Hindawi Publishing Corporation
The Scientific World Journal
Volume 2015, Article ID 107650, 5 pages
http://dx.doi.org/10.1155/2015/107650

Research Article

Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence

Thenmozhi Srinivasan[1] and Balasubramanie Palanisamy[2]

[1]*Department of Computer Applications, Gnanamani College of Technology, AK Samuthiram, Pachal, Namakkal District, Tamil Nadu 637 018, India*
[2]*Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode, Tamil Nadu 638 052, India*

Correspondence should be addressed to Thenmozhi Srinivasan; thenmozhi.s1983@gmail.com

Received 13 April 2015; Accepted 21 April 2015

Academic Editor: Venkatesh Jaganathan

Clusters of high-dimensional data techniques are emerging, according to data noisy and poor quality challenges. This paper has been developed to cluster data using high-dimensional similarity based PCM (SPCM), with ant colony optimization intelligence which is effective in clustering nonspatial data without getting knowledge about cluster number from the user. The PCM becomes similarity based by using mountain method with it. Though this is efficient clustering, it is checked for optimization using ant colony algorithm with swarm intelligence. Thus the scalable clustering technique is obtained and the evaluation results are checked with synthetic datasets.

1. Introduction

Data mining deals with extracting useful information from datasets [1]. In data mining, clustering is a process which recognizes similar description (homogenized) groups of data on the basis of their size (profile). This involves a distance metric, in which the data points in each partition are similar to points in different partitions. A large number of data mining techniques to cluster the data are available. Few of them are CLARANS [2], Focused CLARANS [3], BIRCH [4], DBSCAN [5], and CURE [6]. Clustering is a technique that is required for various applications in pattern analysis, decision making, group and machine document retrieval, learning environment, pattern classification, and image segmentation. Clustering the dataset includes measures distance or similarity measure to partition the dataset, where data inside the clusters is similar to data outside the cluster.

The preprocessing steps are carried out before clustering, in which feature extraction and feature selection are major parts. This reduces the burden of clustering algorithm. The selection process is being done by eliminating the redundant ones. These techniques increase the speed of clustering algorithms and hence performance is improved [6]. Many feature selection techniques are proposed in the literature to reduce the size of the dataset. In those cases, reducing the dimensions using conventional feature selection leads to significant loss of data. In traditional clustering algorithm, the clustering algorithm follows distance function, but the distance function used by the algorithms to all aspects of equal treatment is not of equal importance. Where the selection of feature techniques reduces dimensions by removing irrelevant features they may have to eliminate many of the features associated with the presence of sporadic failure. For this a new class of projected clustering arises in this technique.

A projected clustering is also called subspace clustering [7] which has high-dimensional datasets, a unique group of data points that are correlated with different sets of dimensions, where the focus is to determine a set of attributes for each cluster. A set of relevant features shows the accuracy of the cluster, which has potential applications in e-commerce [8], a computer vision task [9]. There are various approaches proposed for projected clustering in the past. Few of them are CLIQUE, DOC, Fast DOC, PROCLUS, ORCLUS, and HARP [10]. After the clustering process in data mining, there is a need to check whether it is an optimized technique or not. There are various optimization techniques in queue, in which

the ant colony cluster algorithm uses the characteristics of positive feedback, which is a good convergent parallel algorithm.

2. Related Works

A projected hierarchical clustering algorithm called hierarchical approach with automatic relevant dimension selection (HARP) is proposed in [10]. HARP works with the assumption that if the data points are similar in high-dimensional space, they also show the same similarity in lower-dimensional space. Depending on this criterion, if clusters are similar in various numbers of dimensions, they are allowed to merge. The similarity and minimum number of similar dimensions can be controlled dynamically, without the help of parameters given by user. The advantage of HARP is that it determines automatically the dimensions for each cluster without the use of input parameters, whose values are difficult to define. HARP [11] also provides interesting results on gene expression data.

The efficiency of DOC/FastDOC has been studied using FPC in [12] and has proved that it is much faster than the previous method, which also has some drawbacks, such as FPC that fits well when the cluster is in the form of a hypercube with parameter values clearly specified. A type of density based clustering algorithm for projected clustering is proposed in [13], which uses histogram construction known as efficient projective clustering by histograms (EPCH). Dense regions are spotted in each histogram by lowering threshold iteratively; for each data point corresponding to a region in a subspace a signature is generated. By identifying signatures for a large number of data points [13], projected clusters are uncovered. EPCH is a type of compression based clustering algorithm; it is faster and can handle irregular clusters, but in full-dimensional space it cannot compute distance between data points. Fuzzy clustering method (FCM) [14] is based on partition. FCM's main disadvantage is that it features the request for the number of clusters to be generated. In addition, when the data affected by noise is high, it can lead to clusters of poor quality because FCM is highly sensitive to outliers. Prediction in mobile mining for location based services to determine precious information is studied by Venkatesh et al. [15]. Clustering nonspatial data using similarity based PCM (SPCM) is proposed in [14].

3. Proposed Methodology

The proposed methodology is focused to cluster high-dimensional data in projected space. The main idea behind the development of PCM SPCM based application is to integrate with mountain cluster. SPCM has merit that it can automatically produce results clustering without requiring users to determine the number of clusters. Though this is efficient clustering, it is checked for optimization using ant colony algorithm with swarm intelligence. Thus the scalable clustering technique is obtained and the evaluation results are checked with synthetic datasets. This work is done based on the block diagram shown in Figure 1.

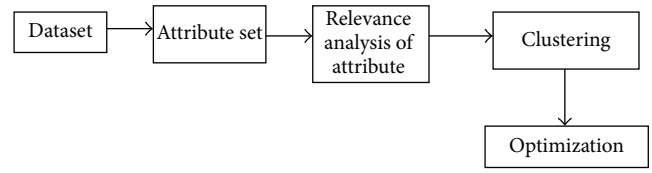

Figure 1: Block diagram of proposed methodology.

Let the dataset be taken as DS, having d-dimensional points, where attributes are denoted as $D = \{D_1, D_2, \ldots, D_d\}$. Then N data points are initialized as

$$Y = \{y_1, y_2, \ldots, y_N\}, \quad \text{where } y_i = y_{i1}, \ldots, y_{ij}, \ldots y_{id}. \quad (1)$$

In this, each y_{ij} $(i = 1, \ldots, N; j = 1, \ldots, d)$ corresponds to the value of y_i data point on D_j attribute.

The cluster will see a considerable number of aspects related to the above relationship in which points are close to each other in a large number. The three main steps are as follows:

(i) attribute relevance analysis,

(ii) clustering,

(iii) optimization.

3.1. Attribute Relevance Analysis. In attribute relevance analysis, cluster structures are displayed by identifying dense regions and their location in each dimension. For projected clustering, a cluster must contain relevant dimensions of the data in which the projection of each point of the cluster is close to a sufficient number of other projected points. The dimensions identified represent potential candidates for the clusters. Using the cluster structure, a region having high density of points is chosen compared to its surrounding regions, which represent the 1D projection of clusters. By detecting dense regions in each dimension the discrimination between dimensions that are relevant and irrelevant can be detected, and sparseness degree is then computed to detect densely populated regions in each attribute.

3.2. SPCM Based Clustering. After the completion of the first step, the dimension of dataset is given for reduction. It incorporates outlier elimination using sparseness estimation and similarity based possibility C-means (SPCM) algorithm. Spatial data [16] defines a location, shape, size, and orientation and it includes spatial relationships whereas nonspatial data is information which is independent of all geometric considerations. In general spatial data are multidimensional and autocorrelated and nonspatial data are one-dimensional and independent. After identification of clusters, by selecting appropriate dimensions the result is refined. The advantage of PCM is the membership function and the number of clusters is independent, and in a noisy environment with outliers it is highly vigorous [4]. When the data is a set of samples drawn from stationary processes, a framework for defining consistency of clustering algorithms is proposed in [17]. An absolute degree of data object and membership

function in a cluster is assumed. There is no dependency for the membership degree on the same data objects in other clusters. So they are independent in this technique, as given in

$$U_{ip} = \frac{1}{1 + \left(d_{ip}^2/\eta_i\right)^{1/(m-1)}}, \quad (2)$$

where u_{ij} is the membership degree for object and d_{ij} is the Euclidean distance between object x_j and prototype v_i. The parameter η_i determines the distance of membership function equals 0.5. It can be expressed as

$$\eta_i = P\frac{\sum_{p=1}^{N} u_{ip}^m d_{ip}^2}{\sum_{p=1}^{N} u_{ip}^m}, \quad (3)$$

where P is assigned to a value of one. The prototype for inner product distance measure is as given in

$$v_i = \frac{\sum_{p=1}^{n} u_{ip}^m x_p}{\sum_{p=1}^{N} u_{ip}^m}. \quad (4)$$

This algorithm follows the proceeding steps.

Step 1. A number of clusters are fixed c; fix $1 < m < \infty$; set iteration counter $l = 1$; initialize possible c-partition $U^{(O)}$; then estimate η_i.

Step 2. Update the prototypes $V^{(l+1)}$ using $U^{(l)}$.

Step 3. Compute $U^{(l+1)}$ using $V^{(l+1)}$.

Step 4. Consider $l = l + 1$.

Step 5. If $|U^{(l+1)} - U^{(1)}| < \varepsilon$, then process stops; else go to Step 2.

3.3. Mountain Method. Yager and Filev [18] proposed the mountain method which is applicable for searching approximate centers in the cluster, where the maximum of density measures are located. In Algorithm 1 one cluster is identified at a time and it is removed by eliminating density based function for other points.

In m-dimensional space, R^m, n data points are chosen as $\{x_1, \ldots, x_n\}$. Let x_{pj} denote the jth coordinate of the pth point, where $p = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. Mountain method discretizes the feature space which forms an m-dimensional grid in hypercube $I_1 \times \cdots \times I_m$ with nodes $N_{(i_1,\ldots,i_m)}$, where $i_i, \ldots, i_m$ chooses values from the set $[1, \ldots, r_1], \ldots, [1, \ldots, r_m]$. The intervals I_j are used to define the range of coordinates x_{pj}, where the intervals I_j are discretized into r_j equidistant points.

The mountain function is calculated for the vertex point V_i defined as given in

$$M^1(V_i) = \sum_{p=1}^{n} e^{-\alpha d(x_p, v_i)}, \quad (5)$$

where $d(x_p, V_i)$ is the distance from the data point x_p to the grid nodes V_i and α is the positive constant. The cluster centers are selected by the nodes having the maximum number of mountain functions. To find the cluster centers the mountain function is defined as given in

$$M^p(V_i) = M^{p-1}(V_i) - M'_{p-1} \sum_{p=1}^{n} e^{-\beta d(x_{p-1}, v_i)}, \quad (6)$$

where M'_{p-1} is taken as the maximal value of mountain function and M_{p-1} and β are positive constants.

3.4. Ant Colony Optimization Based on Swarm Intelligence. The optimization algorithm is designed in such a way that probability function is used as basic pick-up and pick-down. This creates increase in similitude degree of system and maximizes the system's average similitude degree. Dorigo and Socha [19] who proposed ant colony system solve combinatorial optimization problems. Ant colony cluster [20] is motivated by the accumulation of ant bodies and ant larvae classification. The ant colony cluster algorithm uses the positive feedback characteristics. The ant conveying process is followed as the main process in the ant cluster algorithm. To convey the object to the destination, the ant considers the equivalent degree between the current object and the surrounding objects; by this way the ant will not know the other ant's location distributing load status.

Thus the ant conveying process is a simple, flexible, easy, and absolute individual behavior where objects are distributed and divided into several clusters during long time, subsequently as a concurrent process. The ant's observing radius is the main factor that influences this process, and the cluster is more efficient when radius is small [4]. Many isolated points are created, where the combination of the clusters is influenced when the observing radius is too small, which cause the inefficiency of the cluster capability. When the observing radius is large, the algorithm's convergence speed in return is expedited. The main advantage of traditional ant cluster algorithm is the adjustment of observing radius and the ant's "memory" function, and it also benefits in magnitude ameliorating aspect.

4. Experimental Results

The experiment is evaluated using the synthesis dataset implemented in MATLAB R2012 platform. The evaluation of this work is compared with existing method of PCKA for WDBC and MF, and the two datasets WDBC and MF are chosen because they are real and synthetic datasets. The performance results are evaluated using clustering accuracy and immunity to outliers.

Table 1 shows the clustering accuracy of the proposed and existing technique for the WDBC and MF dataset. The proposed technique achieves higher accuracy for both datasets when compared with existing PCKA technique. The reliability of PCKA technique is 90%.

Figure 2 shows the immunity of outlier for the 2 percent of d, where the existing technique PCKA is affected more

```
For every item D_i do
        Place D_i randomly on grid
End For
For all ants do
        Place ant at randomly selected site
End For
For t = 1 to t_max do
        For all ants do
            If ((ant Unladed) & (site S_j occupied by item D_i)) then
                Compute the similarity f(D_i) of D_i in A × A
                Calculate the R_R and generate a random number P
                If R_R > P then
                   Pick up item D_i
                   Takes the f(D_i) and current position
                   Move the ant with the item D_i to random site.
                Else
                  Move the empty ant to random site
                End If
            Else if ((agent carrying) and (site empty)) then
                Compute the similarity f(D_i) of D_i
                Calculate the R_R and generate a random number P
                If R_R > P then
                   Drop item
                End If
            Else
              Move to a randomly selected neighboring site
            End If
        End For
        If ((t > 0.5t_max) and (if t attains radius change condition)) then
            Reduce the radius
         Clusters are generated iteratively and cluster centers are calculated
         Connect the obtained clusters with the same cluster center
         With Poor Similarity Relocate the Items
       End If
End For
Print location of items.
```

Algorithm 1: Ant colony optimization algorithm.

Table 1: Accuracy of clustering.

Dataset	Proposed	Existing PCKA [4]
WDBC	94.52	91.56
MF	93.7	90.45

when compared with proposed technique and Figure 3 shows the immunity of outlier for the 30 percent of d, where the existing technique PCKA is affected more when compared with proposed technique.

5. Conclusion

An efficient high-dimensional dataset clustering is proposed in this work with its optimized results. The results were analyzed through performance evaluation based on real datasets which are synthetic. Here the SPCM method has shown its excellent performance for similarity based clustering in terms of quality, though the noisy environment dealt with outliers.

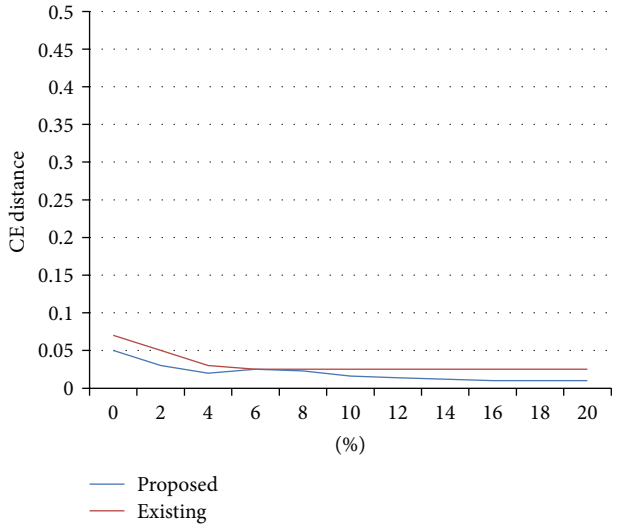

Figure 2: Immunity to outliers for datasets with $l_{\text{real}} = 2$ percent of d.

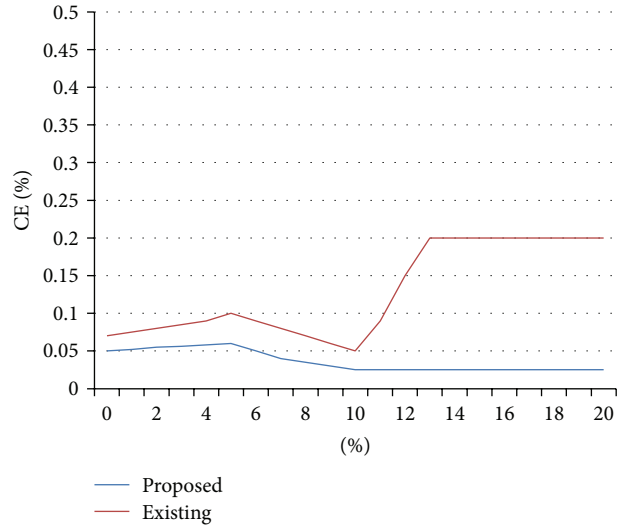

FIGURE 3: Immunity to outliers for datasets with $l_{\text{real}} = 30$ percent of d.

This solves the problem of other fuzzy clustering methods which deals with similarity based clustering applications in the sets. This work is composed of SPCM technique which finds the clusters automatically without users input of number of clusters. And the optimized clustering result is obtained by using ant colony optimized technique with swarm intelligence. It is robust and helps to produce scalable clustering. The results given in the above section show its efficient clustering accuracy and outlier detection. The scope of further research is to deal with datasets that have a large number of dimensions.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data," *Data Mining and Knowledge Discovery*, vol. 11, no. 1, pp. 5–33, 2005.

[2] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park, "Fast algorithms for projected clustering," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '99)*, pp. 61–72, 1999.

[3] M. Ester, H. P. Kriegel, and X. Xu, "A database interface for clustering in large spatial databases," in *Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining*, Montreal, Canada, 1995.

[4] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," in *Proceedings of the ACM SIGMOD Conference on Management of Data*, Montreal, Canada, 1996.

[5] K. Y. Yip, D. W. Cheung, and M. K. Ng, "A review on projected clustering algorithms," *International Journal of Applied Mathematics*, vol. 13, pp. 24–35, 2003.

[6] S. Guha, R. Rastogi, and K. Shim, "CURE: an efficient clustering algorithm for large databases," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 73–84, June 1998.

[7] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[8] M. Bouguessa and S. Wang, "Mining projected clusters in high-dimensional spaces," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 4, pp. 507–522, 2009.

[9] M. L. Yiu and N. Mamoulis, "Iterative projected clustering by subspace mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 176–189, 2005.

[10] K. Y. L. Yip, D. W. Cheung, and M. K. Ng, "HARP: a practical projected clustering algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1387–1397, 2004.

[11] E. Ng, A. Fu, and R. Wong, "Projective clustering by histograms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 369–383, 2005.

[12] C. M. Procopiuc, M. Jones, P. K. Agarwal, and T. M. Murali, "A Monte Carlo algorithm for fast projective clustering," in *Proceedings of the ACM SIGMOD International Conference on Managment of Data*, pp. 418–427, June 2002.

[13] H. Wang, W. Wang, J. Yang, and P. S. Yu, "Clustering by pattern similarity in large data sets," in *Proceedings of the ACM SIGMOD International Conference on Managment of Data (ACM SIGMOD '02)*, pp. 394–405, June 2002.

[14] V. S. Tseng and C.-P. Kao, "A novel similarity-based fuzzy clustering algorithm by integrating PCM and mountain method," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 6, pp. 1188–1196, 2007.

[15] J. Venkatesh, K. Sridharan, and S. B. Manooj Kumaar, "Location based services prediction in mobile mining: determining precious information," *European Journal of Social Sciences*, vol. 37, no. 1, pp. 80–92, 2013.

[16] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, pp. 144–155, Santiago, Chile, 1994.

[17] D. Ryabko, "Clustering processes," in *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*, pp. 919–926, Haifa, Israel, June 2010.

[18] R. R. Yager and D. P. Filev, "Approximate clustering via the mountain method," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 8, pp. 1279–1284, 1994.

[19] M. Dorigo and K. Socha, "An introduction to ant colony optimization," in *Approximation Algorithms and Metaheuristics*, T. F. Gonzalez, Ed., CRC Press, 2007.

[20] K. Y. Yip, D. W. Cheung, M. K. Ng, and K. H. Cheung, "Identifying projected clusters from gene expression profiles," *Journal of Biomedical Informatics*, vol. 37, no. 5, pp. 345–357, 2004.