Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 92 (2016) 442 – 449

Procedia Computer Science

2nd International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2016)

Srikanta Patnaik, Editor in Chief

Conference Organized by Interscience Institute of Management and Technology

Bhubaneswar, Odisha, India

Empirical Study on Clustering Based on Modified Teaching Learning Based Optimization

Sushanta Kumar Panigrahi[a], Sabyasachi Pattnaik[b]*

[a] *Interscience Institute of Management & Technology, Bhubaneswar, Odisha, India*
[b] *Fakir Mohan University, Bhubaneswar, Odisha, India*

Abstract

In this Paper the focus is given on data clustering using Modified Teaching–Learning Based Optimization (MTLBO) a hybridization technique of TLBO. Unlike TLBO, this population based method works on the effect of influence of a teacher on learners to find the optimum solution and it has been used for clustering. The motivation behind the data clustering is to find inherent structure in the data items and grouping then on the basis of their mutual similarity. The effectiveness of the method is tested on many benchmark problems with different characteristics and the results are compared with other population based methods and finally it is implemented on clustering using neural network in data mining.

Peer-review under responsibility of the Organizing Committee of ICCC 2016

Keywords: Artificial neural network; evolutionary algorithm; particle Swarm Optimization; teaching learning based optimization; clustering.

* Corresponding author. Tel.: 91-9861261763
*E-mail address:*ctcsushanta@gmail.com

Peer-review under responsibility of the Organizing Committee of ICCC 2016
doi:10.1016/j.procs.2016.07.338

1. Introduction

Data mining is the field of research whose core exists at the intersection of statistics, machine learning, and databases. In data mining several tasks like classification, association rule mining, clustering, regression, summarization etc .are embedded with. Each of these tasks can be viewed as a type of problem to be solved by a data mining technique. In this work the focus is on data clustering.

The motivation behind the data clustering is to find inherent structure (similarity) in the data items and grouping then on the basis of their mutual similarity. A good clustering is one that achieves- High within-cluster similarity and Low inter-cluster similarity [2]. In other words Similarity among the same cluster should be high as compared to the data objects among different clusters [1]. Similarity measurement is a very important concern in data clustering. It is inversely related to distance.

Clustering technique is used to partition unlabeled scattered data set into groups of similar objects known as clusters. Usually the clusters are different from each other. Unsupervised algorithms are mainly known as clustering algorithms. Clustering techniques can be classified into types such as hierarchical and partitional. The hierarchical clustering is classified into agglomerative and divisive. In hierarchical clustering n objects will be grouped into k clusters by minimizing some measure of dissimilarity in each group and maximizing the dissimilarity of different groups [2:9].

In this paper the focus is on partitional clustering, and in particular the K-means algorithm that is one of the most efficient clustering algorithms. However, the K-means algorithm suffers from drawbacks like many local optima, because it is not convex and it heavily depends on the initial solutions [2:4].

Clustering process starts with randomly generated initial centroids and keeps reassigning the data objects various clusters based on the similarity between the data object and the cluster centroids until a termination criteria is met (e.g., the fixed number of iterations or stability in movement of data points among clusters) [4]. K-Means is the most efficient algorithm in terms of the execution time but it has a drawback that the cluster results are extremely sensitive to the selection of the initial cluster centroids and may converge to the local optimal solution [10, 11]. Bad initialization leads to bad clustering and poor convergence speed. Therefore, the initial selection of the cluster centroids decides the main processing of K-Means and the clustering result of the dataset as well. Considering these limitation, it has been proposed to use meta-optimization to improve the processing capabilities of existing clustering algorithms. Meta-optimization is an approach which allows using the combination of two or more than two algorithms to achieve a common goal. In current scenario, it will be good to utilize any global optimal searching algorithm for generating the initial cluster centroids for K-Means [2:4]. Recently many algorithms have been developed based on evolutionary algorithms like Genetic Algorithm (GA), Tabu Search (TS), Particle Swarm Optimization (PSO) and Simulated Annealing (SA) [21:24]. But the disadvantage is that most of these evolutionary algorithms are very slow to get the optimal solution.

This work presents the improvised Teaching–Learning Based Optimization (TLBO) termed as Modified Teaching–Learning Based Optimization (MTLBO**)** a hybridization technique of TLBO with evolutionary system, such as Adaptive Particle Swarm Optimization (APSO) [26, 27]. Unlike TLBO, this population based method works on the effect of influence of a teacher on learners to find the optimum solution and it has been used for clustering.

2. Cluster analysis

Clustering analysis that is an NP-complete problem to find groups in heterogeneous data by minimizing dissimilarity measures is one of the fundamental tools in data mining, machine learning and pattern classification solutions [12:20]. Clustering in N-dimensional Euclidean space R^N is the process of partitioning a given set of n points into K groups (or, clusters) based on some similarity (distance) metric that is Euclidean distance, which derived from the Minkowski metric (equations 1 and 2).

$$d(x, y) = \left(\sum_{i=1}^{m} | x_i - y_j |^r \right)^{1/r} \quad (1)$$

$$d(x, y) = \sqrt{\sum_{i=1}^{m} \left(x_i - y_j \right)^2} \quad (2)$$

Let the set of *n* points *{X1, X2,..., Xn}* be represented by the set S and the *K* clusters be represented by *C1,C2...,CK* Then:

$Ci \neq \phi$ *for i=1 ,...,K,*

$Ci \cap Cj = \phi$ *for i=1 ,...,K , j=1 ,...,K , and* $I \neq j$

and $\bigcup_{i=1}^{k} c_i = s$

In this study the Euclidian metric has been used as a distance. All of clustering algorithms can be classified into two categories: hierarchical clustering and partitional clustering. Partitional clustering methods are the most popular class of center based clustering methods. The K-means algorithms, is one of the most widely used center based clustering algorithms. To find *K* centers, the problem is defined as an optimization performance function (minimization), *Perf (X, C)*, defined on both the data items and the center locations. A popular performance function for measuring goodness of the *K* clustering is the total within-cluster variance or the total mean-square quantization error (MSE), equation 3.

$$Perf(X, C) = \sum_{i=1}^{N} Min \left\{ \| X_i - C_i \|^2 \mid l = 1,...., K \right\} \quad (3)$$

The steps of the K-means algorithm are as follow [7]:

Step 1: Choose *K* cluster centers *C1, C2, ..., Ck* randomly from *n* points *{X1, X2,..., Xn}*.

Step 2: Assign point *Xi, i=1, 2, ..., n* to cluster *Cj*, $J \in \{1, 2, ..., K\}$ if ||Xi-Cj ||< || X *i- C p* || , *p*=1, 2, …, *K*, and $j \neq p$.

Step 3: Compute new cluster centers *C*1* *C*2* ……*CK* *, as follows:

$$C_i^* = \frac{1}{n} \sum_{x_j \in c_i} X_j, \qquad i=1 , ,...,K, \quad (4)$$

where *ni* is the number of elements belonging to cluster *Ci*.

Step 4: If termination criteria satisfied, stop otherwise continues from step 2

Note that in case the process close not terminates at step 4 normally, then it executed for a mutation fixed number of iterations.

3. Teaching Learning Based Optimization

Generally, the process of Teaching Learning Based Optimization (TLBO) is divided into two parts. The first part considered as the 'Teacher Phase' and the second part considered as the 'Learner Phase'. The 'Teacher Phase' means learners learn from the teacher and the 'Learner Phase' means learners learn through the interaction between them [25].

3.1 Teacher Phase

In our society the best learner is treated as teacher, who has the better knowledge than other learners. Teacher tries to disseminate knowledge among students or learners to enhance their knowledge in the classroom, i.e. the mean of a class increases from MA to MB depending upon the ability of a good teacher. Good teacher ability is estimated by how much he can bring his or her learners up to his level in terms of knowledge. But, practically this is not possible and a teacher can only move the mean of a class up to some extent depending on the capability of the class. This follows a random procedure depending on many factors. Let Mi be the mean and Ti be the teacher at any iteration i. Ti always try to move mean Mi towards its own level, so now the new mean will be designated as

Mnew. The solution is updated according to the difference value between the existing and the new mean and is given by the expression,

$$\text{Difference\_Mean}_i = r_i(M_{new} - T_F M_i) \quad (5)$$

where TF is a teaching factor that decides the value of mean to be changed, and ri is a random number in the range [0, 1]. The value of TF can be either 1 or 2, which is again a heuristic step and decided randomly with equal probability as,

$$T_F = \text{round}[1 + \text{rand}(0,1)\{2-1\}] \quad (6)$$

This difference modifies the existing solution to enhance the mean according to the following expression,

$$X_{new,i} = X_{old,i} + \text{Difference\_Mean}_i \quad (7)$$

3.2 Learner Phase

Learners enhance their knowledge by two different means, one through input knowledge from the teacher and the other through interaction of knowledge between themselves. A learner interacts randomly with other learners with the help of presentations, group discussions and formal communications, etc. A learner learns something new if the other learner has more knowledge than him or her. For population size Pn learner phase is expressed as,

```
For i = 1 : Pn
  Randomly select two learners Xi and Xj, where i <> j
  If f (Xi) < f (Xj)
  Xnew,i = Xold,i + ri(Xi − Xj)
  Else
  Xnew,i = Xold,i + ri(Xj − Xi)
  End If
End For
```

Accept Xnew if it gives a better function value.

4. Modified TLBO

In the modified TLBO (MTLBO) the teacher phase is similar to TLBO algorithm. In the learner phase the algorithm is modified. A learner interacts randomly with other learners with the help of group discussions, presentations, formal communications, etc. A learner learns something new if the other learner has more knowledge than him or her and also he or she follows the best learner as team leader [26]. This representation mimics the PSO activities where the particle update its position by following its previous best as well as global best position found by all particle. In TLBO, a learner learns something new if the other learner has more knowledge than him or her can be treated as learner's previous best position. In modified TLBO in addition to previous best the learner also learns from the best learner acting as team leader unlike PSO. Any learner in the Learner phase modification is expressed as,

```
For i = 1 : Pn
  Randomly select two learners Xi and Xj, where i <> j
  If f (Xi) < f (Xj)
  Xnew,i = Xold,i + ri (Xi − Xj) + ri (Xg − Xi)
  Else
  Xnew,i = Xold,i + ri (Xj − Xi) + ri (Xg − Xi)
  End If
End For
```

where Xg is the Knowledge of best learner acting as team leader. Accept Xnew if it gives a better function value.

5. Application of MTLBO algorithm on Clustering

Our proposed hybridization technique follows two phases; first one is Teacher and second one is Learner phase. Although K-Means is a good option (fast, robust and easier to understand) for local search ability but it didn't work well with global clusters [12:20]. Even its performance is un-consistent at different initial partitions, it produce different results at different initial partitions. These considerations were main objective behind this research. At the initial stage, the MTLBO clustering algorithm is executed to search for the location of clusters' centroid. These locations are derived from Euclidean distance measures for refining and generating the optimal clustering solution. This arrangement is not only resolving the limitations of these algorithms but multiplying the advantages of both algorithms as well.

To apply the MTLBO algorithm on clustering the following steps should be repeated;

Step 1: Initializing the problem and algorithm parameters
Step 2: Initialize each learner to contain N, randomly selected cluster.
Step 3: Compute the objective function using Euclidean distance Eq. (1&2).
Step 4: Compute the fitness (MSE) using Eq. (3) of the population.
Step 5: Determine the best solution of Teacher using Eq. (5).
Step 6: Modify solutions based on the teacher knowledge according to teacher phase using Eq. (6).
Step 7: Update solutions according to learner phase using Eq. (7).and objective function.
Step 8: Go to Step 4 & compute fitness until the maximum iteration number arrives.

5.1 Experimental Studies

The performance and efficiency of the MTLBO model on clustering is evaluated using two real-life application data sets and compared with the TLBO, PSO and K-means algorithms. In stochastic algorithms the effectiveness highly depends on the initial solutions. To overcome these drawback each algorithms performed 100 times individually with randomly generated initial solutions.

5.2 Real Life Datasets Used

IRIS Data (N=150, d=4, K=3): This dataset has 150 points which are random samples of three plant species (length and thickness of its petal and sepal) divided into three distinct classes (Iris Setosa, Iris Versicolor and Iris Virginica). For each clusters we have 50 samples with 4 dimensions.

WINE Data (N=178, d=13, K=3): The wine dataset resulting from chemical analyses performed on three types of wine produced in Italy from grapevines cultivated by different owners in one specific region. The dataset has 178 points with 13 continues attributes.

5.3 Experimental Results & Comparative Studies

The efficiency of the proposed algorithm has been compared with other algorithms by applying them on above datasets. The best solution of 100 runs of each algorithm, number of function evaluation and standard deviation of solutions obtained by applying algorithms on the datasets has been used for comparison. The quality of solution is considered based on the average and worst values of the clustering metric (F_{avg} and F_{worst}). F is the performance of clustering algorithms that has been shown in equation 3. Tables 1 and 2 present a comparison among the results of algorithms.

Table 1 shows the result of algorithms on the iris dataset. MTLBO converges to the global optimum of 96.4687, while the best solutions of TLBO, PSO and K-means are 96.6500, 96.8942 and 97.333 respectively. The standard deviation of the fitness function for this algorithm is 0, which it significantly is smaller than other methods.

Table 1.Result obtained by various algorithms for 100 different runs on Iris data.

Method	Function Value			Standard Deviation	Number of Function Evaluations
	F_{best}	F_{avg}	F_{worst}		
MTLBO	96.4687	96.4552	96.4817	0	2432
TLBO	96.6500	96.6500	96.6500	0	2468
PSO	96.8942	97.2328	97.8973	0.347168	4953
K-means	97.333	106.05	120.45	14.6311	120

Table 2 shows the result of algorithms on the wine dataset. The optimum value is 16188.467 which is obtained in 88% runs of MTLBO algorithm. Noticeably other algorithms fail, except TLBO to attain this value even once within 100 runs. It is found that the MTLBO clustering algorithm is able to provide the same partition of the data points in most of runs. As earlier, the results of the other algorithms are in inferior to that of ours. MTLBO as empirically established proves to be a better clustering method.

Table 2.Result obtained by various algorithms for 100 different runs on Wine data.

Method	Function Value			Standard Deviation	Number of Function Evaluations
	F_{best}	F_{avg}	F_{worst}		
MTLBO	16188.467	16225.453	16265.395	26.461	6228
TLBO	16295.31	16323.17	16345.26	26.824	6316
PSO	16345.9670	16417.4725	16562.3180	85.4974	16532
K-means	16555.68	18061	18563.12	793.213	390

In terms of the number of function evaluations, K-means needs the least number of function evaluations, but the results are less than satisfactory. For the iris dataset, the number of function evaluations of MTLBO, TLBO, PSO and K-means are 2432, 2468, 4953 and 120 respectively. For the wine dataset, the number of function evaluations of MTLBO, TLBO, PSO and K-means are 6228, 6316, 16352 and 390 respectively. These results show that the number of function evaluations of MTLBO is less than those of other evolutionary algorithms. Based on the obtained simulation results, we can conclude that the changes of the number of fitness function evaluations of the proposed algorithm are less than other evolutionary algorithms for all cases.

5.4 Comparison of Time Complexity of TLBO and MTLBO

Keeping no. of data points constant, lets assume n=150, d=2, i=10 and varying no. of clusters, we obtain the following table and graph of time complexity. Where n = number of data point, c = number of cluster, d= dimension, i = number of iteration.

Table 3. Time complexity when number of cluster varying.

Sl.No.	Number of cluster	TLBO time complexity	MTLBO time complexity
1	1	3000	3000
2	2	12000	6000
3	3	27000	9000
4	4	48000	12000

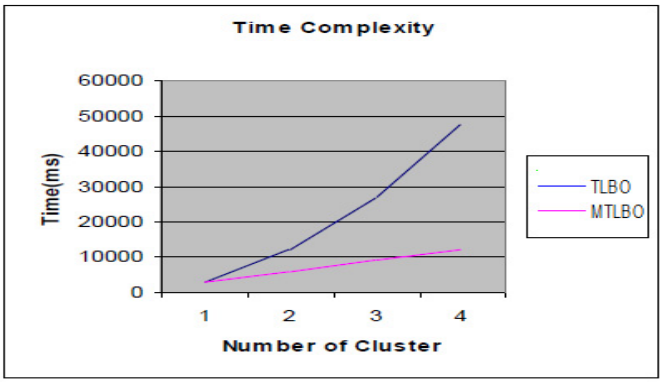

Fig. 1. Time complexity of TLBO and MTLBO by varying no. of clusters.

5.5 Comparison of Space Complexity of TLBO and MTLBO

Keeping no. of data points constant, lets assume n=150, d=2, i=10 and varying no. of clusters, we obtain the following table and graph of space complexity. Where n = number of data point, c = number of cluster, d= dimension, i = number of iteration.

Table 4. Space complexity when number of cluster varying.

Sl.No.	Number of cluster	TLBO space complexity	MTLBO space complexity
1	5	400	2
2	10	600	4
3	15	750	6
4	20	900	8

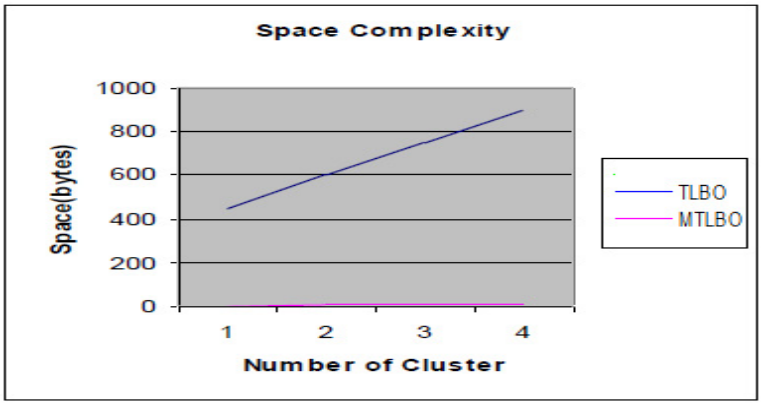

Fig. 2. Space complexity of TLBO and MTLBO by varying no. of clusters.

The above simulation results in the tables demonstrate that the proposed evolutionary algorithm converges to global optimum with a smaller standard deviation and less function evaluations, less time and space complexity, leads naturally to the conclusion that the MTLBO algorithm is a viable and robust technique for data clustering.

6. Conclusion

The clustering analysis is a very important technique and has attracted much attention of many researchers in different areas. The K-means algorithm one of the most efficient clustering method and is very simple that has been applied to many engineering problems. In this work modification of TLBO algorithm MTLBO has applied for solving the clustering problem. The effectiveness of MTLBO method is evaluated using two real life benchmark databases for clustering performance in terms of $F_{best,}$ standard deviation, function evaluation etc. The proposed algorithm has been implemented and tested on well known real life datasets. The result illustrate that the proposed MTLBO optimization algorithm can be considered as an efficient heuristic method to find optimal solutions for clustering problems of allocating N objects to k clusters. The experimental results indicate that the proposed optimization algorithm is at least comparable to the other algorithms in terms of function evaluations, standard deviations and time & space complexity.

References

1. Pavel Berkhin, A Survey of Clustering Data Mining Techniques, pp.25-71, 2002.
2. Jain, M. Murty and P. Flynn. Data Clustering: A Review. ACM Computing Surveys, vol. 31, no. 3, pp. 264-323,1999
3. Jain, R. Duin and J. Mao. Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.1, pp. 4-37, 2000.
4. G. Hamerly and C. Elkan. Learning the K in K-means. In The Seventh Annual Conference on Neural Information Processing Systems, 2003.
5. Jain and R. Dubes. Algorithms for Clustering Data. Prentice Hall, New Jersey, USA, 1988.
6. S. Baek, B. Jeon, D. Lee and K. Sung. Fast Clustering Algorithm for Vector Quantization. Electronics Letters, vol. 34, no. 2, pp. 151-152, 1998.
7. Z. Xiang. Color Image Quantization by Minimizing the Maximum Inter-cluster Distance. ACM Transactions on Graphics, vol. 16, no. 3, pp. 260-276, 1997.
8. D. Judd, P. Mckinley and A. Jain. Large-scale Parallel Data Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 871- 876, 1998.
9. Lee and E. Antonsson. Dynamic Partitional Clustering Using Evolution Strategies. In The Third Asia-Pacific Conference on Simulated Evolution and Learning, 2000.
10. J. Han , M. Kamber, Data Mining, Morgan Kaufmann Publishers, 2001.
11. Arun K. Pujari,Data mining techniques-a reference book ,pg. no.-114-147.
12. K. Jain, "Data Clustering: 50 Years Beyond K-Means, in Pattern Recognition Letters, vol. 31 (8), pp. 651-666, 2010.
13. Rama, P. Jayashree, S. Jiwani, " A Survey on clustering Current status and challenging issues", International Journal of Computer Science and Engineering, vol. 2, pp. 2976-2980.
14. K. Ravichandra Rao, "Data Mining and Clustering Techniques", DRTC Workshop on Semantic Web, Bangalore, 2003.
15. Rui Xu, Donald C. Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions on neural Networks, vol. 16, pp. 645-678, May 2005.
16. S.B. Kotsiantis, P. E. Pintelas, "Recent Advances in Clustering: A Brief Survey" WSEAS Transactions on Information Science and Applications, Vol. 1, No. 1, pp. 73–81, Citeseer, 2004.
17. J. Kelinberg, "An impossibility theorem for clustering", in NIPS 15, MIT Press,2002, pp. 446-453.
18. B.Amiri, M.Fathian. "Integration of self organization feature maps and honey bee mating optimization algorithm for market segmentation", JATIT, Vol. 3, No. 3, Pages 70-86, July, 2007.
19. T. Kohonen, "Self-Organization and Associative Memory", Springer-Verlag, New York, vol 10, Page 811-821, 1988.
20. K. M. Faraoun, A. Boukelif, "Neural networks learning improvement using the K-means clustering algorithm to detect network intrusions", IJCI, Page 161-168, 2006.
21. E. Alba and J.F. Chicano, "Training Neural Networks with GA Hybrid Algorithms", K. Deb(ed.), Proceedings of GECCO'04, Seattle, Washington, LNCS 3102, pp.852-863, 2004.
22. J. Kennedy and R. C. Eberhart, "Particle swarm optimization," Proceedings of the IEEE International Conference on Neural Networks, vol. IV, pp. 1942–1948, 1995
23. R.S. Sexton, B. Alidaee, R.E. Dorsey and J.D. Johnson,"Global optimization for artificial neural networks: a tabu search application", European Journal of Operational Research, (106)2-3, pp.570-584,1998.
24. N.K. Treadgold and T.D. Gedeon, "Simulated annealing and weight decay in adaptive learning: the SARPROP algorithm", IEEE Transactions on Neural Networks, 9:662-668,1998.
25. Rao R.V., Savsani V.J., Vakharia D.P., "Teaching–Learning-Based Optimization: An optimization method for continuous non-linear large scale problems", Information Sciences, 183 (2012) 1–15
26. A Sahu, S. Panigrahi, S. Pattnaik "An Empirical Study on Classification Using Modified Teaching Learning Based Optimization", International Journal of Computer Science and Network, V-2, I-2, 2013.
27. S. Panigrahi, A Sahu, S. Pattnaik "Neuro Structure Optimization Using Adaptive Particle Swarm Optimization ", Elsevier Procedia Computer Science 48 (2015) 802 – 808.