MultiCraft

International Journal of Engineering, Science and Technology
Vol. 6, No. 1, 2014, pp. 20-26

INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND TECHNOLOGY

www.ijest-ng.com
www.ajol.info/index.php/ijest

Notes about the paper entitled "*A hybridized K-means clustering approach for high dimensional dataset*"

A.J. Arriaza-Gómez[1], F. Fernández-Palacin[2], M. Muñoz-Marquez[3], S.M. Pérez-Plaza[4]

[1,2,3,4] Department of Statistics and Operations Research, University of Cádiz, SPAIN
[1]antoniojesus.arriaza@uca.es, [2]fernando.fernandez@uca.es,[3]manuel.munoz@uca.es, [4]sonia.perez@uca.es
[1, 4] Tel. +34-956-012834, [2] Tel. +34-956-012772, [3] Tel. +34-956-012784, Fax. +34- 956-016565

Abstract

In the paper *"A hybridized K-means clustering approach for high dimensional dataset"* Dash, Mishra, Rash and Acharya have presented a new version of the *k*-means algorithm. In it, principal components analysis (PCA) was used before applying the *k*-means algorithm with a new initialization method. The authors compare the results obtained by using the HKMCA and PCA with the results of the original *k*-means, but a direct comparison is not valid as this paper shows.

Keywords: Cluster analysis, *K*-means Algorithm, Principal Component Analysis, Hybridized *K*-means algorithm

DOI: http://dx.doi.org/10.4314/ijest.v6i1.2

1. Introduction

The *k*-means algorithm is a widely used algorithm for classification problems. In order to improve the *k*-means algorithm performance on high dimensionality datasets a new method has been presented in Dash *et al.* (2010). The hybridized *k*-means clustering algorithm (HKMCA) uses the principal components analysis to reduce the dimension of data before applying the *k*-means algorithm. The above mentioned paper also shows a new method to select the initial centroids that seems improves the algorithm behaviour. The *k*-means algorithm looks for a partition of the original data into *k* groups that minimizes the sum of squared distances between the points and the centers of the groups. The main drawbacks of this method are its stochastic behaviour and the high computational effort needed for large data sets. A lot of authors have addressed such issues. The original *k*-means algorithm and the HKMCA results are analysed by using four datasets in Dash *et al.* (2010). In this paper, the methodology used in Dash *et al.* (2010) to compare both cluster procedures is studied in order to clarify some guidelines when two clustering methods are compared. Also we have proved that the HKMCA does not achieve the experimental results showed in Dash *et al.* (2010) when these guidelines are followed.

In Section 1, a review of the paper of Dash *et al.* (2010) is made and the goal of this paper is proposed. Section 2 contains some previous concepts about cluster analysis, the *k*-means algorithm, and principal components analysis. Section 3 describes the procedure proposed in Dash *et al*. (2010). In Section 4, the authors repeat the steps followed in Dash *et al.* (2010) when synthetic dataset is used and prove that the HKMCA has been programmed properly. Experimental activities and its corresponding result discussion have been done in Section 5. Finally, Section 6 is devoted to the final conclusions.

2. Previous concepts

2.1 Cluster Analysis

Cluster analysis is an important part of the knowledge discovery process that is used in many fields of study. The utility of cluster analysis has been proved in many papers, the main goal is to group either the data units into clusters such that the elements within a cluster have a high degree of "natural association" among themselves while the clusters are "relatively distinct" from one another. A type of measure of similarity must be defined which indicates if two data units are in the same cluster.

2.2 K-means algorithm

The *k*-means algorithm is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. It is a non-hierarchical technique that tries to find *k* groups (*k* fixed a priori) from a data set.

Given a natural number *k* and a dataset $X = \{x_1, x_2, ..., x_m\}$ with *m* instances and *n* numeric variables, the *k*-means main objective is to find *k* centroids, one for each group, in order to minimize the within groups sum of squared errors (SSE): $SSE_x(C) = \sum_{i=1}^{m} d^2(x_i, C)$, where $d(x_i, C) = \min_{j=1}^{k}\{d(x_i, c_j)\}$, $C = \{c_1, c_2, ..., c_k\}$, and $d(x_i, c_j)$ is the Euclidean distance between x_i and c_j, and c_j is the *j-th* centroid.

There are many improvements and versions of the *k*-means algorithm. MacQueen (1967), Lloyd (1982), Forgy (1965) and Hartigan-Wong (1979) are the most used versions. Below we describe the classic *k*-means method:

Algorithm:

Step 1 Initialization

Choose *k* points randomly in the featuring space, usually points of dataset are selected. They are called initial centroids or initial seeds.

Step 2 Assignation

For each point in the dataset, find the centroid that is closest, and assign that point to the corresponding cluster.

Step 3 Mean update

Re-calculate *k* new centroids as barycentre of the clusters, using the mean of all points in the same cluster.

Step 4 Stopping rule

Repeat steps two and three until the stopping rule is satisfied. Several stopping rules exist, for instance: when no more changes take place among the centroids, when the decrease of objective function is less than a prefixed threshold or when a prefixed number of iterations is reached.

2.3 Principal components analysis

The main purpose of principal component analysis (PCA) is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variability present in the data set. See Jolliffe (2002). PCA provides an orthogonal linear map which changes the original dataset coordinates to new ones. This linear map defines an orthonormal matrix whose columns are called principal components. The first principal component retains most of the variability present in the dataset; the second principal component retains the second largest variance, and so on. There are different rules to choose the number of variables that should be selected in each case. According to Jolliffe (2002), the most common criterion is selecting the number of principal components such that the percentage of variation accounted for these components within the range 70% to 90%.

Other criteria to select the appropriate number of principal components are:

- Choose those principal components whose variance is greater than a fixed threshold value.
- Select those principal components whose variance is greater than the mean of variances of principal components.

The last criterion is selected in Dash *et al.* (2010).

Using principal components analysis before applying the algorithm with a new initialization method is proposed in Dash *et al.* (2010).

3. Description of the hybridized *k*-means clustering algorithm

For completeness the hybridized *k*-means clustering algorithm (HKMCA) in Dash *et al.* (2010), is included here:

Input: $X = \{d_1, d_2, ..., d_n\}$ // set of n data items.
K // Number of desired clusters.
An array *Cen*[ ] having size *k* initially being empty.

Output: A set of *k* clusters

// Phase-1: Apply PCA to reduce the dimension of the data set
1. Organize the dataset in a matrix *X*.
2. Normalize the data set using Z-score.
3. Calculate the singular value decomposition of the data matrix. $X = UDV^T$
4. Calculate the variance using the diagonal elements of *D*.
5. Sort variances in decreasing order.
6. Choose the p principal components from *V* with largest variances.
7. Form the transformation matrix *W* consisting of those *p* PCs.
8. Find the reduced projected dataset *Y* in a new coordinate axis by applying *W* to *X*.

//Phase-2: Find the initial centroids
9. Set *m*=1.
10. Compute the distance between each data points in the set *Y*.
11. Choose the two data points y_i and y_j such that distance (y_i, y_j) is maximum.
12. $Cen[m] = y_i; Cen[m+1] = y_j; m = m+2;$
13. Remove the two objects y_i, y_j from *Y*.
14. While $(m \le k)$
    1. Find the distance of each object in *Y* to *Cen*[*i*], for *i*=1 to *m*-1.
    2. Find the average of all the distances to the centroid for each object in *Y*.
    3. Choose the data object y_0 having maximum average distance from previous centroids.
    4. $Cen[m] = y_0; m = m+1;$
    5. Remove the object y_0 from *Y*.

// Phase-3: Apply the *k*-means clustering with the initial centroids given in array *Cen*.
15. For each data point, in set *Y*, find the nearest cluster center from list *Cen* that is closest and assign that data point to the corresponding cluster.
16. Update the cluster centres in each cluster using the mean of the data points, which are assigned to that cluster.
17. Repeat the steps 15 and 16 until there are no more changes in the values of the centroids.

4. Reproducibility of experimental results on the datasets

In Dash *et al.* (2010) the hybridized *k*-means clustering algorithm is applied to three well-known datasets: Pima Indian Diabetes, Breast Cancer, SPECTF Heart and a fourth synthetic dataset with 15 data objects and 10 variables. In order to study the behaviour of this method the algorithm has been re-implemented and we have used the same datasets.

The results obtained with the synthetic dataset are shown in Table 2. The SSE value obtained with the HKMCA is the same as the SSE value obtained in Dash *et al.* (2010). The table shows the number of principal components (NPC) selected. The *k*-means algorithm is a phase of the HKMCA. The *k*-means algorithm is applied to the data that are in the subspace generated by these components in the HKMCA case.

Table 1. Synthetic dataset with 15 data objects and 10 variables

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Data 1	1	5	1	1	1	2	1	3	1	1
Data 2	2	5	4	4	5	7	10	3	2	1
Data 3	3	3	1	1	1	2	2	3	1	1
Data 4	4	6	8	8	1	3	4	3	7	1
Data 5	5	4	1	1	3	2	1	3	1	1
Data 6	6	8	10	10	8	7	10	9	7	1
Data 7	7	1	1	1	1	2	10	3	1	1

Table 1 (cont'd). Synthetic dataset with 15 data objects and 10 variables

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Data 8	8	2	1	2	1	2	1	3	1	1
Data 9	9	2	1	1	1	2	1	1	1	5
Data 10	10	4	2	1	1	2	1	2	1	1
Data 11	11	1	1	1	1	1	1	3	1	1
Data 12	12	2	1	1	1	2	1	2	1	1
Data 13	13	2	1	1	1	2	1	2	1	1
Data 14	14	5	3	3	3	2	3	4	4	1
Data 15	15	1	1	1	1	2	3	3	1	1

Table 2. SSE values when the original *k*-means and the hybridized *k*-means clustering algorithms are applied in synthetic dataset

Dataset	No of instances	No of variables	Original *k*-means algorithm SSE	HKMCA SSE	NPC
synthetic	15	10	560.3373	47.80006	3

In Dash *et al.* (2010) the SSE value is averaged out from the results of 10 runs when the original *k*-means algorithm is applied. In order to replicate the results and avoid the different initialization problem the algorithm has been run 1000 times to synthetic dataset. According to Dash *et al.* (2010), the average of SSE is 608.446. This value is one of the possible results, obtained 34. 2% of times (see Table 3) and the average value in our analysis is 560.3373. If the SSE value had been calculated using the *k*-means algorithm with normalized synthetic dataset, then the SSE coefficient would have been much smaller (see Table 4). As data have not been standardized when the original *k*-means has been used in this case, we suppose the procedure used in that paper was the same in the remaining cases.

Table 3. SSE values when the *k*-means algorithm is applied 1000 times and its classification percentages

Sum of Squared Error	506.000	608.446	653.429	602.722	791.000	838.417	841.732
Percentage for each cluster classification	56.1%	34.2%	3.8%	2.2%	1.6%	1.2%	0.9%

For the remaining datasets the algorithm is applied in the original work by using a random sample from the datasets. For this reason, the results would be different with each sample.

5. Experimental analysis

Dash's method achieves a significant reduction of the error when this procedure is compared with the classic *k*-means algorithm. In this new procedure, two phases are highlighted. Firstly, the data are normalized and, secondly, the principal components are calculated and then the *k*-means algorithm is applied in a new subspace X*. This subspace is consisting of those principal components whose variance is greater than the mean of variances of principal components. The direct comparison between the SSE obtained by Dash's method and by the *k*-means algorithm is not possible. This claim is deduced from the following considerations: the SSE value obtained by the classic *k*-means algorithm is different if data are normalized or not and the SSE value grows as the number of principal components selected grows. To illustrate these considerations, two examples are shown.

- Firstly, given a set of m objects and n numeric variables, the objective function value is different if the *k*-means algorithm is applied to normalized data or not. This can be seen in the following case, consider the synthetic dataset and the normalized synthetic dataset and run 1000 times the *k*-means algorithm. The average and the standard deviation of SSE have been calculated for both cases as the following table shows.

Table 4. SSE average and standard deviation values calculated by using the *k*-means algorithm 1000 times in synthetic dataset, normalized and non-normalized

Dataset	SSE Average Original *k*-means algorithm	SSE standard deviation Original *k*-means algorithm
original synthetic dataset	560.5608	76.99678
normalized synthetic dataset	76.98022	14.32716

- Secondly, given a set of m objects and n numeric variables and fixed a cluster classification, the SSE value calculated in the space consisting of the p first principal components is less or equal than the SSE value calculated in the subspace consisting of the $p+q$ first components ($p+q \leq n$). So the SSE value grows as the number of principal components selected grows. As an

example of this, the *k*-means algorithm has been applied on the normalized synthetic dataset. The labels provided by the *k*-means algorithm in the space of all the principal components have been considered. Each centroid has been computed as average of all elements within each cluster. Then, the sum of squared distances between each element and its centroid has been calculated. The different SSE values obtained are shown in Table 5.

Table 5. SSE values calculated in vector spaces of 3 and 10 dimensions

Dataset	**SSE**	**Subspace**
Normalized Synthetic Dataset	47.80006	$\Re^3$
	65.81692	$\Re^5$
	70.50695	$\Re^7$
	71.11372	$\Re^{10}$

Although the comparison between the SSE value in the original space and the SSE value proposed by Dash is possible taking into account the next.

Remark:

Given a cluster classification, the SSE values calculated in the original space and in the space consisting of all the principal components coincide. This happens because the principal components analysis provides an orthonormal transformation and SSE is a sum of squared distances that does not depend on the orthonormal basis selected.

Moreover, in order to achieve the best classification possible a high quality of data is needed. Frequently, the variables of study or sampling methods are not subject to choose. Instead the process begins with a dataset that should be analysed. The variables which do not preserve relevant information must be deleted, this type of variable may be a label or code variable that identifies each case of dataset. Preserving these variables may increase the computational cost and the results of cluster analysis can be altered. About the rows of dataset that have some missing values, there are different options: missing values can be changed by the average of the variable corresponding or the most frequent value. As these options introduce artificial information, removing the complete row is frequently preferred, but it depends on the number of cases in the dataset. For this reason, the duplicate rows or rows with missing values will be eliminated before applying the cluster procedure, as well as the variables that do not provide relevant information. The HKMCA and the original *k*-means algorithm are applied to synthetic dataset below by taking into account the above considerations. Observe that the number of the clusters to make is 2 and it is the same as in Dash *et al.* (2010). In this case, the variable V1 is a variable which identifies each case of the dataset and the variable V10 is almost constant except for just one case, therefore, both variables have been removed in this case (Table 1). Table 6 shows the SSE values obtained.

Table 6. SSE values when the HKMCA and the original *k*-means algorithm is applied to the modified synthetic dataset

Dataset	No of instances	No of variables	Original *k*-means algorithm	HKMCA	
			SSE	SSE	NPC
Modified Synthetic	14	8	45.09221	50.38126	8

The SSE value, obtained by using the original *k*-means algorithm, is averaged out from the results of 1000 runs. The SSE value, when the HKMCA is applied, coincides with the SSE value obtained with the original *k*-means algorithm 13.4% of times. This is shown in Table 7.

Table 7. The SSE values obtained by using the *k*-means algorithm with the normalized synthetic dataset and its classification percentages

Sum of Squared Error	43.7367	50.3813	49.0435	49.8854
Percentage for each cluster classification	78.2%	13.4%	6.1%	2.3%

The results obtained by using the Pima Indian Diabetes, Breast Cancer, SPECTF Heart datasets, are shown in Table 8. The number of groups to create in each case has been provided by the variable that indicates the class of the cases.

Table 8. SSE values obtained by applying the HKMCA and the original *k*-means algorithm to Pima Indian Diabetes, Breast Cancer and SPECTF heart datasets

Dataset	No of instances	No of variables	Original *k*-means Algorithm SSE	HKMCA SSE	HKMCA NCP
Pima Indian Diabetes	768	8	5154.336	5143.613	8
Breast Cancer	585	9	2366.440	2366.723	9
SPECTF Heart	80 (training)	44	2841.563	2959.472	44
	187 (test)	44	6213.734	6210.210	44

The SSE value obtained by using the HKMCA with Pima Indian Diabetes improves the results obtained with the original *k*-means algorithm 39.55 % of times. The original Breast Cancer dataset had 699 cases, but the duplicate rows or rows with missing values have been deleted as well as the first variable that identifies each case, so the final dataset has 585 cases. In this case, the original *k*-means algorithm provides just one classification after running 1000 times this algorithm. The SPECTF dataset is composed of two sets, training and test data. We have repeated the analysis in both cases, deleting previously the class variable. The SSE values obtained when the original *k*-means algorithm is applied to training dataset are smaller than the value obtained by using the HKMCA 98.75 % of times.

5.1 A high dimensional example

Nowadays, datasets encountered in our society are composed of thousands of variables and, in these cases, the component dimension reduction proposed by Dash could be needed. In this sense, data sets used by the author in the previous paper have not very high dimensionality. In order to analyse the utility of the procedure proposed by Dash when it is applied to high dimensional data sets, a new example is included in this paper. This dataset, called *mfeat*, consists of features of handwritten numerals (0-9) extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images. These digits are represented in terms of 649 features. The procedure proposed by Dash, applied to this new data set always produces the same result, since it contains the initial seeds. According Dash's procedure, in a subspace of 77 principal components, it is obtained a value of SSE = 635619.5. However, as already mentioned, this value is not directly comparable to the value provided by the original algorithm *k*-means. The value considered in the original space of the normalized data, that it is comparable with the result of the *k*-means algorithm, is SSE = 790886.5. In order to contrast the real differences between both methods each algorithm has been run 10 times. The average, the minimum and the maximum values and the standard deviation of the SSE values have been calculated as the following table shows.

Table 9. SSE average, minimum, maximum and standard deviation values calculated using the *k*-means algorithm and the HKMCA 10 times in *mfeat* dataset

Sum of Squared Error (SSE)	Mean	min	max	sd
k-means	783370	782566.9	786854.4	1315.108
HKMCA	790886.5	790886.5	790886.5	0

Moreover, as an example, the time employed in both cases has been calculated to have a reference about the time complexity. The time employed for the initialization process is also included.

Table 10. Time complexity using the *k*-means algorithm in *mfeat* dataset

Time (sec.)	mean	sd
k-means	49.072	4.255408
HKMCA	65.456	1.178492

The results show that the HKMCA does not improve the standard method, furthermore it requires more computational time.

6. Conclusions and final remarks

This paper makes different remarks about the method of comparison followed in Dash *et al.* (2010) and the experimental results obtained by using our guidelines. The Pima Indian Diabetes, Breast Cancer, SPECTF Heart and the synthetic datasets have been employed to compare both procedures. According to results obtained and subsequent computational analysis we can conclude that on average the results obtained by using the HKMCA coincide with the results obtained with the original *k*-means algorithm. However, the HKMCA is a deterministic procedure and it always provides the same result in each dataset. The HKMCA does not

improve the *k*-means results as much as Dash *et al.* state, since the results obtained in Dash *et al.* (2010) using the HKMCA and the *k*-means algorithm are not comparable. Finally, in order to analyse the utility of the procedure proposed by Dash when it is applied to high dimensional data sets, a new example is included. This example shows that the HKMCA does not provide improvements.

Nomenclature

HKMCA	Hybridized *K*-Means Clustering Algorithm
PCA	Principal Components Analysis
NPC	Number of Principal Components
SPECTF	Single Proton Emission Computed Tomography Features
SSE	Sum Square of Errors

References

Bache, K., Lichman, M. 2013. *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Dash R., Mishra D., Rath A.K., Acharya M., 2010. A hybridized *K*-means clustering approach for high dimensional dataset. *International Journal of Engineering, Science and Technology,* Vol. 2, No. 2, pp. 59-66.

Forgy E.W., 1965. Cluster Analysis of Multivariate data: efficiency vs interpretability of classifications. *Biometrics*, Vol. 21, pp. 768-769.

Hartigan J.A., Wong M.A., 1979. Algorithm AS 136: A *k*-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics),* Vol. 28, No. 1, pp 100-108.

Jolliffe I.T., 2002. *Principal Components Analysis*. 2nd ed., New York: Springer-Verlag.

Lloyd S.P., 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, Vol. 28, pp. 129-137.

MacQueen J.B., 1967. Some methods of classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297.

Peña D., 2002. *Análisis de Datos Multivariantes.* McGraw-Hill Interamericana de España, S.A.U.

Biographical notes

A. J. Arriaza-Gómez. He had a degree in mathematics in 2005 and a Master's degree in mathematics in 2008, both at the University of Cádiz. He is currently an Assistant Professor in the Department of Statistics and Operations Research at the University of Cádiz, Spain and he is a PhD student in the same Department. His current area of research includes Data Mining, and Optimization.

F. Fernández-Palacín. Degree in mathematics from the University of Seville in 1981. He begins to work in the Department of Economie at the Cádiz University. He earned a PhD in 1994 under the supervision of Antonino García-Rendón and José Muñoz-Pérez in the field of location theory. From October 1995 he held the position as Head of College in the Department of Mathematics, University of Cádiz (now Department of Statistics and Operations Research). His current area research includes Location Theory, Multivariate Analysis, Statistical Processing and Data Mining. He belongs to the national network of researchers in Location Theory and has participated in several research projects. He currently holds the position as vice-rector for prospective, quality and communication at the University of Cádiz.

M. Muñoz-Márquez. Degree in mathematics from the University of Seville in 1991. He begins to work in the Department of Statistics and Operations Research at the same university. He earned a Ph.D. in 1995 under the supervision of Emilio Carrizosa and Justo Puerto Albandoz Priego in the field of location theory. From October 1995 he held the position of Associate Professor type IV in the Department of Mathematics, University of Cádiz (now Department of Statistics and Operations Research) until 1997 when he obtained a position as Head of College. His current area research includes Location Theory, Optimization, Surveys, Statistical Data Analysis and Data Mining. He belongs to the national network of researchers in Location Theory and he has participated in several research projects and organizing scientific events of multiple national and international levels. He currently holds the position responsible for educational technology at the University of Cádiz.

S.M. Pérez-Plaza. She had a degree in mathematics in 2006 and a Master's degree in mathematics in 2008, both at the University of Cádiz. She is currently an Assistant Professor in the Department of Statistics and Operations Research at the University of Cádiz, Spain and she is a PhD student in the same Department. Her current area of research includes Data Mining, and Optimization.

Received February 2012
Accepted September 2012
Final acceptance in revised form September 2013