Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

Projected Clustering Particle Swarm Optimization and Classification

Satish Gajawada[+] and Durga Toshniwal

Department of Electronics and Computer Engineering,
Indian Institute of Technology Roorkee, Roorkee, India.

Abstract. Supervised learning algorithms are trained with labeled data only. But labeling the data can be costly and hence the amount of labeled data available may be limited. Training the classifiers with limited amount of labeled data can lead to low classification accuracy. Hence pre-processing the data is required for getting better classification accuracy. Full dimensional clustering has been used in literature as pre-processing step to classification methods. But in high dimensional data different clusters may exist in different subspaces of the dataset. Projected Clustering Particle Swarm Optimization (PCPSO) finds optimal centers of subspace clusters by optimizing a subspace cluster validation index. In this paper we use PCPSO method to find subspace clusters present in the dataset. The subspace clusters found and limited amount of available labeled data are used to label the large amount of unlabelled data that is present in the dataset. Various classification methods are then applied on the data pre-processed by using PCPSO. In this paper we propose PCPSO-Classification method. Various new classification methods like PCPSO-Naive bayes, PCPSO-Multi layer perceptron and PCPSO-Decision table can be obtained by using different classification methods like Naive bayes, Multi layer perceptron and Decision table respectively in the classification stage of proposed PCPSO-Classification method. When the dataset contains subspace clusters and labeling the data is costly due to which available labeled data is limited then the structure of data may be used along with available limited labeled data to label the large amount of unlabeled data. After pre-processing the data the amount of labeled data is not limited. We applied PCPSO-Naive bayes, PCPSO-Multi layer perceptron and PCPSO-Decision table methods on synthetic datasets and found classification accuracy improved significantly compared to using Naive bayes, Multi layer perceptron and Decision table for classification with limited available labeled data for training classifiers. The subspace clusters found by PCPSO can be used for different types of pre-processing for solving different problems before applying classification methods on datasets. In this paper we considered the problem of limited labeled data and using PCPSO to find subspace clusters which are used for labeling large amount of unlabeled data with the help of available limited labeled data.

Keywords: Projected clustering, Particle swarm optimization, Pre-processing, Classification.

1. Introduction

Clustering algorithms divide the dataset into set of disjoint clusters. Traditional clustering methods tend to fail when applied on high dimensional data due to various problems associated with clustering in high dimensional data [1]. 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 in high dimensional data [2]. In subspace clustering one object may be assigned to multiple subspace clusters but in projected clustering one point can belong to only one subspace cluster. Projected clustering is preferred over subspace clustering when partition of points is required [3]. Classification methods learn to classify new objects by using a set of training objects which contain class labels. Clustering methods have been used in literature together with classification methods for improving classification [4]. But in high dimensional datasets different clusters may exist in different subspaces. Hence there is a scope to explore using of clustering methods which find subspace clusters to enhance classification performed by classifiers.

[+] Corresponding author. Tel.: +91-9997090355.
E-mail address: gajawadasatish@gmail.com.

Particle swarm optimization (PSO) has been applied for solving complex optimization problems. In PSO, the solutions to a given problem are represented by particles. Each particle is associated with a position and velocity. The positions of particles are evaluated by fitness function. The velocities of the particles are calculated using the historical best positions of the population and positions of particles are updated in each iteration. The particles move towards better regions through their own effort and with the cooperation of other particles [5].

In this paper, we propose hybrid methods for classification using PCPSO and classification methods. The proposed methods have been applied on synthetic data sets and it has been observed that the proposed methods gave better accuracy compared to directly applying classification methods on datasets.

This paper is organized as follows: Related work is given in Section 2. The explanation of proposed method is present in Section 3. Section 4 contains results and discussion. Finally, we draw conclusions in Section 5.

2. Literature Review

Aggarwal et al. [6] proposed PROCLUS which is a k-medoid like clustering algorithm. Procopiuc et al. [7] developed DOC by considering a projected cluster to be a fixed length hypercube of width w containing atleast α points. 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. Satish Gajawada et al. [13] proposed PCPSO for finding optimal cluster centers of subspace clusters. Subspace clusters can be found by using optimal cluster centers given by PCPSO.

Antonia Kyriakopoulou et al. [14] clustered both training and testing data before the classification step, in order to extract the structure of the whole dataset. Data is pre-processed by adding artificial meta-features based on the clustering result. A more efficient classifier was built by applying classification method on the processed data. Recently, Patil et al. [4] proposed effective framework for prediction of disease outcome using clustering and classification. Labels have been assigned through clustering and assigned labels were matched with given labels to preprocess the data. Classification has been done on preprocessed data. The proposed framework obtained promising classification accuracy as compared to other methods found in literature. Hence clustering methods which find clusters in full dimensional space have been used as pre-processing step for classification methods. But for high dimensional datasets with subspace clusters clustering methods which find clusters that exist in subspaces of dataset need to be used to extract structure of the dataset which can be used for enhancing results given by classification methods.

3. Proposed Method

Figure 1 shows proposed PCPSO-Classification method. Section 3.1 explains PCPSO-Classification method.

3.1. Description of Proposed Method

PCPSO-Classification method consists of two stages. In the first stage PCPSO [13] is applied on whole dataset to find subspace clusters that are present in the high dimensional dataset. The limited amount of labelled data and subspace clusters obtained by PCPSO are used to find labels of unlabelled points. In the second stage classification methods like Naive Bayes [14], Decision table [14] and Multi layer perceptron [14] are applied for building a classifier using data pre-processed by PCPSO.

Various steps in PCPSO-Classification method are given below:

- Particle encoding for PCPSO: 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.

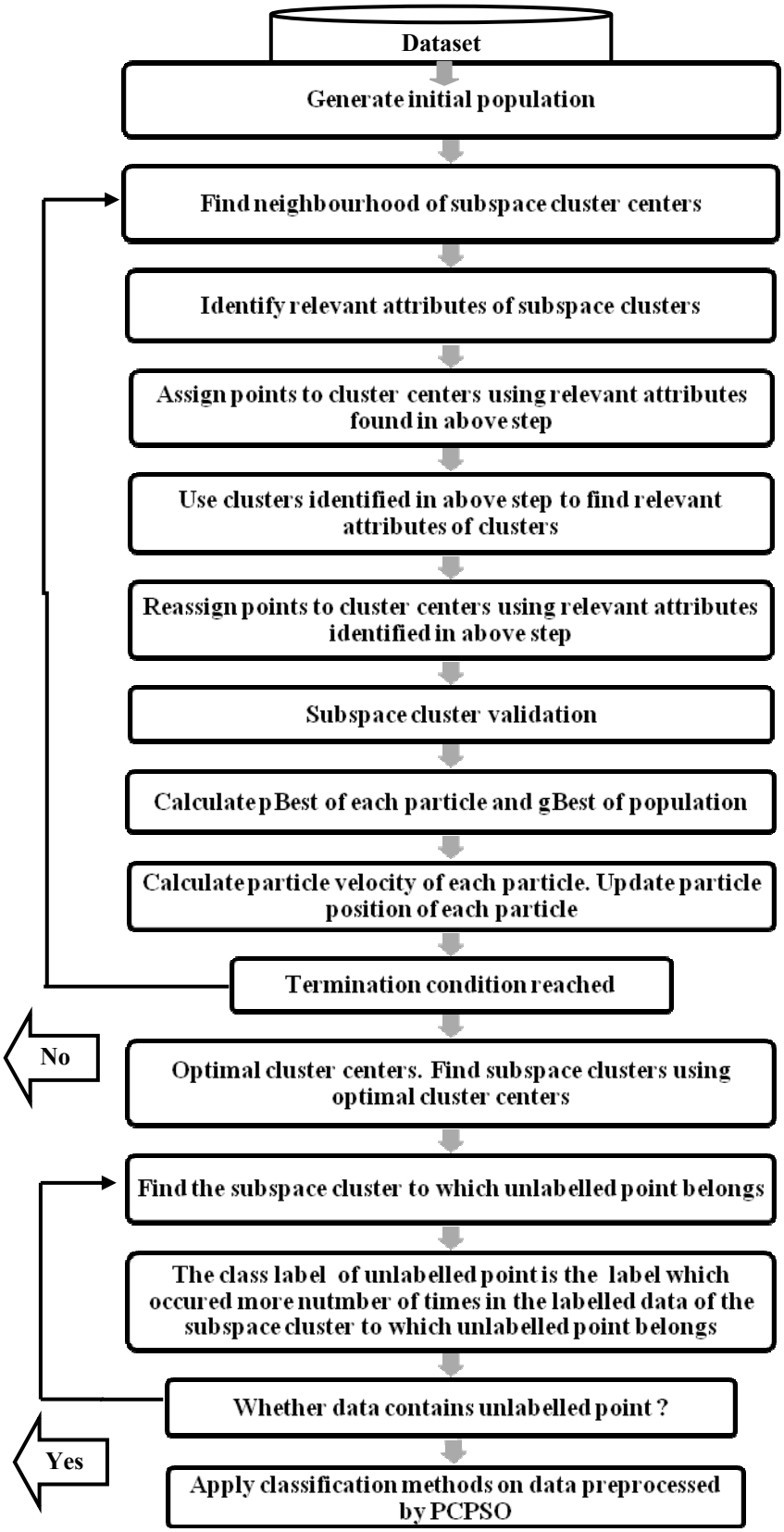

Fig. 1: Proposed PCPSO-Classification method

- Fitness function for PCPSO: Decoding particle gives K cluster centers of subspace clusters. Neighbourhood of centers is obtained by using method described in [6]. 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 cluster validation index has been taken as fitness value of the particles.
- PCPSO method is applied on complete dataset which contains labeled and unlabelled points to get optimal cluster centers of subspace clusters. Subspace clusters are obtained by using optimal cluster centers given by PCPSO.
- Label the unlabelled points in the subspace clusters using the labeled points in those clusters. The class label which occurred more number of times in the labeled data of the subspace cluster to which unlabelled point belongs is taken as the class label of the unlabelled point.
- Classification method is applied on the data preprocessed by PCPSO.

4. Experimental Results

We applied proposed PCPSO-Classification method on synthetic datasets [15]. PCPSO has been used as pre-processing step for various classification methods. Results obtained for a synthetic dataset having 9 subspace clusters with 14 average numbers of relevant dimensions per subspace cluster are discussed below. Randomly 97 percent points have been selected as unlabelled data. Although synthetic dataset contains labels for all the points we have considered labels of less number of points to show the advantage of proposed method when dataset contains labels only for less number of points.

Table 1: Classification accuracy obtained by using different classifiers

Classification Method	Accuracy
Naive bayes	84.8093
Multi layer perceptron	90.4977
Decision table	78.9916

Table 2: Classification accuracy obtained by using different classifiers in our proposed method

Proposed Classification Method	Accuracy
PCPSO-Naive bayes	91.2853
PCPSO-Multi layer perceptron	96.1755
PCPSO-Decision table	91.8495

Table 1 shows results obtained by directly applying classification methods on the dataset with available 3 percent labeled data as training data. Table 2 shows the results obtained after pre-processing the dataset with PCPSO and then applying classification methods with 10-fold cross validation on pre-processed data.

From Table 1 we can observe that directly performing classification with Decision table gave accuracy around 79 percent. But PCPSO-Decision table gave more than 91 percent accuracy which we can find from Table 2. This large difference in classification accuracy between proposed PCPSO-Decision table and Decision table classification is due to that fact that the data has very limited labeled data. But pre-processing the data by using PCPSO to get structure present in the dataset can be helpful to increase the amount of available labels. After processing the data using PCPSO the amount of labeled data is not limited and hence

applying classification method on processed data showed significant improvement in accuracy. From Table 1 and Table 2 we can find that PCPSO-Naive bayes, PCPSO-Multi layer perceptron showed improvement compared to Naive bayes and Multi layer perceptron respectively.

There is scope for other kind of pre-processing steps with projected clustering before classification stage. Working in this direction will lead to new projected clustering-Classification methods which can yield better results compared to directly applying classification methods on datasets.

5. Conclusion

In this paper we proposed PCPSO-Classification method. The results obtained on synthetic datasets showed that pre-processing the high dimensional data which has very less amount of labeled data with PCPSO can improve accuracy of classifiers significantly. Our future work includes creation of other new methods where projected clustering is used in combination with classification methods to improve the results compared to applying classification methods directly on high dimensional data with subspace clusters. Our future work also includes creation of various methods like HC-PCPSO-Classification where hierarchical clustering is used as initialization method for PCPSO to get better results compared to PCPSO-Classification. There is scope for other things like using Projected Clustering Differential Evolution (PCDE) for pre-processing high dimensional data by creating methods like PCDE-Classification and HC-PCDE-Classification similar to PCPSO-Classification and HC-PCPSO-Classification.

6. References

[1] L. Parsons, E. Haque, H. Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explor*. 2004, **6**: 90-105.

[2] G. Moise, A. Zimek, P. Kroger, H.P. Kriegel, J. Sander. Subspace and projected clustering: experimental evaluation and analysis. *Knowl. Inf. Syst.* 2009, **3**: 299-326.

[3] H.P. Kriegel, P. Kroger, A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data.* 2009, **3**: 1-58.

[4] B.M. Patil, R. C. Joshi and Durga Toshniwal. Effective framework for prediction of disease outcome using medical datasets: clustering and classification. *Int. J. Computational Intelligence Studies* 2010, **1** (3): 273-290.

[5] Y. Lu, S. Wang, S. Li, C. Zhou. Particle swarm optimizer for variable weighting in clustering high-dimensional data. *Mach. Learn.* 2011, **82**: 43-70.

[6] C.C. Aggarwal, C.M. Procopiuc, J.L. Wolf, P.S. Yu, J.S. Park. Fast algorithms for projected clustering. In: *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*. 1999, pp. 61-72.

[7] C.M. Procopiuc, M. Jones, P.K. Agarwal, T.M. Murali. A Monte Carlo algorithm for fast projective clustering. In: *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*. 2002, pp. 418-427.

[8] C. Bohm, K. Kailing, H.P. Kriegel, P. Kroger. Density connected clustering with local subspace preferences. In: *Proceedings of the 4th International Conference on Data Mining (ICDM)*. 2004, pp. 27-34.

[9] M. Ester, H.P. Kriegel, J. Sander, X. Xu. 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, pp. 291-316.

[10] X. Cui, T.E. Potok, P. Palafhingal. Document clustering using particle swarm optimization. In: *IEEE Swarm Intelligence Symposium*. 2005, pp 185-191.

[11] D.W. Van der Merwe, A.P. Engelhrecht. Data Clustering using Particle Swarm Optimization. In: *Congress on Evolutionary Computation*. 2003, pp 215-220.

[12] Y Lu, S Wang, S Li, C Zhou. Text Clustering via Particle Swarm Optimization. In: *IEEE Swarm Intelligence Symposium*. 2009, pp 45-51.

[13] Satish Gajawada, Durga Toshniwal. Projected Clustering Using Particle Swarm Optimization. In: *2nd International Conference on Computer, Communication, Control and Information Technology (C3IT).* 2012.

[14] A. Kyriakopoulou, T. Kalamboukis. Using clustering to enhance text classification. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2007, pp. 805-806.

[15] E. Muller, S. Gunnemann, I. Assent, T. Seidl. Evaluating Clustering in Subspace Projections of High Dimensional Data. In: *Proc. 35th International Conference on Very Large Data Bases (VLDB)*. 2009.