High Dimensional Data Clustering Using Hybridized Teaching-Learning-Based Optimization

TUSHARKANTA SAMAL, ARCHANA BARAL and H. S. BEHERA

Department of Computer Science and Engineering, Veer Surendra Sai University of Technology (VSSUT) Burla, Sambalpur, Odisha, INDIA.

(Received on: May 16, 2013)

ABSTRACT

TLBO is a newly introduced population based optimization algorithm inspired by passing on knowledge within a classroom environment. The most prominent feature of TLBO is that it does not require the algorithm parameters to be tuned. This paper proposes a Hybridized Teaching-Learning-Based Optimization (HTLBO) Technique in order to abstract/gain useful information from high dimensional data. The new HTLBO algorithm is evaluated on some high dimensional datasets and compared to the performance of Hybridized K-means Clustering Algorithm and Genetic Algorithm.

Keywords: Clustering, K-means Algorithm, Teaching-Learning-Based Optimization (TLBO), Dimensionality Reduction, Principal Component Analysis, Hybridized-TLBO.

I. INTRODUCTION

Clustering\(^1\) is an unsupervised classification technique in which patterns or feature vectors or data items are organized in clusters or groups based on similarity, such that the data within the same cluster have high degree of similarity and data belonging to different clusters have high degree of dissimilarity. Clustering techniques can be classified into two types: hierarchical and partitional. Hierarchical clustering generates a hierarchy of partitions in which each partition is nested within the partition at the next level in the hierarchy. On the other hand, partitional clustering generates a single partition. This paper focus on K-means clustering algorithm\(^2,3\) which is a well-known partitioning based, non-hierarchical clustering technique. It is well-known because of its simplicity in which the given data is partitioned into a specified number of clusters by minimizing the within cluster variation. However, k-means is
iterative and hill climbing and may converge to a local minima and it is sensitive to the selection of initial cluster centroids.

Teaching-Learning-Based Optimization (TLBO) is a new evolutionary optimization algorithm, which can be used to find global solutions for clustering. TLBO does not require the algorithm parameters to be changed for better performance. It consists of two phases: ‘Teacher Phase’ and ‘Learner Phase’. The ‘Teacher Phase’ means learning from the teacher and the ‘Learner Phase’ means learning through the interaction between the learners.

One of the major problems in clustering is the ‘curse of dimensionality’. One solution is to preprocess the data using some dimension reduction technique.

This paper proposes Hybridized Teaching-Learning-Based Optimization. First, one of the dimension reduction techniques such as the PCA is applied to the data. Using the principal components, data can be mapped to a lower dimensional subspace. Then clustering algorithm can be applied to the subspace.

This paper is organized as follows: Section 2 describes Dimension Reduction Technique, Section 3 describes the K-means algorithm, Section 4 describes Genetic Algorithm, Section 5 describes the Teaching-learning-based optimization Technique, Section 6 contains the proposed algorithm, Section 7 demonstrates the experimental results and Section 8 contains the conclusion.

II. DIMENSIONALITY REDUCTION

Increase in data and their storage capabilities have led to information overload. In our day-to-day life we face high-dimensional data (dimension of the data is the number of variables or features or attributes associated with it) which leads to a number of problems, one of them is “all the variables are not required” to understand the underlying phenomena of data. So in order to handle data adequately many applications need to reduce the dimension of the data without losing the original content of the data. In mathematical terms, if we are given with a n-dimensional random variable $X=(x_1, x_2... x_n)$ then we have to find a lower dimensional representation as $X_{\text{new}} = (x_1, x_2... x_{n_0})$, such that, $n_0 \leq n$.

There are two major types of dimension reduction techniques: Linear and Non-Linear. Here we will consider a Linear dimension reduction technique, i.e., PCA (Principal Component Analysis).

A. Principal component analysis

It is a widely used statistical technique for unsupervised dimension reduction by examining the interrelations among a set of variables. It is based on covariance matrix of the variables. PCA (also called Karhunen-Loeve or K-L method) reduce the dimension of the data by finding some orthogonal linear combinations (PCs) of the original variables with largest variance.

Procedure
1. The input data are normalized.
2. PCA computes the orthonormal vectors (PCs) that provide a basis for the normalized input data.
3. The PCs are sorted in decreasing order of “significance” or “strength”.
4. The size of the data can be reduced by
eliminating the weaker components that is those with low variance.
PCs may be used as inputs to cluster analysis.

III. K-MEANS CLUSTERING ALGORITHM

K-Means algorithm uses K number of clusters to characterize the data. They are determined by minimizing the sum of squared errors:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \| X_n - M_k \|^2$$  \hspace{1cm} (1)

Where X is the data matrix, M is the centroid of the cluster, n is the number of data points, k is the number of clusters, \| \| is the Euclidean distance.

**Procedure**
1. Initialize: Choose K centers (data points) at random from X, say this matrix C.
2. Search for nearest neighbor: Each point in X is assigned or mapped to its closest center in C, which results in K number of clusters.
3. Update centroid: For each cluster, compute the centroid /center /median. This is that point in the cluster such that the distances to the remaining point in the cluster is as small as possible. This results in C with K new centers.
4. Stopping condition: Repeat step 2 and 3, until the matrix C changes or by minimizing equation (1).

In Hybridized K-means (HKM) clustering algorithm\textsuperscript{10}, first PCA is applied to map the data to a lower dimensional subspace. Then the K-means clustering algorithm can be applied to the subspace to find the clusters.

IV. GENETIC ALGORITHM

Genetic algorithms (GAs) are randomized search and optimization techniques guided by the principles of evolution and natural genetics\textsuperscript{12}. They have a large amount of implicit parallelism and can perform search in complex, large and multimodal landscapes, and provide near-optimal solutions for objective or fitness function of an optimization problem. In GAs, the parameters of the search space are encoded in the form of strings or chromosomes, collection of such strings are called Population. The fitness function which represents the degree of goodness of the string is associated with it. Based on the principle of survival of the fittest, a few of the strings are selected and then some biologically inspired operators like crossover and mutation are applied to those strings to yield a new generation of strings. The process of selection, crossover and mutation continues for a fixed number of generation or until a termination condition is satisfied\textsuperscript{4,6}.

**Procedure**
1. Initialize each chromosome to contain K randomly chosen points from the dataset.
2. Calculate fitness function for each chromosome.
3. Perform Crossover and Mutation.
4. Select fittest chromosomes from current generation to next generation.
5. Repeat step 3 and 4, until maximum
generation is reached or some termination criteria is satisfied.

6. Select the best chromosome in the population which will act as the resultant clusters.

In Hybridized GA (HGA) clustering algorithm, first PCA is applied to map the data to a lower dimensional subspace. Then the GA clustering algorithm can be applied to the subspace to find the clusters.

V. TEACHING-LEARNING-BASED OPTIMIZATION

This optimization technique works on the principle of teaching and learning mechanism in which a teacher is generally considered as a highly learned person who shares knowledge with the learners and moreover, learners can also learn from interaction between themselves. Like all other nature-inspired population based optimization methods, TLBO is also a population-based method that uses a population of solutions (P) with a pre-defined size of the population (popsize) to proceed to the global solution. Each learner (L_i) is considered as a single possible solution in the population. The number of subjects (d) refers to the design variables or dimension of the problem. The learners result in TLBO corresponds to the fitness of the solution.

TLBO consist of two phases: 'Teacher Phase' and 'Learner Phase'. The 'Teacher Phase' means learning from the teacher and the 'Learner Phase' means learning through the interaction between the learners.

A. Teacher phase

During Teacher phase, the best learner (L_{teacher}) is considered as a teacher. The teacher always circulate knowledge among learners (L_i), which will in turn increase the knowledge level of the class and help learners to get good marks or grades. A good teacher is one who brings his or her learners up to his or her own level in terms of knowledge. But in practice this is not possible, so the teacher increases the mean value (L_{mean}) of the class up to some extent depending on the capability of the class. This follows a random process depending on many factors as shown below:

Let teacher L_{teacher} will try to move mean L_{mean} towards its own level, so now the new mean will be L_{mean}_{new}. The solution is updated according to the difference between the existing and the new mean given by

$$\text{Difference mean} = r \left( L_{\text{teacher}} - T_F \cdot L_{\text{mean}} \right)$$  \hspace{1cm} (2)

Where \( T_F \) is a teaching factor that decides the value of mean to be changed and its value can be either 1 or 2 which is again a heuristic step and decided randomly with equal probability as \( T_F = \text{round}(1+\text{rand}(0,1)) \), and \( r \) is a random number in the range [0,1]. The difference modifies the existing solution according to the following expression,

$$L_{i,\text{new}} = L_i + \text{Difference}_\text{mean}$$  \hspace{1cm} (3)

Accept \( L_{i,\text{new}} \) if it gives a better solution.

B. Learner phase

Learners increase their knowledge by two different means: one through input
from the teacher and the other through interaction between themselves. A learner (L$_i$) interacts randomly with other learners (L$_j$) 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. If L$_i$ is better than L$_j$, then L$_j$ will be moved towards L$_i$ as in equation (4). Otherwise, L$_i$ will move towards L$_j$ as in equation (5).

$$L_{i,new} = L_i + r (L_i - L_j)$$ \hspace{1cm} (4)

$$L_{j,new} = L_j + r (L_j - L_i)$$ \hspace{1cm} (5)

Accept $L_{i,new}$ if it gives a better solution.

The algorithm continues until reaching the maximum number of generations.

VI. PROPOSED ALGORITHM

This paper proposes Hybridized Teaching-Learning-Based Optimization (HTLBO). First, one of the dimension reduction techniques such as the PCA is applied to the data. Using the principal components, data can be mapped to a lower dimensional subspace. Then clustering algorithm can be applied to the subspace.

The Hybridized Teaching-Learning-Based Optimization (HTLBO) consist of three stages. In the first stage, PCA is applied to the dataset. In the second stage, TLBO algorithm is applied to find the initial cluster centroids and in the third stage, K-means clustering algorithm is used to find the clusters. The proposed model is shown in Fig.1.

Algorithm

Input: $X = \{x_1, x_2, x_3, \ldots, x_n\}$ \hspace{1cm} // $x_i$ is the data item and $n$ is the number of data items.

Set the value of $K$ \hspace{1cm} // $K$ is the number of clusters.

$\text{popsize} = 100$

Output: Set of $K$ clusters.

Halting criteria: In this paper, the Teacher and Learner Phase are executed until $max\_gen$ (maximum number of generations is reached) is reached or 98% of the population are same.

// Stage-1: Apply PCA to reduce the dimension of the dataset
1. Load the dataset in $X$.
2. Normalize the dataset using min-max.
3. Find the Covariance matrix.
4. Find the Eigen vector.
5. Find the mean of Eigen vector.
6. Find the Index of Eigen vector greater than the mean of Eigen vector.
7. Find the column of index represented in step(6), which will act as the principal components.
8. Replace $X$ with the principal components.

// Stage 2: Apply TLBO to find initial cluster centers
9. Initialize '$P$' randomly from the dataset.
10. Evaluate the population $P$ / Calculate the fitness function of the population.

11. while Halting criteria not satisfied do
12. for $i=1$ to $\text{popsize}$ do

Figure 1: Hybridized TLBO
13. Apply Teacher modification to the solutions as in equation (3).
14. Evaluate the modified solution and if it is better, then replace the existing solution with the modified one immediately.
15. Apply Learner modification to the solutions as in equation (4) and (5).
16. Evaluate the modified solution and if it is better, then replace the existing solution with the modified one immediately.
17. end for
18. endwhile
19. Select the best chromosome in the population which will act as the initial clusters centroids.

// Stage 3: Apply K-means Clustering Algorithm to find clusters of the dataset
20. Calculate the Euclidean distance and assign data points to its nearest cluster.
21. Calculate the mean of data points in a cluster which is the new cluster centroid.
22. Repeat step 20 and 21 until there is no change in the cluster centroid.

VII. EXPERIMENTAL ANALYSIS

The efficiency of the HTLBO algorithm on clustering high dimensional datasets has been tested on several well known real world high dimensional datasets and two artificial datasets and compared with Hybridized K-means and GA clustering algorithms. The information of datasets are given below where ‘n’ represents the number of data points in dataset, ‘K’ represents the number of cluster for the dataset and ‘d’ represents the dimension of dataset or number of attributes in dataset. The artificial datasets are:

Data1 (n=100, d=10, K=2): Designed randomly by taking uniform distribution between 0-2 for 50 points and 1-4 for next 50 points.

Data2 (n=400, d=20, K=2): Designed randomly by taking uniform distribution between 0-5 for 150 points and 6-11 for next 150 points.

The real datasets taken from UCI machine repository are:
Lung cancer data (n=32, d=56, K=3): This dataset has 32 points in which all predictive attributes are nominal, and takes integer value 0,1,2 or 3.
Wine data (n=178, d=13, K=3): This dataset is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The dataset has 178 points with 13 attributes.
Iris data (n=150, d=4, K=3): This dataset has 150 points which are random samples of three species of the iris flower such as: setosa, versicolor, and virginica.
Glass data (n=214, d=10, K=3): This dataset has 214 points used to know whether the glass was a type of "float" glass or not. The study of classification of types of glass was motivated by criminological investigation.

Concepts Used

- **Fitness Function**- It is sum of the distance of each point to its respective cluster centers, termed as Clustering Metric (M) [5] and aim to minimize ‘M’.
- **Dimension of population (N)**- N=Number of clusters (K) * Dimension of each cluster centroid.
Flowchart for proposed algorithm:

Start
- Take input data X and number of clusters K.
- Normalize the data using min-max Normalization.
- Apply the steps of PCA to the input data.
- Initialize the population randomly from the input data.
- Evaluate the population.
- Compute the mean of the population.
- Identify the best solution (teacher).
- Find new solution based on equation (3).
- Is new solution better than existing?
  - Yes
    - Randomly select $L_i$ and $L_j$ and find out which one is better.
  - No
    - Replace the existing solution with the new solution.
- Convergence occur?
  - Yes
    - Find the best solution which is the initial centroid.
    - Assign data points to its nearest cluster.
    - Find new centroids.
    - Centroid changes?
      - Yes
        - Clusters found
      - No
        - Replace the existing solution with the new solution.
  - No
    - Reject new solution.
- Find new solution based on equation (4) and (5).
- Is new solution better than existing?
  - Yes
    - Replace the existing solution with the new solution.
  - No
    - Reject new solution.

The following parameters are used for comparison:
- the clustering metric [8] ‘M’ for K clusters $C_1, C_2, \ldots, C_K$ is given by
  \[ M(C_1, C_2, \ldots, C_K) = \sum_{i=1}^{K} \sum_{z_j \in C_i} ||x_j - z_i|| \]
  where the objective is to minimize ‘M’.
- the intra-cluster distances, i.e. mean of maximum distance between two data vectors within a cluster of clusters i.e.
  \[ \frac{1}{N_c} \sum_{c=1}^{N_c} [\max_{z_p, z_q \in C} d(Z_p, Z_q)] \]
  where the objective is to minimize the intra-cluster distances.
• the inter-cluster distances, i.e. minimum distance between the centroids of the clusters, where the objective is to maximize the distance between clusters.
• NOI, i.e. average Number of Iterations required to find the optimal solution.

For all the Clustering algorithm, we randomly initialize cluster centroids. The cluster centroids are randomly fixed from the data vectors in the dataset.

Since the algorithms are stochastic in nature, the results of two successive runs usually do not match. Hence, 10 independent runs (having random initial generation) of each algorithm are taken. All the experiment codes are implemented in MATLAB. The experiments are conducted on a core i3, 3GB memory laptop in Windows 7 environment.

Table I: Comparison in terms of Clustering Metric for GA and TLBO.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Algorithms</th>
<th>Clustering Metric</th>
<th>NOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUNG CANCER</td>
<td>GA</td>
<td>127.2208</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>TLBO</td>
<td><strong>126.6174</strong></td>
<td>98</td>
</tr>
<tr>
<td>WINE</td>
<td>GA</td>
<td>1.69E+04</td>
<td>910</td>
</tr>
<tr>
<td></td>
<td>TLBO</td>
<td><strong>1.63E+04</strong></td>
<td><strong>871</strong></td>
</tr>
<tr>
<td>IRIS</td>
<td>GA</td>
<td>97.3462</td>
<td>317</td>
</tr>
<tr>
<td></td>
<td>TLBO</td>
<td><strong>97.3377</strong></td>
<td><strong>97</strong></td>
</tr>
<tr>
<td>GLASS</td>
<td>GA</td>
<td>3.89E+03</td>
<td>545</td>
</tr>
<tr>
<td></td>
<td>TLBO</td>
<td>3.89E+03</td>
<td>545</td>
</tr>
<tr>
<td>DATA1</td>
<td>GA</td>
<td>231.22</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>TLBO</td>
<td><strong>219.2048</strong></td>
<td>98</td>
</tr>
<tr>
<td>DATA2</td>
<td>GA</td>
<td>1.99E+03</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>TLBO</td>
<td><strong>1.93E+03</strong></td>
<td>98</td>
</tr>
</tbody>
</table>

Table I shows the comparison between GA and TLBO in terms of Clustering metric and NOI, with SD (Standard Deviation) to show the variance from the mean. For all the datasets, except for glass dataset results show that TLBO is efficient in finding optimal solutions in comparison to GA. For glass dataset, both are comparable to each other.

Table II shows the comparison between HKM, HGA and HTLBO in terms of Clustering metric, Intra-Cluster Distance and Inter-Cluster Distance with SD. For all the datasets, results show that HTLBO is efficient in finding optimal solutions in comparison to HGA and HKM. Fig. 2 shows the effect of PCA on the attributes of the datasets and Fig. 3 gives the convergence behaviour of different clustering algorithms for lungcancer dataset and reflects the superiority of HTLBO over other algorithms. From Fig. 4 and Fig. 5, this can be concluded that it is not always possible to visualize a high-dimensional data. So PCA can be used to reduce the attributes and view it in 2-D or 3-D plane.
Table II: Comparison in terms of Clustering Metric for Hybridized K-means, GA and TLBO.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Algorithms</th>
<th>Clustering Metric</th>
<th>Intra-Cluster Distance</th>
<th>Inter-Cluster Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>LUNG CANCER</td>
<td>HKM</td>
<td>4.2226</td>
<td>0.4042</td>
<td>0.3699</td>
</tr>
<tr>
<td></td>
<td>HGA</td>
<td>4.0899</td>
<td>0.0255</td>
<td>0.3594</td>
</tr>
<tr>
<td></td>
<td>HTLBO</td>
<td><strong>4.0514</strong></td>
<td><strong>0.0047</strong></td>
<td><strong>0.3504</strong></td>
</tr>
<tr>
<td>WINE</td>
<td>HKM</td>
<td>29.0868</td>
<td>0.9382</td>
<td>0.5281</td>
</tr>
<tr>
<td></td>
<td>HGA</td>
<td>28.1855</td>
<td>0.0208</td>
<td>0.5209</td>
</tr>
<tr>
<td></td>
<td>HTLBO</td>
<td><strong>28.0904</strong></td>
<td>0</td>
<td>0.5223</td>
</tr>
<tr>
<td>IRIS</td>
<td>HKM</td>
<td>9.0829</td>
<td>0.9079</td>
<td>0.2226</td>
</tr>
<tr>
<td></td>
<td>HGA</td>
<td>8.5763</td>
<td>0.0145</td>
<td>0.2386</td>
</tr>
<tr>
<td></td>
<td>HTLBO</td>
<td><strong>8.5188</strong></td>
<td><strong>0.0403</strong></td>
<td><strong>0.2190</strong></td>
</tr>
<tr>
<td>GLASS</td>
<td>HKM</td>
<td>29.6769</td>
<td>0.3543</td>
<td>0.6920</td>
</tr>
<tr>
<td></td>
<td>HGA</td>
<td>28.7749</td>
<td>0.1214</td>
<td>0.6907</td>
</tr>
<tr>
<td></td>
<td>HTLBO</td>
<td><strong>28.5354</strong></td>
<td><strong>0.2817</strong></td>
<td><strong>0.6634</strong></td>
</tr>
<tr>
<td>DATA1</td>
<td>HKM</td>
<td>7.4685</td>
<td>0.0098</td>
<td>0.2512</td>
</tr>
<tr>
<td></td>
<td>HGA</td>
<td>7.4544</td>
<td>0.1224</td>
<td>0.2500</td>
</tr>
<tr>
<td></td>
<td>HTLBO</td>
<td><strong>7.4387</strong></td>
<td><strong>0.2817</strong></td>
<td><strong>0.6634</strong></td>
</tr>
<tr>
<td>DATA2</td>
<td>HKM</td>
<td>25.9007</td>
<td>0.0997</td>
<td>0.3475</td>
</tr>
<tr>
<td></td>
<td>HGA</td>
<td>25.8998</td>
<td>0.0200</td>
<td>0.3475</td>
</tr>
<tr>
<td></td>
<td>HTLBO</td>
<td><strong>25.7186</strong></td>
<td><strong>0.0200</strong></td>
<td><strong>0.3475</strong></td>
</tr>
</tbody>
</table>

Figure 2: Plotting of original and reduced number of attributes for 1.lungcancer, 2.wine, 3.iris, 4.glass, 5.data1, and 6.data2 datasets.
Figure 3: Convergence for Lungcancer dataset.

Figure 4: Clusters formed with lungcancer dataset (original) by TLBO.

Figure 5: Clusters formed with reduced lungcancer dataset by HTLBO.
VIII. CONCLUSION

This paper represents some of the problems related to K-means clustering algorithm. It illustrates that TLBO can be used to find global solutions for clustering efficiently in comparison to GA. The proposed Hybridized Teaching-Learning-Based Optimization (HTLBO) can overcome some of the problems related to K-means algorithm such as, sensitive to the selection of the initial centroids, convergence to a local optimal solution and the ‘curse of dimensionality’.

REFERENCES