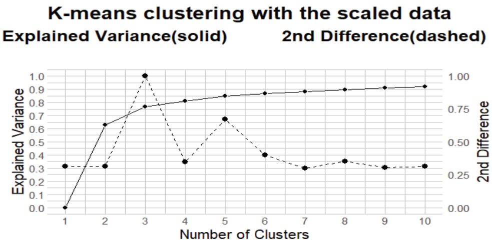

Fig. 2 Elbow point using explained variance

Step 2 in the algorithm is to check the explained variance in the each PCA component.

Table 2. Explained variance in the PCA subspace

	pc1	pc2	pc3	pc4
EV	0.729624	0.228508	0.036689	0.005178
CEV	0.729624	0.958132	0.994821	1.000000

Here, EV is explained variance in each PCA component an CEV is cumulative sum of EV. Using equation (8), we have j = 2. Thus in step 3 in the algorithm we check from j = 1 to j = 3 for finding optimal combination of subspace for elbow point and explained variance in the clustering.

Figure 3, 4, 5 and 6, show the elbow point using explained variance.

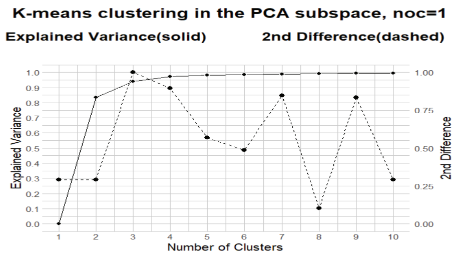

Fig. 3 Elbow point using EV at (3, 0.94032) with number of principal components 1

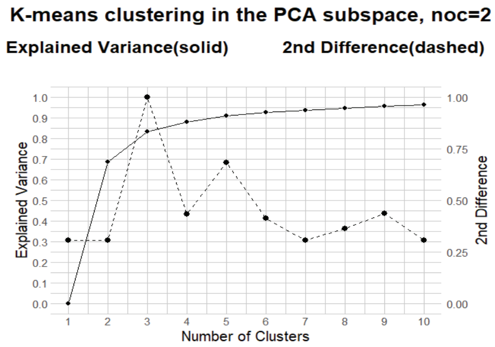

Fig. 4 Elbow point using EV at (3, 0.83347) with number of principal components 2

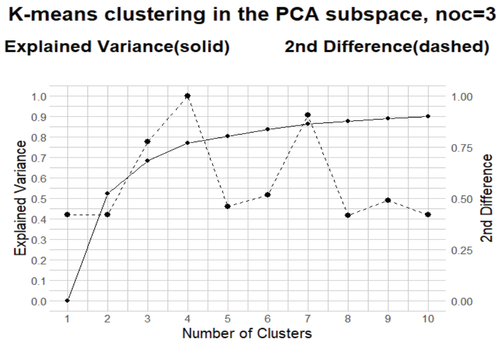

Fig. 5 Elbow point using EV at (4, 0.87852) with number of principal components 3

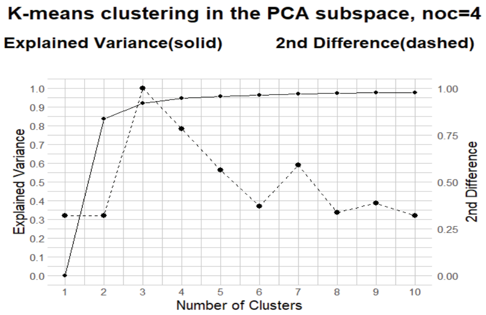

Fig. 6 Elbow point using EV at (3, 0.91913) with number of principal components 4

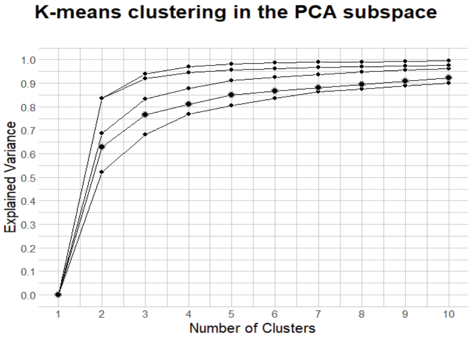

Fig. 7 Elbow point using EV local

Table 3. Explained variance of k-means clustering in the PCA subspace and scaled data(SO) with local measure

k	SO	PC1:1	PC1:2	PC1:3	PC1:4
1	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.6294	0.8355	0.6867	0.5217	0.8359
3	0.7670	0.9403	0.8335	0.6814	0.9191
4	0.8099	0.9710	0.8785	0.7676	0.9458
5	0.8487	0.9810	0.9106	0.8038	0.9567
6	0.8667	0.9862	0.9250	0.8372	0.9634
7	0.8822	0.9893	0.9358	0.8639	0.9684
8	0.8959	0.9916	0.9457	0.8763	0.9713
9	0.9057	0.9933	0.9542	0.8878	0.9743
10	0.9213	0.9945	0.9620	0.8991	0.9764

Figure 7 summarizes k-means clustering in the PCA subspace using local measure and in the scaled data with larger dots. The numerical details are shown in the table 3. Here, the shaded cell denotes the elbow point. First principal component(PC) only gives the maximum explained variance. We can choose the first 2 PCs as next candidate.

Figure 8 summarizes k-means clustering in the PCA subspace using global measure and in the scaled data with larger dots. Here, $Var_E(j)$ is factored by the $Var_{PCA}(j)$. PC1:2 and PC1:4 are satisfy the constraint of equation (9). The numerical details are shown in the table 4.

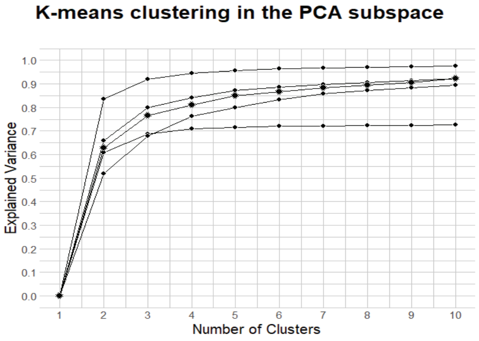

Fig. 8 Elbow point using EV global

Table 4. Explained variance of k-means clustering in the PCA subspace and scaled data(SO) with global measure

k	SO	PC1:1	PC1:2	PC1:3	PC1:4
1	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.6294	0.6096	0.6579	0.5190	0.8359
3	0.7670	0.6861	0.7986	0.6779	0.9191
4	0.8099	0.7085	0.8417	0.7636	0.9458
5	0.8487	0.7158	0.8725	0.7996	0.9567
6	0.8667	0.7196	0.8863	0.8329	0.9634
7	0.8822	0.7218	0.8966	0.8594	0.9684
8	0.8959	0.7235	0.9061	0.8718	0.9713
9	0.9057	0.7247	0.9142	0.8832	0.9743
10	0.9213	0.7256	0.9217	0.8944	0.9764

Ⅴ. Conclusions

In this paper, k-means clustering algorithm in the PCA subspace is developed. We present an explanation in principle why principal component analysis and k-means clustering are sequentially performed. Next plan is to apply the algorithm to the practical high dimensional data using the local measure and the global measure.

References

[1] R. Duda and P. Hart, *Pattern Classification and Scene Analysis.* New York: John Wiley & Sons, 1973.

[2] R. L. Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, no. 4, 1953 pp. 267-276.

[3] R. C. Gonzalez and R. E. Woods, *Digital Image Processing.* Reading, MA: Addison-Wesley, 1992.

[4] S. Cen, J. Yoo, and C. Lim, "Electricity Pattern Analysis by Clustering Domestic Load Profiles Using Discrete Wavelet Transform," *Energies.* vol. 15. no. 4, 2022, pp. 1350(1-18).

[5] Y. Tong, I. Aliyu, and C. Lim, "Analysis of Dimensionality Reduction Methods Through Epileptic EEG Feature Selection for Machine Learning in BCI," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 13, no. 6, 2018, pp. 1333-1342.

[6] Y. Kim, S. Park, and D. Kim, "Research on Robust Face Recognition against Lighting Variation using CNN," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 12, no. 2, 2017, pp. 325-330.

[7] J. Kim and C. Kim, "Image Retrieval System of semantic Inference using Objects in Images," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 7, 2016, pp. 677-684.

[8] J. Park and S. Lee, "An Image Processing Mechanism for Disease Detection in Tomato Leaf," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 4, no. 5, 2019, pp. 959-968.

[9] B. Kim, H. Yoon, and J. Lee, "A Study on the Distribution of Cold Water Occurrence using K-Means Clustering," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 2, 2021, pp. 371-378.

[10] C. Lee, "Enhancement of the k-Means Clustering Speed by Emulation of Birds' Motion in Flock," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 9, no. 9, 2014, pp. 965-970.

[11] C. Lee, "The Effect of the Number of Phoneme Clusters on Speech Recognition," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 9, no. 11, 2014, pp. 1221-1226.

[12] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*. vol. 7, no. 2, 1936, pp. 179-188.

[13] D. Dua and C. Graff, "UCI Machine Learning Repository[http://archive.ics.uci.edu/ml]," University of California, School of Information and Computer Science, Irvine, CA., 2019.

[14] J. Yoo, "A Unified Bayesian Tikhonov Regularization Method for Image Restoration," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 11, 2016, pp. 1129-1134.

[15] J. Yoo, "An Extension of Unified Bayesian Tikhonov Regularization Method and Application to Image Restoration," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 15, no. 11, 2020, pp. 161-166.

저자 소개

류재흥(Jae-Hung Yoo)

1981년 한양대학교 기계공학과 졸업(공학사) (BE in Mechanical Engineering from Hanyang Univ. in 1981)

MA in Computer Science from Univ. of Detroit in 1986

PhD in Computer Science from Wayne State Univ. in 1993

Joined as a faculty member in the Dept. of Computer Engineering, Yosu Nat. Univ. in 1994

Became a faculty member in the Dept. of Computer Engineering, Chonnam Nat. Univ. in 2006

※(Main research areas: Artificial Neural Networks, Pattern Recognition, Machine Learning, Image Processing and Computer Vision)