Projected Clustering Using Particle Swarm Optimization

Satish Gajawada\textsuperscript{a}, Durga Toshniwal\textsuperscript{a}

\textsuperscript{a}Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, India

Abstract

Clustering methods divide the dataset into groups of similar objects, where objects in the same group are similar and objects in different groups are dissimilar. Traditional clustering techniques that find clusters in full dimensional space may fail to find clusters in high dimensional data due to various problems associated with clustering on high dimensional data. Subspace and projected clustering methods find clusters that exist in subspaces of dataset. These methods provide solutions to challenges associated with clustering on high dimensional data. Projected clustering methods output subspace clusters where one point in the dataset belongs to only one subspace cluster. Points may be assigned to multiple subspace clusters by subspace clustering methods. Projected clustering is preferable to subspace clustering when partition of points is required. Particle swarm optimization (PSO) has been proven to be effective for solving complex optimization problems. In this paper, we propose a Projected Clustering Particle Swarm Optimization (PCPSO) method to find subspace clusters that are present in the dataset. In PCPSO, Particle swarm optimization has been used to find optimal cluster centers by optimizing a subspace cluster validation index. In this paper, kmeans has been used to find neighbourhood of subspace cluster centers. The proposed method has been used to find subspace clusters that are present in some synthetic datasets.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of C3IT

Keywords: Particle swarm optimization; projected clustering; k-means clustering; high dimensional data

1. Introduction

Clustering refers to the process of dividing dataset into groups of similar objects. Each group is called as a cluster. The objects in the same cluster are more similar and objects in different clusters are dissimilar [1]. Traditional clustering methods which find clusters in full dimensional space tend to fail when applied on high dimensional data. This is because of various problems associated with clustering in high dimensional data [2]. Subspace and projected clustering methods find clusters that exist in subspaces of dataset. These methods have emerged as a possible solution to the challenges associated with
clustering high dimensional data [3]. One point can belong to only one subspace cluster in projected clustering. In subspace clustering a single point can belong to more than one subspace cluster [4].

Particle swarm optimization (PSO) has been proven to be a promising technique for solving complex optimization problems. In PSO, each particle in the search space represents a solution to a given problem. Each particle is associated with a position and velocity. Fitness function evaluates the positions of particles. The velocities of the particles are calculated according to the historical best positions of the population and positions of particles are updated in each iteration. The particles fly towards better regions through their own effort and with the cooperation of other particles [5].

In this paper, we propose a method to find subspace clusters using Particle swarm optimization. We have used a subspace cluster validation index as fitness of particles. The proposed method has been applied on synthetic data sets and it has been observed that the proposed method found subspace clusters that are present in these datasets.

This paper is organized as follows: In Section 2, related work is given. Proposed method is explained in Section 3. Section 4 contains results and discussion. Finally, we draw conclusions in Section 5.

2. Related Work

Aggarwal et al. [6] proposed PROCLUS which is a k-medoid like clustering algorithm. It randomly determines set \( M \) of scattered medoids. In the cluster refinement phase subspace is determined for each of \( k \) current medoids and points are then assigned to the closest medoid considering the relevant subspace of each medoid. The bad medoids are replaced by new medoids from set \( M \) as long as clustering quality increases. Relevant dimensions are found again from the clusters identified and points are reassigned to centers. Points that are too far away from their closest medoid are identified as noise points in postprocessing step. Procopiuc et al. [7] considered a projected cluster in method DOC to be a fixed length hypercube of width \( w \) containing atleast \( \alpha \) points. DOC selects an arbitrary point \( p \) and then it randomly selects tentative cluster for \( p \). Attributes for which projections of all the tentative cluster’s members on this attribute are within distance \( w \) from the projection of \( p \) on this attribute are taken as relevant attributes. All data points that fall within distance \( w \) from \( p \) on all attributes which are relevant for tentative cluster together form a projected cluster. Bohm et al. [8] proposed PreDeCon which uses a specialized distance measure and a full dimensional density based clustering algorithm known as DBSCAN [9].

Particle swarm optimization has been applied to clustering problems. Cui et al. [10] presented a hybrid PSO algorithm using PSO and K-means. The clustering result from PSO clustering method has been used for giving initial seeds to K-means clustering in this hybrid approach. Van der Merwe et al. [11] developed a new PSO based clustering algorithm where K-means clustering is used to seed the initial swarm. Recently, Lu et al. [5] proposed a PSO-based algorithm called PSOVW to solve the variable weighting problem in soft projected clustering of high-dimensional data. In [12] PSOVW is extended to handle the problem of text clustering. In this paper, we propose PCPSO for hard projected clustering.

3. Proposed Method

In this section our proposed PCPSO method is explained. Figure 1 shows proposed PCPSO method.

3.1 Description of proposed PCPSO method

- **Input parameters**
  
  The number of subspace clusters (k) and average number of relevant dimensions per subspace cluster (l) are input parameters required by PCPSO in addition to parameters required by PSO.
- **Encoding of particles**
  Each particle is of length K where K represents number of subspace clusters. Each dimension of the particle can take any value from the set \{1, 2, 3, 4, ..., N\} where N represents number of points in the dataset. Each particle encodes K points from input dataset which are selected as cluster centers of K subspace clusters.

- **Initial population**
  The initial population is selected randomly. Each dimension of the particle can take any value from the set \{1, 2, 3, 4, ..., N\} where N represents number of points in the dataset.

- **Fitness evaluation**
  Each solution in the population has a fitness value which depends on the function to be optimized. Decoding particle gives K cluster centers of subspace clusters. K-means clustering is applied on input dataset with K cluster centers given by PSO as initial seeds. Cluster centers are updated to those centers.
given by K-means clustering. Clusters obtained by K-means clustering on input dataset are taken as neighborhood points of corresponding cluster centers. Neighbourhood points are then used for identifying relevant attributes for subspace clusters. Points are assigned to centers using relevant attributes found. Relevant attributes are found again using clusters obtained after assigning points to centers. The points are again reassigned to centers using relevant attributes found. Subspace clusters obtained are validated using subspace cluster validation. Subspace cluster validation index has been taken as fitness value of the particles.

- Updating positions of particles
  The velocities of the particles are calculated according to the best positions (pBest of particle and gBest of population). The positions of particles are updated in each iteration using velocities of particles calculated in that iteration.

- Termination
  If termination condition is satisfied then algorithm terminates and returns optimal cluster centers.

4. Experimental results

We applied PCPSO method on synthetic datasets. Results obtained for a dataset having 9 subspace clusters with 14 average number of relevant dimensions per subspace cluster are discussed below.

Table 1. Matching points between output and input clusters

| Output cluster obtained by PCPSO | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|---------------------------------|----|----|----|----|----|----|----|----|
| Input cluster that is matched with output cluster | D  | I  | A  | H  | B  | F  | G  | E  |
| No. of points in input cluster | 152| 150| 300| 153| 152| 151| 149| 153|
| No. of points in output cluster | 160| 155| 317| 167| 260| 153| 161| 154|
| No. of common points between output and input clusters | 152| 150| 300| 153| 152| 151| 149| 153|

Table 2. Number of mismatched dimensions between output cluster and input cluster

| Output cluster | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|----------------|----|----|----|----|----|----|----|----|
| Matched input cluster | D  | I  | A  | H  | B  | F  | G  | E  |
| No. of mismatched dimensions | 0  | 0  | 1  | 0  | 3  | 0  | 0  | 0  |

Subspace clusters are obtained by running steps in fitness function of PCPSO once again with optimal cluster centers given by PCPSO. Table 1 shows that output subspace clusters 2 to 9 obtained by PCPSO has common points with input subspace clusters A to I (excluding cluster C ) that are present in dataset. Cluster C got merged with cluster B and identified as output cluster 6. Set of noise points are identified as output cluster 1. Hence cluster 1 is not shown in Table 1 and Table 2. From Table 1 we can observe that the sizes of input clusters and number of common points are equal. This shows that each output cluster (except output cluster 1) has matched to input cluster that is present in dataset. The size difference between output clusters and input clusters is due to noise points present in the dataset. From Table 2 we can observe that there are only 4 mismatches between relevant dimensions of subspace clusters and relevant dimensions of subspace clusters identified by using PCPSO.
5. Conclusion

In this paper, we proposed Projected Clustering Particle Swarm Optimization (PCPSO) method for projected clustering. We applied PCPSO on synthetic datasets to obtain subspace clusters present in these datasets. Our future work includes using different methods for encodings for particles, neighbourhood generation, identifying relevant attributes and different subspace cluster validation indices in our PCPSO method. Results can be improved by adding a step for removing noise points before subspace cluster validation step in PCPSO method. Our future work also includes combining hierarchical clustering and PCPSO for improving results by creating methods like HC-PCPSO.

References

1. Das S, Abraham A, Konar A. Automatic Clustering Using an Improved Differential Evolution Algorithm. IEEE Transactions On Systems, Man, And Cybernetics-Part A: Systems And Humans 2008; 38:218-237.
2. Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: A review. SIGKDD Explor 2004; 6:90-105.
3. Moise G, Zimek A, Kroger P, Kriegel HP, Sander J. Subspace and projected clustering: experimental evaluation and analysis. Knowl. Inf. Syst. 2009; 3:299-326.
4. Kriegel HP, Kroger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl. Discov. Data. 2009; 3:1-58.
5. Lu Y, Wang S, Li S, Zhou C. Particle swarm optimizer for variable weighting in clustering high-dimensional data. Mach. Learn. 2011; 82: 43-70.
6. Aggarwal CC, Procopiuc CM, Wolf JL, Yu PS, Park JS. Fast algorithms for projected clustering. In Proceedings of the ACM International Conference on Management of Data (SIGMOD), 1999, p 61-72.
7. Procopiuc CM, Jones M, Agarwal PK, Murali TM. A Monte Carlo algorithm for fast projective clustering. In Proceedings of the ACM International Conference on Management of Data (SIGMOD), 2002, p 418-427.
8. Bohm C, Kailing K, Kriegel HP, Kroger P. Density connected clustering with local subspace preferences. In Proceedings of the 4th International Conference on Data Mining (ICDM), 2004, p 27-34.
9. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), 1996, p 291-316.
10. Cui X, Potok TE, Palafhingal P. Document clustering using particle swarm optimization. In 2005 IEEE Swarm Intelligence Symposium, 2005, p 185-191.
11. Van der Merwe DW, Engelrecht AP. Data Clustering using Particle Swarm Optimization. The 2003 Congress on Evolutionary Computation, 2003, p 215-220.
12. Lu Y, Wang S, Li S, Zhou C. Text Clustering via Particle Swarm Optimization. IEEE Swarm Intelligence Symposium, 2009, p 45-51.