Improved Cluster Partition in Principal Component Analysis Guided Clustering

S. M. Shaharudin
Dept. of Mathematical Sciences,
Universiti Teknologi Malaysia,
Malaysia.

N. Ahmad
Dept. of Mathematical Sciences,
Universiti Teknologi Malaysia,
Malaysia.

F. Yusof
Dept. of Mathematical Sciences,
Universiti Teknologi Malaysia,
Malaysia.

ABSTRACT

Principal component analysis (PCA) guided clustering approach is widely used in high dimensional data to improve the efficiency of K- means cluster solutions. Typically, Pearson correlation is used in PCA to provide an eigen-analysis to obtain the associated components that account for most of the variations in the data. However, PCA based Pearson correlation can be sensitive on non-Gaussian distributed data, which involve skewed observations such as outlying values. Thus, applying PCA based Pearson correlation on such data could affect cluster partitions and generate extremely imbalanced clusters in a high dimensional space. In this study, Tukey's biweight correlation based on M-estimate approach in PCA is used as an alternative to Pearson correlation. This approach is more resistant to outlying values as it examines each observation and down weight those that lie far from the center of the data. In particular two major features are highlighted: (1) fewer components are retained and imbalanced clusters at the recommended cumulative percentage of variation threshold is avoided; (2) the cluster quality with respect to external, internal and relative criteria as shown in Rand, Silhouette and Davies-Bouldin indices, outperform that of the clusters from PCA based Pearson correlation.

General Terms

Data Structures and Algorithms.

Keywords

Tukey's biweight, K-means, Principal Component Analysis.

1. INTRODUCTION

Cluster analysis has been used in various disciplines such as biology, marketing and hydrology to partition observations of similar patterns to the same cluster and dissimilar patterns to different clusters[1]. Principal component analysis (PCA) is a reduction dimension technique which is often used as a pre-processing method to guide the process of grouping items in order to improve the efficiency and accuracy of cluster solutions [2][3]. The main idea of PCA is to reduce the dimensionality of data set consisting of a large number of interrelated variables, while retaining as much as possible the variation present in the data set.

A typical approach in PCA requires the use of configuration points of entities between the rows and column of the data based on Pearson correlation matrix. Pearson correlation matrix is calculated by finding the covariance of variables and dividing it by the square root of the product of the variances. As each pair of observations is equally weighted, Pearson correlations can be sensitive on non-Gaussian distributed data, which could involve skewed observations such as outlying values [4]. Thus, applying PCA based Pearson correlation on such data could affect cluster partitions and generate extremely imbalanced clusters in a high dimensional space.

In this paper, weighting on the observations is used as a resistant measure by introducing a Tukey's biweight correlation matrix as an alternative to Pearson correlation matrix in PCA to provide a robust cluster partitions with respect to cluster validity indices.

2. METHODOLOGY

2.1 K-Means Clustering Algorithm

K- means is a method in cluster analysis to partition observations into *k* pre-determined number of disjoint clusters. This algorithm consists of two separate steps run iteratively until convergence. The first step is to define *k* centroids for each cluster and the next step is to assign each data object to the nearest centre. Euclidean distance method is generally used to determine the distance between each data points and the cluster centres. The *k* means clustering algorithm works as follows:

Step 1: Choose *k* randomly from the data set as initial cluster centre

Step 2: Calculate the distance between each data points and assign each item to the cluster which has the closest centroid. Recalculate the cluster centre for each cluster until convergence criterion is met.

2.2 Principal Component Analysis

PCA is designed to reduce the dimension of large data matrix to a lower dimension by retaining most of the original variability in the data [5]. This is achieved by converting a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called principal components. The first principal component accounts for as much of the variation in the original data. Then each succeeding component accounts for as much of the remaining variation subject to being uncorrelated with the previous component.

Covariance or correlation matrix derived from the data matrix plays an important role in PCA to calculate its eigenvalues and eigenvectors to obtain the associated components that account for most of the variations in the data. [6]. For the purpose of this study, correlation matrix is used. It is generally recommended taking at least 70% of cumulative percentage of total variation as a benchmark to cut off the eigenvalues in a large data set for extracting the number of components [7]. The reduced matrix is the component matrix of eigenvector “loadings” which defines the new variables consisting of linear transformation of the original variables that maximizes the variance in the new axes.

The steps involved in PCA algorithm are as follows:

Step 1 : Obtain the input matrix.

Step 2 : Calculate its correlation matrix.

Step 3 : Calculate the eigenvectors and eigenvalues of the correlation matrix.

Step 4 : Select the most important principal components based on cumulative percentage of total variation.

Step 5 : Derive the new data set

2.2.1 Pearson correlation matrix

In applications such as environmental sciences and climatology, Pearson correlation is typically used in PCA for calculating its eigenvectors and eigenvalues [8],[9],[10],[11]. Pearson correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. Typically, Pearson correlation is used to measure the distance (or similarities) before implementing a clustering algorithms. The Pearson correlation coefficient between two vectors of observations is as follows :

$$r_{ij} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)(X_j - \bar{X}_j)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_i)^2 \sum_{i=1}^{n}(X_j - \bar{X}_j)^2}} \qquad (1)$$

where X_i and X_j refer to the vectors of observations in matrix data X with n observations, with $\bar{X}_i$ and $\overline{X}_j$ refer to the mean of the vectors.

2.2.1 Tukey's biweight correlation

Tukey's biweight correlation is based on Tukey's biweight function that relies on M-estimators used in robust correlation estimates. M-estimate has a derivative function, ψ which determines the weights assigned to the observations in the data set. It has the ability to down weight observations to reflect its influence from the centre of the data [12]. The derivative function is derived as follows:

$$\psi(u) = \begin{cases} u(1-u)^2 & |u| \ \leq 1 \\ 0 & |u| \ > 1 \end{cases} \qquad (2)$$

It can be seen that if $|u|$is large enough, then $\psi(u)$ reduces to zero. One of the important aspects to measure the resistance to outlying data values of M-estimators is its breakdown point. A breakdown point is the smallest fraction of contamination that can cause an inaccurate result [13]. In this study, Tukey's biweight with breakdown points at 0.0, 0.2, 0.4 and 0.5 are compared and breakdown point of 0.4 performs the best. According to [14], typically a breakdown point of 0.4 performs better in most situation and the result is more accurate and efficient when compared to a lower breakdown point.

The biweight estimate of correlation is produced by first calculating the location estimate, $\tilde{T}$ and then updating the shape estimate, $\tilde{S}$. The (i,j)th element of $\tilde{S}$, i.e. $\tilde{s}_{ij}$ acts as a resistant estimate of the covariance between the two vectors, X_i and X_j. The biweight correlation of these two vectors is calculated as follows :

$$\tilde{r}_{ij} = \frac{\tilde{s}_{ij}}{\sqrt{\tilde{s}_{ii}\tilde{s}_{jj}}} \qquad (3)$$

with

$$T_n^{(k+1)} = \frac{\sum_{i=1}^{n} X_i w\left(u_{i^{(k)}}\right)}{\sum_{i=1}^{n} w(u_{i^{(k)}})} \qquad k = 0,1,2, \ldots \qquad (4)$$

$$S_n^{(k+1)} = \frac{\sum_{i=1}^{n} w\left(u_{i^{(k)}}\right)\left(X_i - T^{(k+1)}\right)(X_i - T^{(k+1)})^t}{\sum_{i=1}^{n} w(u_{i^{(k)}})(u_{i^{(k)}})} \qquad (5)$$

where $T_n^{(k+1)}$is a location vector and $S_n^{(k+1)}$is a shape matrix such that $k = 0,1,2, \ldots.$

Thus, a PCA based Tukey’s biweight correlation for K-means cluster analysis is more likely to produce a better cluster partition that is more resistant to outlying values than Pearson correlation in PCA.

3. PROPOSED METHOD

As Pearson correlation is likely to be more sensitive to non-Gaussian distributed data, Tukey's biweight correlation in PCA on the original data set in a PCA guided clustering setting is proposed. Before proceeding, the original data matrix is standardized by a robust location and scale estimator to avoid any masking or swamping effect [15].

The reduced data set is then applied to K-Means cluster analysis to obtain cluster partitions. K means method requires specifying the number of clusters before the algorithm is applied. To overcome this problem, Calinski and Harabasz Index [16] is used as a measure to determine the optimal number of cluster partition for the input data. This is indicated by the maximum value of the index.

The steps involved in the proposed algorithm are as follows :

Step 1 : Obtain the input matrix.

Step 2 : Standardize the observation with median and mean absolute deviation (MAD), i.e.

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}}{median\left(\left|x_{ij} - median\left(x_{ij}\right)\right|\right)} \qquad (6)$$

such that x_{ij} refer to elements in the input matrix.

Step 3 : Set the breakdown point for the Tukey's biweight correlation at 0.4

Step 4 : Calculate the Tukey's biweight correlation matrix.

Step 5 : Calculate the eigenvectors and eigenvalues of the correlation matrix.

Step 6 : Select the most important principal components based on cumulative percentage of total variation.

Step 7 : Derive the new data set

Step 8 : Calculate Calinski and Harabasz index in new data set to determine the best number of cluster

Step 9 : Apply K-means method to new data set

4. EXPERIMENTAL RESULTS

The proposed algorithm is evaluated on a set of daily rainfall data, which is known to have a non-Gaussian distribution and is likely to be skewed [17]. The data is obtained from the Department of Irrigation and Drainage, Malaysia, i.e. *Jabatan Pengairan dan Saliran (JPS)* for the period 1975-2007. The daily rainfall data from 75 stations over Peninsular Malaysia are located at different geographical coordinates on four regions, east, southwest, west and northwest. In this study, the occurrence of episodes on extreme rainfall event described as torrential rainfall is focused on. It was therefore necessary to choose some criteria that would lead to the establishment of a threshold, in order to allow for a clear distinction between what constitutes a day of torrential rainfall in the Peninsular Malaysia region and what does not. The most common threshold applied for this purpose in a tropical climate is an amount of 60 mm/day. The filtered days with rainfall that exceeds 60 mm in at least 2% of the stations are used. This criterion yielded 250 days and 15 rainfall stations, which is an adequate number to represent the main torrential centers. In this study, clustering results achieved by the PCA based Pearson correlation and PCA based Tukey's biweight correlation are compared.

Table 1. Number of components retained in PCA and the number of clusters based on Pearson and Tukey's Biweight Correlation

Cum. %	Number of components		Number of cluster, *k*	
	Tukey's biweight	**Pearson**	**Tukey's biweight**	**Pearson**
60	11	12	12	2
65	13	14	12	2
70	15	19	10	2
75	22	26	6	2
80	28	35	2	2

Table 1 shows the relationship between cumulative percentage of variance and number of clusters obtained using two different approaches in PCA based correlation matrix. From Table 1, it can be seen that the number of components between two different approaches in correlation matrix differ at the same level of cumulative percentage of variation. It appears that Tukey's biweight correlation requires less number of components to extract in order to achieve at least 70% of cumulative percentage of variation compared to Pearson. For instance, 28 components is retained with Tukey's as compared to 35 with Pearson's at 80% cumulative percentage of variation. In climate data, extracting too many components is not favorable as it may reflect variations of low frequency or spatial scale that are not important [8][18].

In terms of cluster partitions, Table 1 also shows that in contrast to Pearson's, Tukey's biweight correlation is more sensitive to the number of clusters according to the number of components retained. The number of clusters as a result of PCA-based Pearson correlation, appear to stabilize at only two clusters regardless of the cumulative percentage of variation used. In climatology studies particularly in identifying rainfall patterns, it is more reasonable to obtain more than two cluster partitions to explain the various types of rainfall patterns. Thus, two clusters clearly is not appropriate as it mask the true structure of the data.

Table 2. Indices to measure the quality of clustering results

Correlation	Rand Index	Silhouette Index	Davies-Bouldin Index
Tukey's biweight	0.55	0.1	2.02
Pearson	0.53	0.04	4.78

In order to evaluate the cluster solutions, the clustering output at 70% cumulative percentage of variation on PCA-based Tukey's biweight correlation (10 clusters) and Pearson correlation (2 clusters) are chosen respectively. These clusters are evaluated based on three fundamental criteria of quality cluster as prescribed by [19]: external criteria, internal criteria and relative criteria using Rand Index, Silhouette Index, Davies-Bouldin Index respectively. As a guideline, a higher value of Rand and Silhouette index and a lower value of Davies-Bouldin index should indicate a good quality cluster. Table 2 illustrates that Tukey's biweight correlation show a relatively better clustering results in terms of the three indices when compared to Pearson correlation. Figure 1 illustrates the cluster partitions for both approaches, set at 5 clusters on a two-dimensional scatter plot matrix. The figure shows a much clearer partition of membership clusters based on PCA based Tukey's biweight correlation when compared to that of Pearson.

5. CONCLUSION

In this paper, PCA based Tukey's biweight correlation has been proposed to guide the cluster solution of K-means clustering method. The purpose is to introduce an alternative correlation matrix due to the issues when dealing with non-Gaussian distributed data particularly when the data is skewed in nature. This study shows a substantial improvement in the cluster partition with PCA based Tukey's biweight correlation than Pearson's to avoid inaccurate imbalanced clusters in high dimensional space. The quality of clustering results has been proven by the validity indices to indicate better internal, external and relative cluster quality.

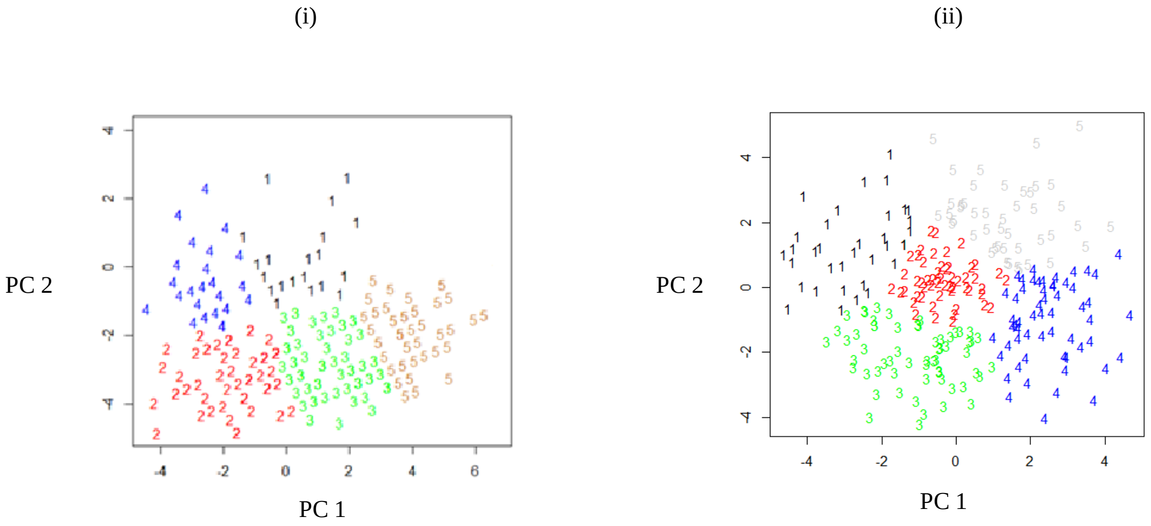

Figure 1 : Scatter plot of the data with respect to their clustering features based on (i) Tukey's biweight correlation matrix and (ii) Pearson correlation matrix

6. ACKNOWLEDGMENTS

The authors would like to thank Universiti Teknologi Malaysia and the Malaysian Ministry of Higher Education for their financial funding through GUP grant Vote No. (Q.J130000.2626.04J09).

7. REFERENCES

[1] Marghny, M.H., Abd El-Aziz, R.M., Taloba, A.I. 2011. An Effective Evolutionary Clustering Algorithm: Hepatitis C Case Study. International Journal of Computer Applications. Vol 34-No.6.

[2] Indhumathi, R. and Sathiyabama, S.2010. Reducing and Clustering High Dimensional Data Through Principal Component Analysis. International Journal of Computer Applications. Vol 11-No.8.

[3] Ding, C. and Xiaofeng, H. 2004. K-means Clustering via Principal Component Analysis. In proceedings of the 21st International Conference on Machine Learning, Canada.

[4] Kendall, M. G. and Stuart, A. 1958. The Advanced Theory of Statistics. New York..

[5] Everitt, B. S. and Dunn, G. 2001. Applied Multivariate Data Analysis. London: Arnold Publisher.

[6] Neware, S., Mehta, K. and Zadgaonkar, A.S. 2013. Finger Knuckle Identification using Principal Component Analysis and Nearest Mean Classifier. International Journal of Computer Applications. Vol 70-No.9.

[7] Jolliffe, I.T. 2002. Principal Component Analysis (2nd ed.). New York,Inc. : Springer-Verlag.

[8] Penarrocha, D., Estrela, M.J., and Millan, M. 2002. Classification of Daily Rainfall Patterns in a Mediterranean Area with Extreme Intensity Levels: The Valencia Region. Internation Journal of Climatology, Vol 22, 677-695.

[9] Romero, R., Ramis, C., and Guijarro, J.A. 1999. Daily Rainfall Patterns in the Spanish Mediterranean Area: An Objective Classification. International Journal of Climatology. Vol 19, 95-112.

[10] Sumner, G., Guijarro. J.A., and Ramis, C. 1995. The Impact of Surface Circulations on The Daily Rainfall Over Mallorca. International Journal of Climatology. Vol 15, 673–696.

[11] Wickramagamage, P. 2010. Seasonality and spatial pattern of rainfall of Sri Lanka: Exploratory factor analysis. International Journal of Climatology. Vol 30, 1235-1245.

[12] Hardin, J., Mitani. A., Hicks. L. and Vankoten. B. 2007. A Robust Measure of Correlation Between Two Genes on A Microarray. BMC Bioinformatics. Vol 8, 220.

[13] Rousseeuw, P. J. and Leroy, A. M. 2003. Robust Regression and Outlier Detection. New Jersey: John Wiley & Sons, Inc.

[14] Owen, M. 2010. Tukey's Biweight Correlation and the Breakdown. Thesis. Pomona College.

[15] Choulakian, V. 2001. Robust Q-Mode Principal Component Analysis in L_1. Computational Statistics & Data Analysis. Vol 37, 135-150.

[16] Maulik, U. 2002. Performance Evaluation of SomeClustering Algorithms and Validity Indices. IEEE Transactions of Pattern Analysis and Machine Intelligence. Vol. 24, No. 12.

[17] Cui, K. 2012. Semiparametric Gaussian Variance-Mean Mixtures for Heavy-Tailed and Skewed Data. ISRN Probability and Statistics, vol. 2012, Article ID 345784, 18 pages, 2012. doi:10.5402/2012/345784

[18] Mimmack, G.M., Mason S.J. and Galpin, J.S.2002. Choice of Distance Matrices In Cluster Analysis : Defining Regions. Journal of Climate. Vol 14, 2790-2797.

[19] Everitt, B. S. , Landau, S. and Leese, M. 2001. Cluster Analysis. London: Arnold Publisher.