Clustering of datasets using PSO-K-Means and PCA-K-means

Anusuya Venkatesan
Manonmaniam Sundaranar University
Tirunelveli- 605014, India
anusuya_s@yahoo.com

Latha Parthiban
Computer Science Engineering Department,
S.S.N College of Engineering,
Chennai – 600 004. India
lathparthiban@yahoo.com

Abstract

Cluster analysis plays indispensable role in obtaining knowledge from data, being the first step in data mining and knowledge discovery. The purpose of data clustering is to reveal the data patterns and gain some initial insights regarding data distribution. K-means is one of the widely used partitional clustering algorithms and it is more sensitive to outliers and do not work well with high dimensional data. In this paper, K-means has been integrated with other approaches to overcome the shortcomings hereby improving the accuracy of clustering. In this paper, basic k-means and the combination of k-means with PCA and PSO are applied on various datasets from UCI repository. The experimental results of this paper show that PSO-K-means and PCA-K-Means improves the performance of basic K-means in terms of accuracy and computational time.

Keywords: Clustering, PSO, PSO-K-means, QPSO

1. Introduction

Data clustering is a technique in which data with similar characteristics are grouped together to form clusters. Clustering has been studied by many researchers for a long time and been applied in areas such as pattern recognition, gene expression analysis, customer segmentation, educational research and etc. Clustering techniques are generally categorized into hierarchical, partitional, density based and grid based clustering [10].

Hierarchical Clustering: These methods start with each point being considered a cluster and recursively combine pairs of clusters (subsequently updating the inter-cluster distances) until all points are part of one hierarchically constructed cluster.

Divisive or Partitional Clustering: Partitional clustering, on the other hand, performs a partition of patterns into *K* number of clusters, such that patterns in a cluster are more similar to each other than to patterns in different clusters. Recent studies have shown that partitional clustering algorithms are more suitable for clustering large datasets.

Density Based Clustering: in which neighbours are grouped based on their density in the area.

Grid Based clustering: which quantizes the space into finite number of areas and then they perform their operations in each area separately.

K-means is the most commonly used partitional clustering algorithm but the major problem of K-means is the selection of the initial partition and its convergence to local optima.The basic K-means algorithm is simple and fast. The time complexity of K-means is O(l* k*n), where l is the number of iterations , k is the number of clusters and n is the number of data items. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of k and compare the clustering results.

Algorithm 1: The k-means clustering algorithm
The algorithm operates on a set of d-dimensional vectors, D = $\{\mathbf{x_i} \mid i = 1, . . . , N\}$, Where $\mathbf{X_i} \in R^d$ denotes the ith data point.

The algorithm is initialized by picking k points in R^d as the initial k cluster representatives or "centroids".

1. Choose K centroids at random or using hierarchical clustering.
2. Make initial partition of objects into K clusters by assigning objects to closest centroid.
3. Calculate the centroid(mean)of each of the K clusters.
   a) For object i, Calculate its distance to each of the centroids.
   b) Allocate object i to cluster with closest centroid.
   c) If object was reallocated, recalculate centroids based on new clusters.
4. Repeat step3 for object i= 1 to N
5. Repeat 3 & 4 until no reallocations occur.
6. Assess cluster structure for fit and stability.

The key idea of using Particle Swarm Optimization (PSO) is to create a population of candidate solutions to an optimization problem, which is iteratively refined by alteration and selection of good solutions for the next iteration. Candidate solutions are selected according to a fitness function, which evaluates their quality with respect to the optimization problem.

The objects or particles are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

The paper is organized as follows. The section 2 of this paper briefs the procedure and functionality of PSO while section 3 deals with PSO-K-means, a combination of PSO and K-means to perform clustering. The combination of principal component analysis and kmeans is described in Section 4. The experimental clustering results of various datasets implemented in MATLAB are produced in section 5.

2. Particle Swarm Optimization

PSO is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995[1][5][7], inspired by social behaviour of bird flocking or fish schooling and has been rapidly applied to data mining tasks such as classification and clustering to optimize the results. Clustering using PSO has been applied in wireless sensor networks [2], tested against random search and simulated annealing, and found to be more robust. PSO has also been applied in document clustering [3] which demonstrated that the hybrid PSO algorithm generated more compact clusters in comparison to the K-means algorithm. Combining K-means and PSO[4] for data clustering achieved not only fast convergence to optimum solution but also higher accuracy. PSO-Kmeans is successfully demonstrated on Bus standards for static and transient security evaluation[12].The variants of PSO and its applications are proposed in [8] and [9]. A disadvantage of the global PSO is it tends to be trapped in a local optimum under some initialization conditions[6].

In PSO, the potential solutions, called particles, searches the whole space guided by its previous best position(pbest) and best position of the swarm(gbest). The velocity and position of the particles are updated based on its best experience. The i^{th} particle of swarm is represented as $X_i = (X_{i1}, X_{i2}, \ldots, X_{iD})$ while the velocity for i^{th} particle is represented as $V_i = (V_{i1}, V_{i2}, \ldots, V_{iD})$. The best previous position (the position giving the best fitness value) of the i^{th} particle is recorded and represented as $pi = (p_{i1}, p_{i2}, \ldots , p_{iD})$. At each step, the particles are manipulated according to the following equations (1) and (2):

$$V'_{id}=\omega V_{id}+c_1 rand()(P_{id}-X_{id})+c_2 rand()(P_{gd}-X_{id}) \qquad (1)$$

ω represents the inertia weight to control the speed of each generation of particles and $X'_{id} = X_{id} + V'_{id}$ (2)

$c1$,$c2$ are two positive constants known as cognitive and social components. The velocity is calculated based on previous velocity, cognitive component and social components. The pseudo code of PSO is presented below;

```
Generate random population of N
 solutions(particles);
For(i=0;i< Swarm Size;i++)
Evaluate fitness f(x_i);
Initialize the value of weight factor, ω;
while (termination condition is not true)
{
for(i=0;i< Swarm Size;i++)
{
if(f(x_i)>pbest_i) pbest_i=x_i;
 if(f(x_i)>gbest_i) gbest_i=x_i;
 Update(Position x_i, Velocity v_i);
 Evaluate f(x_i);
}
}
```

3. PSO-K-means

3.1 Procedure of PSO-K-means

1. In the context of PSO-K-means clustering, before initializing the particles, the data points are randomly assigned to K clusters first.
2. Particle fitness is evaluated based on clustering criteria.

$$F(j) = 1/N\sum_{1}^{K}\sum_{x\in K_i}\left\|x_j - C_i\right\|^2 \qquad (3)$$

The fitness function of the particle j, between a data point $x_{j=1,N}$ and the cluster center C_i , to minimize the sum of squared distances from all points to their cluster centres, would result in compact clusters. N represents the total number of data points in the clustering process.
3. The velocity and position of the particle are modified using eqn (1) and (2).
4. The new generation is optimized by K-means as given below.
 a) The data set is reassigned to clusters according to nearest rule.
 b) Cluster centroids, fitness value are recalculated and positions are updated.
5. If the position is satisfactory or the maximum iteration is reached, the process is stopped. Otherwise, return to step 2.

Cluster Validity

In order to improve the PSO-K-means as an automatic one, the validity measure is calculated as the ratio between intra cluster distance and inter cluster distance as in eqn (4). The clustering which gives a minimum value for the validity measure gives the ideal value of *K*.

$$\text{Validity}_{\text{Cluster}} = \text{Intra} / \text{Inter} \qquad (4)$$

4. PCA-K-Means

Principal component analysis is invented by Karl Pearson[11] and the general idea of using PCA is to reduce the dimensionality of data prior to some other statistical processing. Dimension reduction is closely related to unsupervised clustering. PCA based data reduction outperforms traditional noise reduction techniques and has 5. been applied to clustering gene expression profiles[13] . PCA 6. has also been demonstrated in the applications of face recognition, image compression, and to find patterns in high dimensional data.

The main objective of using PCA is dimensionality reduction of data and to find new underlying variables. An approach to find or extract key components from original data influences the division of clusters. Principal component analysis is one such approach constructs a linear combination of a set of vectors that can best describe the variance of data.

It is more common that PCA can be used to project the data into lower dimension subspace by picking up the data with the largest variances. After finding reduced datasets Kmeans is applied to perform clustering.

4.1 Procedure of PCA-Kmeans

1. X represents the original data matrix, Y represents the centered data matrix Y=(y1,y2…yn) where $y_i = x_i - \bar{x}$ represents centered data matrix where $\bar{x} = \sum_i x_i / n$.
2. Find covariance matrix YYT.
3. Choose first p principal components having largest variances.
4. Form transformation matrix W consisting of p principal components.
5. Find the reduced projected dataset D .
6. Partition the D data points into K number of clusters using K-means ;where K is the desired number of clusters. .

5. Experimental Results

In order to evaluate the performance PCA-Kmeans and PSO-K-means, experiments were conducted on various datasets from UCI repository. Fig. 1 shows Kmeans clustering of Iris dataset and the histogram of breastcancer after clustering by PSO-Kmeans is shown in Fig. 2.

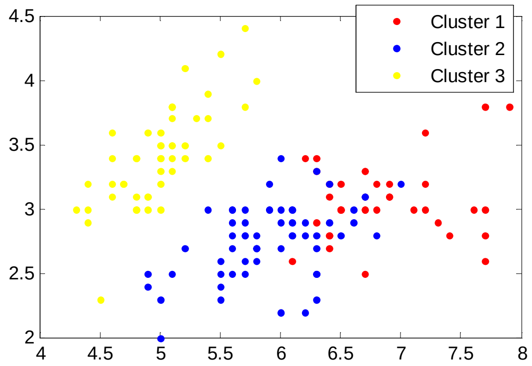
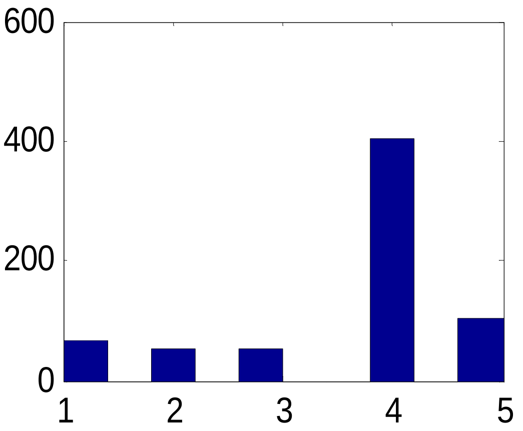


Fig. 1 K-means clustering on Iris dataset



7.

Fig. 2. Histogram of PSO-K-means clusters of Breast Cancer 8.for K=5

9. The overlapping data are very less when PSO-K-means is being applied on Iris compared to basic kmeans . Fig. 3 shows the clustering results of Iris dataset using PSO-K-means for the parameters N=50; c1=1.2; c2=1.2; wmax=0.9; wmin=0.4; M=200; K=3, Fig. 4 shows the bestfit of objective functions for the population of size 100 and first three principal components are considered for PCA-Kmeans and the clustering results of PCA-Kmeans for K=3 shown in Fig. 5and Fig. 6. The distance between cluster centers(intra) and distance between data points within a cluster(Inter) calculated by Kmeans, PCA-Kmeans and PSO-Kmeans are mentioned in Table1. The number of clusters for all datasets is set as K=3 and D represents the dimension. And also the computation time T taken by all three methods on different datasets is given in Table1.

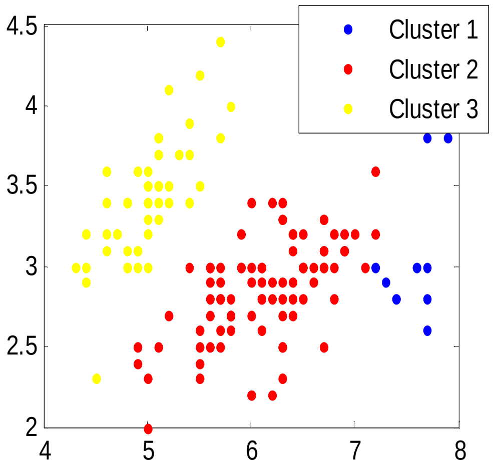


Fig. 3. Clustering results of PSO-K-means for K=3.

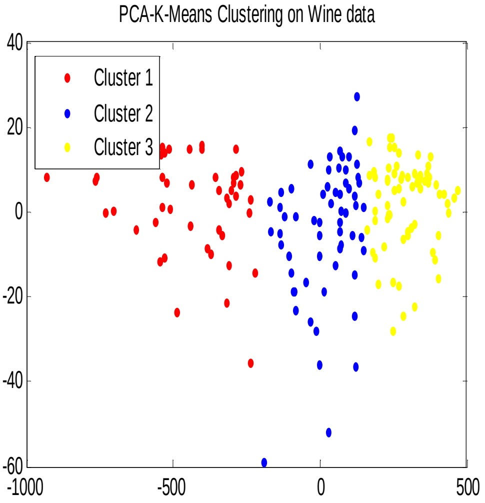


Fig.6. PCA-Kmeans of Glass for K=3

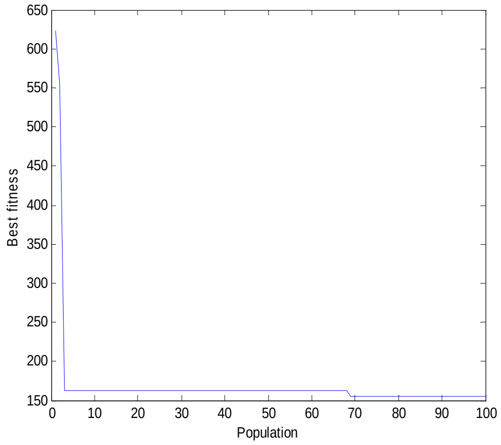


Fig.4. Optimal trajectory fitness of Iris

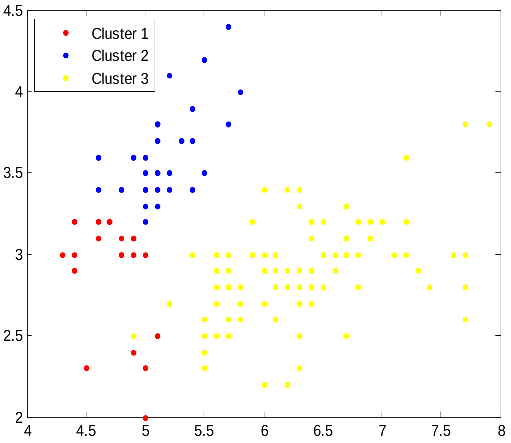


Fig.5. PCA-Kmeans of Iris for K=3

Table 1 Clustering results of Kmeans , PSO-Kmeans and PCA-Kmeans

Data sets	Number of Rows	D	Clusters	Methods	Inter	Intra	T(sec)
Iris	150	4	3	K-means	0.3391	141.9761	0.1172
				PSO-K-means	0.3021	182.65	0.2098
				PCA-K-Means	0.2907	202.94	0.10053
Wine	178	13	3	K-means	49.1408	2.1952e+004	0.2133
				PSO-K-means	49.1408	8.421e+003	0.1998
				PCA-K-Means	49.1407	7.0474e+003	0.1553
Yeast	1484	8	3	K-means	0.0324	176.8620	0.4178
				PSO-K-means	0.0321	170.0321	0.7865
				PCA-K-Means	0.0334	171.0200	0.6530
Breast Cancer	683	9	3	K-means	1.1707	619.3305	1.7723
				PSO-K-means	1.1452	900.3425	0.9231
				PCA-K-Means	1.1610	885.4082	0.1817
Glass	214	10	3	K-means	0.3919	2.5895e+003	0.2277
				PSO-K-means	0.4532	143.9837	0.4421
				PCA-K-Means	0.5505	120.5666	0.2783
Liver disorders	345	7	3	K-means	11.8121	984.89	0.7076
				PSO-K-means	11.3256	991.23	0.6789
				PCA-K-Means	10.5584	7.7847e+003	0.4081
ecoli	336	8	3	K-means	0.0556	23.8656	0.2086
				PSO-K-means	0.0556	47.4532	0.2341
				PCA-K-Means	0.0552	50.2604	0.1811

6. Conclusion

The limitations of K-means can be overcome by the two approaches PSO-K-Means and PCA-K-means. PSO -K-Means performs faster clustering as well as can avoid local minima being trapped by basic Kmeans. Hence, the proposed combination of PSO-Kmeans utilizes the globalised searching of PSO and fast convergence of Kmeans. PCA-Kmeans generates clusters with minimum sum square Error (SSE) in less computation time. The results of Kmeans, PCA-Kmeans and PSO Kmeans on various data sets have been analyzed in this paper. The experiments show that PSO-K-means and PCA-Kmeans are effective methods for partitioning large data sets.

References

[1] Kennedy, J. and Eberhart, R.C, “Particle Swarm Optimization”, in Proceedings of IEEE International conference on Neural Networks, Piscataway, New Jersey, pp. 1942-1948, 1995.

[2] Tillett, J. C. Rao, R. M. Sahin, F. and Rao,T.M, “Particle swarm optimization for clustering of wireless sensors”, in Proceedings of Society of Photo-Optical Instrumentation Engineers, Vol. 5100, No. 73, 2003.

[3] Cui, X. Palathingal, P. and Potok, T. E, “Document clustering using particle swarm optimization”, in IEEE Swarm Intelligence Symposium, Pasadena, California, pp. 185–191, 2005.

[4] Alireza, Ahmadyfard and Hamidreza Modares, “Combining PSO and k- means to enhance data clustering”, 2008 International Symposium on Telecommunications, pp. 688-691, 2008.

[5] M. Clerc, and J. Kennedy, “The particle swarm - explosion, stability, and convergence in a multidimensional complex space”, IEEE Transactions on Evolutionary Computation, vol. 6(1), pp. 58-73, 2002.

[6] Sun j, Feng B, Xu W, “Particle Swarm Optimization with particles having Quantum Behavior”, in Proceedings of Congress on Evolutionary Computation, Portland(OR,USA), pp.325-331, 2004.

[7] J. Kennedy, “Some issues and practices for particle swarms”, in IEEE Swarm Intelligence Symposium, pp. 162-9, 2007.

[8] R. Poli, J. Kennedy,T. Blackwell, “Particle swarm optimization”, Swarm Intelligence, vol. 1(1), pp. 33-57, 2007.

[9] R. Poli, “An analysis of publications on particle swarm optimization applications”, Essex, UK: Department of Computer Science, University of Essex, May - Nov. 2007.

[10] Halkidi, M., Y. Batistakis, and M. Vazirgiannis, “On Clustering Validation Techniques”, J. Intell. Inf. Syst., 2001. 17(2-3): p. 107-145.

[11] Chris Ding and Xiaofeng He, “ K-Means Clustering via Pricipal Component Analysis”, in proceedings of the 21st International conference on machine learning, Banff, Canada, 2004.

[12] S.Kalyani K.S Swarup, “Particle Swarm Optimization based K-Means Clustering approach for security assessment in power systems”, Elsevier, Expert Systems with Applications Volume 38, Issue 9, pp. 10839-10846, Sep 2011.

[13] Chris Ding , Xiaofeng He, “K-means clustering via principal component analysis”, ICML '04 Proceedings of the twenty-first international conference on Machine learning ACM New York, ISBN:1-58113-838-5, 2004.