Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3, November 2015
ISSN 2074-8523

PCAWK : A Hybridized Clustering Algorithm Based on PCA and WK-means for Large Size of Dataset

Fatemeh Boobord[1], Zalinda Othman[1], and Azuraliza Abubakar[1]

[1]Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
University Kebangsaan Malaysia, Selangor Darul Ehsan, Malaysia
e-mail: fboobord@gmail.com, zalinda@ukm.edu.my
azuraliza@ukm.edu.my

Abstract

Real world data usually has a variation of size of dimensionality. The dimensionality needs to be reduced for handling the dimensionality of data. The dimensionality reduction changes the presentation of dimensional data variation to a meaningful presentation. In this paper, a method based on the principle component analysis and WK-means called "PCAWK" is proposed. Firstly, PCA is used to reduce the redundant dimensionality of dataset and then, the WK-means algorithm that is a hybrid of Invasive Weed Optimization (IWO) and the K-means algorithm utilizes the reduced dataset to obtain the optimal clusters. The proposed algorithm is tested on 5 real word instances and the results are compared with the PCAK algorithm. The proposed algorithm generally has better performance in most datasets.

Keywords*: Clustering, K-means, Large Size Data, Metaheuristics, Principle Component Analysis.*

1 Introduction

Nowadays, people want to store or represent large amount of information as data. One of the important means in dealing with these data is to mine or extract knowledge from large amounts of data that stated data mining. Mining data refers to the process that

discovers a small set of valuable knowledge from a great deal of raw data. Data mining is simply a vital step of the process of knowledge discovery.

One of the most important tasks for organizing raw data is classify or cluster them. Clustering performs major and indispensable role in data mining. The procedure of Clustering is assigning a set of similar physical or abstract objects into a cluster. Objects belong to a cluster have the most similarities to each other, whereas objects in different clusters have dissimilarities to one another. In other words, Data analysis is an unsupervised learning in which the class label of each object is not known. Clustering recognizes the natural hidden groups in a set of objects. Against classification that is supervised learning, clustering is an affordable scheme to analyse large databases because finding the class label for a large number of data is more costly.

Typically, the clustering method is classified into two categories: partitioning and hierarchical. K-means is the most famous partitioning clustering algorithm. For a given dataset with n objects, the K-means method groups the dataset to k clusters ($k \leq n$) in which each partition represents a cluster. Classifying the data to k groups should satisfy two requirements: (1) each cluster must have at least one object or tuple, (2) each object must be a member of just one group. The K-means method constructs k groups from the initial given dataset. It then applies the mean of data to relocate the objects among the groups iteratively for improving the clusters. The process of relocation continues until the criterion appears. A good criterion is that the objects in the same group have the most similarities to each other and the objects in different groups are very different. It is a popular and simple algorithm, with linear time complexity. It has been used over 50 years, and it is still widely used.

However, the number of distance calculation increases exponentially when the dimensionality of data grows. When the dimensionality increases, only a limited number of features are relevant to certain cluster and the irrelevant features may lead to wrong clustering. Thus the data reduction method is an essential pre-processing method for data clustering in large number of features.

Dimensionality reduction transforms high-dimensional data into a meaningful illustration of reduced dimensionality that reflects the intrinsic dimensionality of data. The dimensionality reduction technique is divided into two groups: feature selection (FS) and feature reduction (FR). The feature selection algorithm focuses on finding a subset of the most reprehensive features based on some objective functions in discrete data. As the FS algorithm is always greedy, sometimes it cannot find the optimal solution. Meanwhile, the feature reduction algorithm reduces the features by projecting the original high-dimensional data into a lower-dimensional space by algebraic transformations. The FR algorithms are used to find the optimal solution in continuous space, but the computational complexity is more comparative to the FS algorithms.

Different types of FR methods have also been proposed. PCA is a commonly used feature reduction method in terms of minimizing the reconstruction error.

K-means is the most famous partitional algorithm for clustering low dimensional data. Often, it does not work well in high dimensional data. The computational complexity of the K-means algorithm also increases in high dimensionality. Therefore, to improve the efficiency of the K-means, several methods have been proposed. Tajunisha et al. proposed for a data reduction method based on PCA in which they used the heuristic method to reduce the complexity of the k-means algorithm [1]. George proposed for a method that performs dimensionality reduction using PCA as the pre-processing step to data clustering. Then, they integrate the COP-KMEANS with the reduced dataset to produce good and accurate clusters [2]. Dash et al. proposed for an approach to use the Principal Component Analysis (PCA) method as the first phase for K-means clustering. Then, they proposed for a new method to find the initial centroids to make the algorithm more effective and efficient [3]. Napoleon et al. used PCA on the original data set and obtained a reduced dataset containing possibly uncorrelated variables. They proposed for the principal component analysis and linear transformation to be utilized for dimensionality reduction and initial centroid, where it is applied to the K-Means clustering algorithm [4]. A method proposed by Behara et al., uses the Canonical variate analysis to reduce the dimensionality of dataset. Then, a clustering technique is applied using a modified k-means clustering. For initializing the initial centroids, they made use of the genetic algorithm [5]. Paubhu et al. proposed for a method to use the Principal component Analysis to reduce the dataset from high dimensional to low dimensional. The new method is used to find the initial centroids based on the variance of data in each dimension [6]. Behara et al. presented a PSO optimized k-means algorithm with improved PCA for clustering high dimensional dataset [7].

By increasing the dimensionally, the K-means algorithm loses its efficiency. To solve this problem, we proposed for a method based on the PCA data reduction and WK-means [8] called (PCAWK). In this work, after dimensionality reduction, the complexity of K-means algorithm is reduced by IWO which is an evolutionary optimization algorithm [9]. Then, the proposed algorithm shows better performance compared to the PCAK algorithm. This paper is organized as follows: Section 2 presents a review of the clustering problem, principle component analysis, invasive weed optimization and new proposed algorithm. Section 3 provides the results of the proposed algorithm on several datasets.

2. Methodologies

2.1 Clustering Problem

K-means is a simple, fast and very popular clustering method. The procedure for this algorithm first start with placing K objects randomly as initial cluster centres (K is a fixed number as a parameter). Next, the objects are assigned to their closest cluster centre. Then, the algorithm calculates the average of each cluster as a new cluster centre. The last two stages continue until a termination condition is reached. These steps are shown in Figure 1. The goal of the K-means algorithm is to minimize the sum of the distance between cluster centres and objects over all K clusters as follows: [10]

$$pref(X,C) = \sum_{i=1}^{N} \min\{\| X_i - C_i \|^2 l = 1,...,K\} \quad (1)$$

Where, *pre(X,C)* is a performance function (fitness function) of the K-means method that defines both data items and centre locations. $i\ X$,$i=1,\ldots,N$ is a data object and $l\ C$,$l=1,\ldots,K$ *is* a cluster centre [10].

2.2 Principle Component Analysis

Let a dataset consists of data vectors that are defined by n features or dimension. The Principle component analysis (PCA) or Karhunen-loeve method looks for k, n-dimensional orthogonal vectors which can be used for presenting data, whereas $k \leq n$. The original data is then projected onto a smaller space, resulting in dimensionality reduction. The PCA linearly combines the essence of features which maximize the variance of the linear combination that is uncorrelated with previous PCs. Each component that is mined will account for a maximal amount of variance in the dataset which does not account for the previous components, and it will not be correlated with the last components. After mean centring, the data for each attribute, PCs are calculated according to the Eigen value decomposition of a data covariance matrix.

2.3 Invasive Weed Optimization Algorithm

The invasive weed optimization was introduced by Mehrabian et al. in 2006 [9]. The IWO is inspired by the colonization of invasive weeds. It is an evolutionary optimization algorithm [9]. The simulation process of IWO starts by distributing a finite number of seeds in the search area (each seed is regarded as a solution). Every seed grows to a plant and produces new seed according to its fitness. The produced seed is linearly decreased from the fittest plant to unfit (inappropriate) one and dispread over the search area by the normally distributed random numbers, with a mean value of zero and varying variance. The variance is calculated according to equation (2):

$$\delta_{\text{iter}} = \frac{(iter_{\max} - iter)^n}{(iter_{\max})^n}(\delta_{initial} - \delta_{final}) + \delta_{final} \tag{2}$$

Whereas δ_{iter} is the standard deviation of the current iteration, $iter_{max}$ is the maximum number of iteration and n is the non-linear modulation index. The process of generating seed is continued until the maximum population is reached. Now, only the fittest plant can go on to produce seed, and others will be removed. The algorithm ended when the termination criterion (a maximum number of iteration or a sufficiently good fitness) is met.

2.4 Proposed Method

To use the merits of PCA and Wk-means algorithms, a hybrid approach for data clustering is presented in this chapter, which is called "PCAWK". The process of the algorithm consists of three stages where at the first stage, we normalize data to scale and fall data within a small-specified range. The process of normalization is as follows:

Let A is a set of k cluster. An attribute value S of an attribute S is normalized to S' using the Z-score. The Z-score formula is shown in Equation 3.

$$S'=(S-\text{mean}(A))/\text{std}(A) \tag{3}$$

Whereas STD is the standard deviation of set A and mean is the average of A. The normalization of the method has two advantages; first, data centring to decrease the square mean error of approximated input data and secondly, it scales data by standardizing the variables to have unit variance before the analysis takes place. The normalization also prevents certain features from dominating the analysis because of their large numerical values.

In the second step, the PCs will be calculated by the singular value decomposition (SVD) of the normalized data. The number of PCs is the same as the original variables. To select the best PCs, we calculate the corresponding variance, percentage of variances and cumulative variances in percentage. Then, we select PCs that have variances less than the means variance and ignore the others. The transformation matrix is formed from selected PCs, and the obtained matrix is used for normalized data to reduce the dataset that can be used for further analysis.

In the third step, the reduced dataset will be used as an initial solution for the Wk-means algorithm. Let A is a set of n data point $A=\{a_1,\ldots,a_n\}$, Then the process of PCAWK algorithm is as shown in Figure 1.

3 Experimental Results

In this section, the performance of the proposed algorithm is shown based on the sum of square error value in Table1, and the best, average and worst solutions and standard deviations of PCAWK are compared to PCAK [1] for 20 runs. The proposed algorithm is tested on 5 different datasets taken from the UCI machine learning repository. The performance of proposed algorithm is also compared to PCAK in charts Fig 2 – 6 based on the sum of square error.

As seen from the results in Table 1, for Wine dataset, the PCAWK provides the average value of 266.9676 for the objective function, while the PCAK algorithm obtained 267.9972. In the case of the Best, Worst and standard deviation, the PCAWK also has better performance. This shows/reflects that PCAWK can find higher quality clusters compared to other algorithms. On the Cancer dataset, the Best, Average and worst solutions for PCAWK are 2065.30, 2068.66 and 2076.3 respectively, which show better performance compared to other methods. The

Step 1 Normalize data points using z-score

1.1. Finding the mean and standard deviation of data points, according to the following formula respectively:

$$\mu_A = \Sigma a / n$$

$$\sigma_A = \sqrt{\frac{\sum_{i=1}^{n}((a_i - \mu_A)^2)}{n-1}}$$

1.2. Finding elements of the data point matrix by the formula:

$$V' = (a_i - \mu_A)/\sigma_A$$

Step 2. Apply PCA to decrease the dimension of data set

2.1. Using singular value decomposition for data matrix. A=UDV'.
2.2. Calculate the variance of diagonal matrix D.
2.3. Sort variance in decreasing order.
2.4. Select p principle components from V with largest variances.
2.5. Form transformation matrix w, using p selected PCs.
2.6. Find reduced matrix of data set by using transformation matrix w.

Step 3. Using Wk-means to cluster the reduced data set.

3.1 . Select an initial population from reduced dataset.
3.2 . Calculate the fitness of population.
3.3 . Each individual produces new seed according to it fitness.

3.4 . If the maximum population is reached, only the fittest individual can produce seed.
3.5 . Select the optimal solution as the initial solution of K-means.
3.6 . Assign the object to the nearest cluster centre.
3.7 . Calculate the average of new cluster centre.
3.8 . Repeat stages 3.6 and 3.7 until the average of cluster isfixed.

Fig. 1: The stage of PCAWK algorithm

USCencus90 has a better performance in the average, worst and standard deviation for the PCAWK algorithm with 9788.7, 9868.4 and 43.9379, while the PCAK has obtained 10329.39, 12113 and 737.6753 respectively. In the musk2000 dataset, we can also see a better performance for PCAWK in the average, worst and standard deviation that are 4690.03, 4730.9 and 23.8549 respectively. The PCAK at the same time obtained 4706.92, 5267.3 and 295.3462 values respectively. In the SPECTF Heart dataset, the PCAWK is better than PCAK with 1186 for the best solution and 1189.533 for the average value.

Table 1. Results obtained by proposed algorithm on the tested dataset

Dataset	Criteria	PCAK	PCAWK
Wine	Best	267.7058	266.000
	Average	267.9972	266.9676
	Worst	268.525	267.2110
	Std	0.308379	0.2421
Cancer	Best	2076.60	2064.7
	Average	2077.47	2068.66
	Worst	2078	2076.3
	Std	0.566708	3.8733
USCencus90	Best	9632.3	9740
	Average	10329.39	9788.7
	Worst	12113	9868.4
	Std	737.6753	43.9379
musk2000	Best	4566.80	4656.3
	Average	4706.92	4690.03
	Worst	5267.3	4730.9
	Std	295.3462	23.8549

SPECTF Heart	Best	1190.1	1186
	Average	1190.28	1189.533
	Worst	1191	1208.4
	Std	0.36935	6.1777

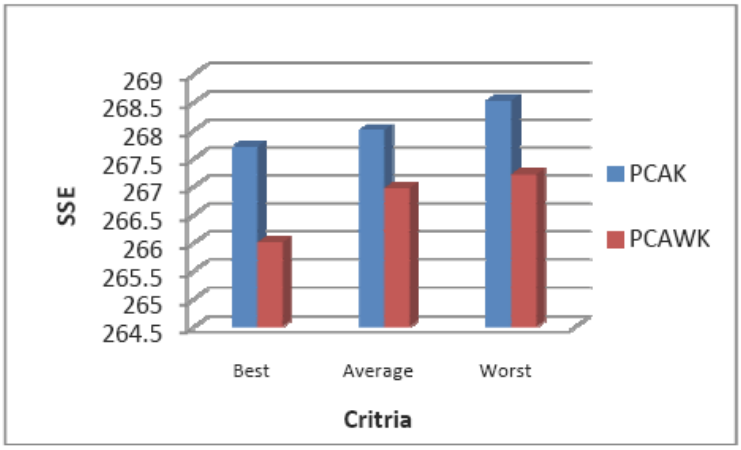

Fig. 2: SSE results on Wine data set

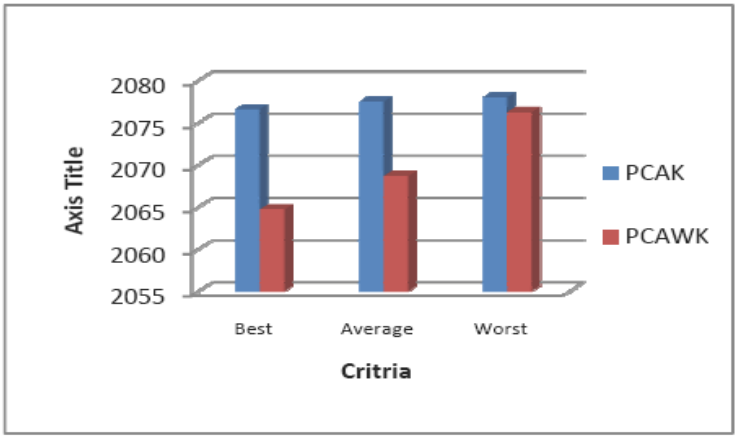

Fig. 3: SSE results on Cancer data set

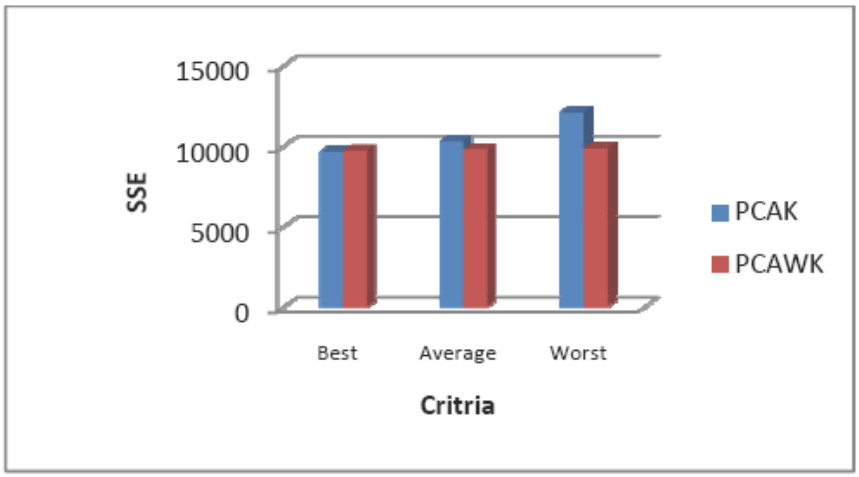

Fig. 4: SSE results on USCencus90 data set

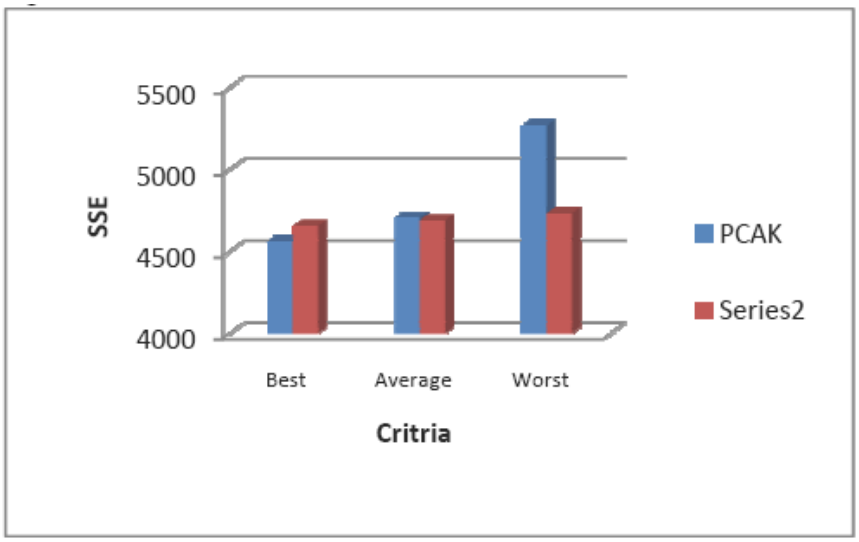

Fig. 5: SSE results on musk2000 data set

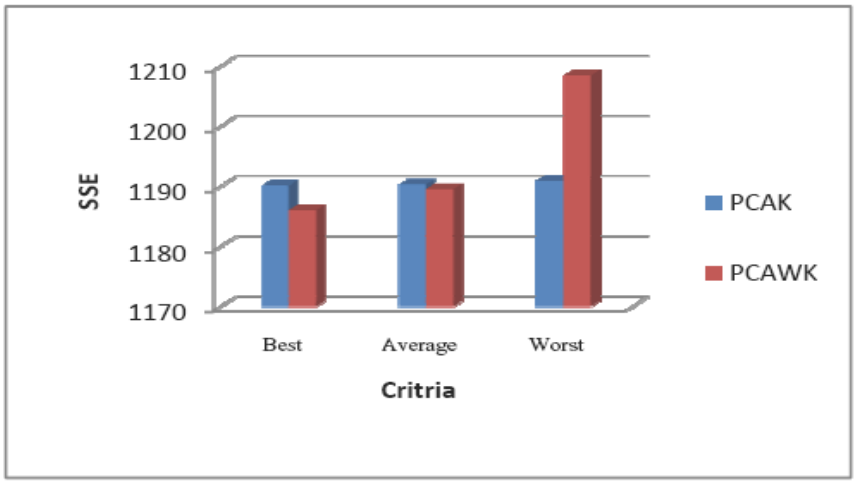

Fig. 6: SSE results on SPECTF Heart data set

4 Conclusion

This paper proposed for a new clustering method based on the PCA and Wk-means algorithms. PCA is a feature selection method that reduces the dimensionality by selecting and decreasing the redundant features and WK-means is a method for clustering the objects. To increase the capability of K-means algorithm in high dimensionality, the proposed method firstly used PCA to reduce the dataset and then, the WK-means is used to produce optimal clusters. The proposed algorithm is tested on 5 real- instance datasets and compared to the PCAK algorithm. The experimental results show that the PCAWK method outperformed in most of the cases in comparison to the PCAK algorithm.

ACKNOWLEDGEMENTS
This research was supported by University Kebangsaan Malaysia, under grant number FRGS/1/2013/ICT02/UKM/01/1.

References

[1]Tajunisha and Saravanan, "An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means," International Journal of Database Management Systems, vol. 3, 2011.

[2]A. George, "Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm," International Arab Journal of Information Technology, vol. 10, 2013.

[3] R. Dash, et al., "A hybridized K-means clustering approach for high dimensional dataset," International Journal of Engineering, Science and Technology, vol. 2, pp. 59-66, 2010.

[4] D. Napoleon and S. Pavalakodi, "A New Method for Dimensionality Reduction using KMeans Clustering Algorithm for High Dimensional Data Set," International Journal of Computer Applications, vol. 13, pp. 0975-8887, 2011.

[5]H. S. Behera, et al., AN IMPROVED HYBRIDIZED KMEANS CLUSTERINGALGORITHM (IHKMCA) FOR HIGHDIMENSIONAL DATASET & IT'S PERFORMANCE ANALYSIS, vol. 3, pp. 0975-3397, 2011.

[6] P. Prabhu and N. Anbazhagan, "Improving the Performance of K-Means Clustering For High Dimensional Data Set," International Journal on Computer Science and Engineering, vol. 3, pp. 0975-3397, 2011.

[7] H. S. Behera, et al., "A New Improved Hybridized K-MEANS Clustering Algorithm with Improved PCA Optimized with PSO for High Dimensional Data Set," International Journal of Soft Computing and Engineering, vol. 2, pp. 2231-2307, 2012.

[8] F. Boobord, et al., "A WK-means approach for clustering problem," International Arab Journal of Information Technology, 2013.

[9] A. R. Mehrabian and C. Lucasc, "A novel numerical optimization algorithm inspired from weed colonization," Ecological Informatics vol. 1, pp. 355-366, 2006.

[10]A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, pp. 651–666, 2010.