A New Clustering Algorithm with the Convergence Proof

Hamid Parvin, Behrouz Minaei-Bidgoli, and Hosein Alizadeh

School of Computer Engineering,
Iran University of Science and Technology (IUST),
Tehran, Iran
{parvin,b_minaei,halizadeh}@iust.ac.ir

Abstract. Conventional clustering algorithms employ a set of features; each feature participates in the clustering procedure equivalently. Recently this problem is dealt with by Locally Adaptive Clustering, LAC. However, like its traditional competitors the LAC method suffers from inefficiency in data with unbalanced clusters. In this paper a novel method is proposed which deals with the problem while it preserves LAC privilege. While LAC forces the sum of weights of the clusters to be equal, our method let them be unequal. This makes our method more flexible to conquer over falling at the local optimums. It also let the cluster centers to be more efficiently located in fitter places than its rivals.

Keywords: Subspace clustering, Weighted Clusters, Features Weighting.

1 Introduction

Data clustering or unsupervised learning is an essential and also an ill-posed, NP-hard problem. The objective of clustering is to group a set of unlabeled objects into homogeneous groups or clusters which are overall called a data partition or partitioning [6-7] and [9]. Clustering methods are applied to a set of patterns to partition the set into some clusters such that all patterns within a cluster are similar to each other and different from the members of other clusters. In other words, it is intended to have minimum inter-cluster variances and maximum between-cluster variances [14].

The partitional clustering algorithms such as k-means and k-medoid use exactly one point as the representative of each cluster. They partition the set of input patterns into non-overlapping clusters. They consider one point per cluster and iteratively update the representative of the cluster by placing it at the average (medoid) of the cluster. There is a proof for their convergence to a local minimum. The most well-known partitional clustering algorithm, k-means, is discussed by Jain and Dubes in detailed theoretically [9]. The term "k-means" was first used by James MacQueen [12].

The curse of dimensionality is an ever challenge in all supervised and also unsupervised learning methods. In high dimensional spaces, it is highly likely that, for any given pair of points within the same cluster, there exist at least a few dimensions on which the points are far apart from each other. As a consequence, distance functions that equally use all input features may not be effective. Furthermore, it is very likely that each cluster in a real dataset is correlated with a subset of features meanwhile some others are correlated with other subsets of features. There are three ways to deal

A. König et al. (Eds.): KES 2011, Part I, LNAI 6881, pp. 21–31, 2011.

with the curse of dimensionality. One: it could be addressed by requiring the user to specify a subspace [16]. Two: Another way is to reduce the dimensionality of the input space by feature reduction algorithms like PCA. Three: and finally, it can be dealt with by LAC (Locally Adaptive Clustering) algorithm. The first is error-prone, while the second is not fit in some cases and also not feasible in some other cases [4-5]. The LAC algorithm firstly developed by Domeniconi, has been shown that is suitable to solve the problem to a great extent.

Although the third method conquers the problem of imbalanced feature-variance of each cluster, it is not capable of solving the imbalanced inter-cluster variances. Two other methods also are not capable of solving it. It is very common to face a dataset including some dense clusters as well as some sparse ones. However, all the previous methods are not capable of managing the problem. While if the clustering mechanism can take into account the densities of clusters simultaneously with the importance of each feature for them, then the mechanism can overcome the problem. In this paper a novel clustering algorithm is proposed which assigns an importance weight to each cluster as well as a weight vector to all features per each cluster. Each feature along which a cluster is loosely correlated receives a small weight; consequently, the feature becomes less participant in distance function for that cluster than the others. In contrary, each feature along which data are strongly correlated receives a large weight which results in participating in distance function with more effect. Our method benefits from LAC and also uses the cluster importance concurrently.

Our contributions are four-fold.

1. We propose a novel clustering error criterion which takes into consideration the inter-cluster variances.
2. We propose a novel clustering algorithm using the proposed criterion.
3. We prove the convergence of the algorithm.
4. We support our method with some results over a number of real datasets.

2 Related Work

Subspace clustering algorithms are considered as extensions of feature selection methods. They attempt to find clusters in different subspaces of the same dataset. Like feature selection methods, subspace clustering algorithms require a search method and an evaluation criteria. Indeed, the subspace identification problems are related to the problem of finding quantitative association rules that also identify interesting regions of various attributes [13] and [17]. If we consider only subsets of the features, i.e., subspaces of the data, then the clusters that we find can be quite different from one subspace, i.e., one set of features, to another. In addition, the subspace clustering must somehow limit the scope of the evaluation criteria so as to consider different subspaces for each different cluster [16].

Subspace clustering algorithms are divided into two categories according to their searching approaches for subspaces. A brute force approach might be to search through all possible subspaces and apply cluster validation techniques to determine the subspaces with the best clusters [9]. This is not feasible because the subset generation problem is intractable [1] and [10]. Another choice of search technique is heuristic based approaches. This categorization is demonstrated hierarchically in Fig 1 [16].

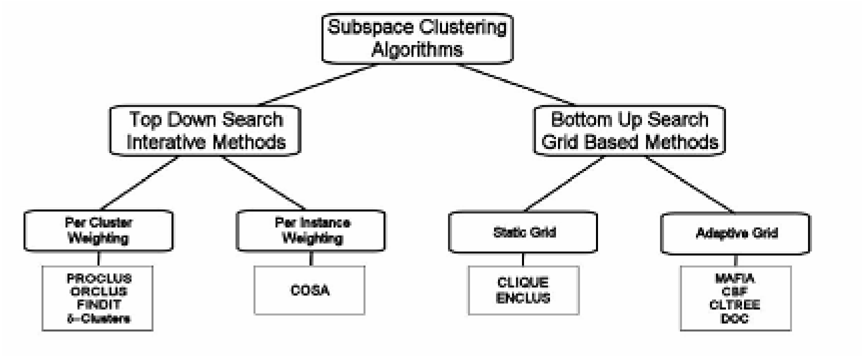

Fig. 1. Categorization subspace clustering algorithms

As shown in Fig 1, the first division in the hierarchy splits subspace clustering algorithms into two groups, the top-down search methods [8] and [11] and bottom-up search methods. The second division is based on the way that the algorithms apply a measure of locality with which they are to evaluate subspaces.

3 Proposed Criterion of Clustering

Consider a set of points in some space of dimensionality D. A weighted cluster C_i is a subset of data points, together with a vector of weights $w_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iD})$, and also a co-efficient d_i such that the points in C are closely clustered according to the L_2 norm distance weighted using $\boldsymbol{w}_i$ and applied co-efficient $\boldsymbol{d}_i$. The component $\boldsymbol{w}_{ij}$ measures the degree of participation of feature j to the cluster C_i. Where d_i is diameter of cluster i, in other words, weight of cluster C_i. If the points in C_i are well clustered along feature j, $\boldsymbol{w}_{ij}$ is large, otherwise it is small. Also for d_i, if the cluster C_i is a big cluster, d_i will be small number. It means if a cluster is big (or has high variance), its distances will be degraded, and otherwise they will be enlarged. Clustering algorithm now faces with the problem "how to estimate the weight vector $\boldsymbol{w}$ for each cluster and co-efficients vector $\boldsymbol{d}$ for clusters of dataset".

Now, the concept of cluster is not based only on points, but also involves a weighted distance metric, i.e., clusters are discovered in spaces transformed by $\boldsymbol{w}$ and simultaneously finding the volumes of clusters. For each cluster we now have a $\boldsymbol{w}$ vector reflecting the importances of features in the cluster and also a weight d standing as its diameter. The effect of $\boldsymbol{w}$ is to transform distances so that the associated cluster is reshaped into a dense hypersphere of points separated from other data. While the weight d is to transform data so that the clusters with low diversities become denser and clusters with high diversities become bulkier. In traditional clustering, the partition of a set of points is induced by a set of representative vectors named centroids or centers. But now the clusters need the centroids plus $\boldsymbol{w}$ vectors and diameters $\boldsymbol{d}$.

Definition. Given a set S of N points $\boldsymbol{x}$ in the D-dimensional Euclidean space, a set of k centers $\{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k\}$, $\boldsymbol{c}_j \in D$, $j=1,\ldots,k$, coupled with a set of corresponding weight vectors $\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_k\}$, $\boldsymbol{w}_j \in D$, $j=1,\ldots,k$, and also a set of diameters d_j, $\boldsymbol{d} \in [0,1]^k$, partition S into k sets $\{S_1,\ldots,S_k\}$:

$$S_j = \left\{ x \mid \sum_{i=1}^{D} d_j \left(w_{ji} \left(x_i - c_{ji} \right)^2 \right) < \sum_{i=1}^{D} d_l \left(w_{li} \left(x_i - c_{li} \right)^2 \right), \forall l \neq j \right\} \qquad 1$$

where w_{ji} represents the ith components of vectors $\boldsymbol{w_j}$ as well as c_{ji} is ith dimension of $\boldsymbol{c_j}$, respectively, finally d_j stands for diameter of cluster j (ties are broken randomly).

The set of centers and weights are optimal with respect to the Euclidean norm, if they minimize the error measure:

$$E_1(C,D,W) = \sum_{j=1}^{k} \sum_{i=1}^{D} d_j \left(w_{ji} \frac{1}{\left| S_j \right|} \sum_{x \in S_j} \left(x_i - c_{ji} \right)^2 \right) \qquad 2$$

subject to the constraints $\forall j \sum_i w_{ji} = 1$ and $\sum_j d_j = 1$. $\boldsymbol{C}$ and $\boldsymbol{W}$ are $(D \times k)$ matrices whose column vectors are $\boldsymbol{c}_j$ and $\boldsymbol{w}_j$, respectively, i.e., $\boldsymbol{C} = [\boldsymbol{c}_1 \ldots \boldsymbol{c}_k]$ and $\boldsymbol{W} = [\boldsymbol{w}_1 \ldots \boldsymbol{w}_k]$, and $|S_j|$ is the cardinality of set S_j. Solving the equation below, the vector $\boldsymbol{D}$ will be non-zero for all clusters and zero for one cluster and all data points will be assigned to the cluster corresponds to zero in $\boldsymbol{D}$ vector.

$$(C_*, D_*, W_*) = \arg\min_{(\mathrm{C,D,W})} E_1(\mathrm{C,D,W}) \qquad 3$$

Our objective, instead, is to find diametric clusters partitioning, where the unit weight gets distributed among all clusters according to the respective variance of data within each cluster. One way to achieve this goal is to add the regularization term $\sum_{j=1}^{k} d_j \log d_j$ which represents the negative entropy of the diameter distribution for clusters [8]. It penalizes solutions with minimal (zero) diameter on the single cluster. The resulting error function will be as follow:

$$E_2(C,D,W) = \sum_{j=1}^{k} \sum_{i=1}^{D} d_j \left(w_{ji} \frac{1}{\left| S_j \right|} \sum_{x \in S_j} \left(x_i - c_{ji} \right)^2 \right) + h_1 \sum_{j=1}^{k} d_j \log d_j \qquad 4$$

subject to the constraint $\sum_j d_j = 1$, where h_1 is parameter of criterion error. But now as it is said the solution of above error criterion will results in maximal (i.e., unit) weight on the feature with smallest variance in each clusters [4]. So, we add the regularization term $\sum_{i=1}^{D} w_{ij} \log w_{ij}$ to our proposed error criterion. Then the proposed error criterion will be as follow:

$$E_3(C,D,W)=\sum_{j=1}^{k}\sum_{i=1}^{D}\left(d_j w_{ji}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i-c_{ji})^2+h_2\sum_{i=1}^{D}w_{ij}\log w_{ij}\right)+h_1\sum_{j=1}^{k}d_j\log d_j \quad 5$$

subject to the constraints $\forall j\sum_i w_{ji}=1$ and $\sum_j d_j=1$. The coefficient h_1, $h_2 \geq 0$ are the parameters of the procedure. Parameters h_1 *and* h_2 control how much the distribution of weight values will deviate from the uniform distribution. We can solve this constrained optimization problem by introducing the Lagrange multipliers λ_j (one for each constraint) and μ for $\sum_j d_j=1$, and minimizing the final (unconstrained now) error criterion:

$$\begin{aligned}E(C,D,W)&=\sum_{j=1}^{k}\sum_{i=1}^{D}\left(d_j w_{ji}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i-c_{ji})^2+h_2\sum_{i=1}^{D}w_{ij}\log w_{ij}\right)\\&+h_1\sum_{j=1}^{k}d_j\log d_j+\sum_{j=1}^{k}\lambda_j(1-\sum_{i=1}^{D}w_{ij})+\mu(1-\sum_{j=1}^{k}d_j)\end{aligned} \quad 6$$

For a fixed partition *P*, fixed c_{ji} and fixed d_j, we compute the optimal $w_{\square\square}$ by setting $\frac{\partial E}{\partial w_{ji}}=0$ and $\frac{\partial E}{\partial \lambda_j}=0$. We obtain:

$$\frac{\partial E}{\partial w_{ji}}=d_j\frac{1}{|S_j|}\sum_{x\in S_j}(x_i-c_{ji})^2+h_2\log(w_{ji})+h_2-\lambda_j=0 \quad 7$$

$$\frac{\partial E}{\partial \lambda_j}=1-\sum_{i=1}^{D}w_{ij}=0 \quad 8$$

solving equation *n* with respect to w_{ji} we obtain

$$h_2\log(w_{ji})=-d_j\frac{1}{|S_j|}\sum_{x\in S_j}(x_i-c_{ji})^2-h_2+\lambda_j \quad 9$$

$$w_{ji}=\exp(-\frac{d_j}{h_2}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i-c_{ji})^2)/\exp(1-\frac{\lambda_j}{h_2}) \quad 10$$

substituting this expression in equation 8 yields to:

$$\frac{\partial E}{\partial \lambda_j}=1-\exp(\frac{\lambda_j}{h_2})\sum_{i=1}^{D}\exp(-\frac{d_j}{h_2}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i-c_{ji})^2-1)=0 \quad 11$$

solving with respect to λ_j we obtain

$$\lambda_j=-h_2\log(\sum_{i=1}^{D}\exp(-\frac{d_j}{h_2}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i-c_{ji})^2-1))=0 \quad 12$$

solving w_{ji}* with considering w_{ji} and λj yields to:

$$w_{ji}^{*} = \frac{\exp(-\frac{d_j}{h_2}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2)}{\sum_{i=1}^{D}\exp(-\frac{d_j}{h_2}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2)} \qquad 13$$

For a fixed partition P, fixed c_{ji} and fixed w_{ji}, we compute the optimal $d_{\square}$ by setting $\frac{\partial E}{\partial d_j} = 0$ and $\frac{\partial E}{\partial \mu} = 0$. We obtain:

$$\frac{\partial E}{\partial d_j} = \sum_{i=1}^{D}\left(w_{ji}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2 \right) + h_1 \log(d_j) + h_1 - \mu = 0 \qquad 14$$

$$\frac{\partial E}{\partial \mu} = (1 - \sum_{j=1}^{k} d_j) = 0 \qquad 15$$

again, solving first equation with respect to d_j we obtain

$$h_1 \log(d_j) = -\sum_{i=1}^{D}\left(w_{ji}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2 \right) - h_1 + \mu \qquad 16$$

$$d_j = \frac{\exp(-\sum_{i=1}^{D}\left(\frac{w_{ji}}{h_1}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2 \right))}{\exp(1 - \frac{\mu}{h_1})} \qquad 17$$

substituting this expression in equation 15 yields to:

$$1 - \sum_{j=1}^{k}\frac{\exp(-\sum_{i=1}^{D}\left(\frac{w_{ji}}{h_1}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2 \right))}{\exp(1 - \frac{\mu}{h_1})} = 1 - \exp(\frac{\mu}{h_1})\sum_{j=1}^{k}\exp(-\sum_{i=1}^{D}\left(\frac{w_{ji}}{h_1}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2 \right) - 1) = 0 \qquad 18$$

solving with respect to μ we obtain

$$\mu = -h_1 \log(\sum_{j=1}^{k}\exp(-\sum_{i=1}^{D}\left(\frac{w_{ji}}{h_1}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2 \right) - 1)) = 0 \qquad 19$$

solving d_j* with considering d_j and μ yields to:

$$d = \frac{\exp(-\sum_{i=1}^{D}\left(\frac{w_{ji}}{h_1}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2\right))}{\sum_{j=1}^{k}\exp(-\sum_{i=1}^{D}\left(\frac{w_{ji}}{h_1}\frac{1}{|S_j|}\sum_{x\in S_j}(x_i - c_{ji})^2\right))} \qquad 20$$

For a fixed partition P and fixed w_{ji}, we compute the optimal $c_{ji}{}^*$ by setting $\frac{\partial E}{\partial c_{ji}} = 0$.

We obtain:

$$\frac{\partial E}{\partial c_{ji}} = \frac{2d_j w_{ji}}{|S_j|}\sum_{x\in S_j}(c_{ji} - x_i) = \frac{2d_j w_{ji}}{|S_j|}\left(|S_j| c_{ji} - \sum_{x\in S_j} x_i\right) = 0 \qquad 21$$

Solving with respect to c_{ji} gives

$$c_{ji}^* = \frac{1}{|S_j|}\sum_{x\in S_j} x_i \qquad 22$$

Proposition. *When $h_1 = 0$ and $h_2 = 0$, the error function E3 is reduced to E1; when $h_1=\infty$ and $h_2=\infty$, the error function E3 is reduced to SSE.*

4 Search Strategy

We need to provide a search strategy to find a partition P that identifies the final partition. Our approach is similar to k-means and LAC where they iteratively improve the quality of initial centroids, weights and diameters, by investigating the space near the centers to estimate the dimensions that matter the most. Specifically, we proceed as follows. We start with *well-scattered* points in S as the k centroids: we choose the first centroid at random, and select iteratively the others so that they are as far as possible from each so far-selected centroids. We initially set all weights and all diameters to $\tilde{1}D$ and $1/k$ respectively. Given the initial centroids $\boldsymbol{c}_j$, for j =1,…, k, we compute the corresponding sets S_j as given in the definition above. Then we use the S_j to compute weights W_j. The computed weights are used to compute the diameters D_j, and finally S_j, D_j and W_j are used to update centroids $\boldsymbol{c}_j$.

Input. N points $\mathbf{x} \in R^D$, k, and h_1, h_2.

1. Start with k initial centroids $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_k$;
2. Set $d_j = \tilde{1}k$, for each centroid $\boldsymbol{c}_j$, j =1,…, k
3. Set $w_{ji} = \tilde{1}D$, for each centroid $\boldsymbol{c}_j$, j =1,…, k and each feature i =1,…,D
4. For each centroid $\boldsymbol{c}_j$, compute S_j considering w_{ji} and d_j;
5. Compute new weights w_{ji} using $\boldsymbol{c}_j$ and d_j.
6. For each centroid $\boldsymbol{c}_j$, compute $\boldsymbol{d}_j$ considering w_{ji} and S_j;
7. Compute new centroids.
8. Iterate 4,5,6,7 until convergence.

Table 1. Experimental results. * indicates dataset is normalized with mean of 0 and variance of 1, *N*(0,1).

	Simple Methods (%)					
Dataset	Single Linkage	Average Linkage	Complete Linkage	K-means	Fuzzy K-means	Proposed Algorithm
Breast Cancer*	65.15	65.15	65.15	95.37	55.34	**98.42**
Bupa*	58.26	57.68	57.68	54.49	37.77	**59.36**
Glass*	35.05	37.85	37.85	45.14	49.07	**53.41**
Wine	38.76	37.64	39.89	96.63	96.63	**98.11**
Yeast*	34.38	35.11	38.91	40.20	35.65	**47.34**
Iris	66.67	67.33	38.00	82.80	89.33	**94.33**
SAHeart*	65.15	65.37	64.72	63.12	45.19	**69.03**
Ionosphere*	63.82	67.52	65.81	70.66	53.22	**72.84**
Galaxy*	25.70	25.70	25.70	29.88	29.41	**36.19**

5 Experimental Results

This section evaluates the result of applying proposed algorithm on 9 real datasets available at USI repository [15]. The final performance of the clustering algorithm is evaluated by re-labeling between obtained partition and the ground true labels and then counting the percentage of the true classified samples. Table 1 shows the performance of the proposed method comparing with the most common base methods. To reach these results for proposed clustering algorithm, by trial and error we turn to the best parameter values for h_1 and h_2 per each dataset. As it can be seen in Fig. 2 the parameters h_1 and h_2 are set for all dataset in the ranges [0.06,0.15] and [0.09,0.27]. These ranges are obtained by trial and error to be the best options for all datasets.

The four first columns of Table 1 are the results of some base clustering algorithms. The results show that although each of these algorithms can obtain a good result over a specific dataset, it does not perform well over other datasets. But well setting of parameters of proposed method leads to its perfect superiority to most of well-known clustering algorithms. The only drawback of the proposed algorithm is the sensitivity to its two parameters. For a comprehensive study, look at the Fig 2 and Fig. 3. In Fig. 2, the normalized mutual information (*NMI*) values between output labels of WLAC algorithm and real labels for different values of parameters h_1 and h_2 over Iris dataset are depicted. It is worthy to mention that the diagram is a cut from its best positions over all possible values for its parameters (h_1 and h_2 both are defined in the range $[0,\infty]$). In Fig. 3, the normalized mutual information (*NMI*) values between output labels of LAC algorithm and real labels for different values of parameters h over Iris dataset are depicted. Note that the diagram is again a cut from its best positions over all possible values for its parameter (h is defined in the range $[0,\infty]$, i.e. y is defined in range $[1,\infty]$ and x is defined in range [1,10]). For a more detailed view of Fig. 3, look at the Fig. 4. It is obvious that in the best setting of parameter for LAC algorithm in the Iris dataset, it can't outperform WLAC algorithm with just a near best setting of its parameters, i.e. as it is presented in Fig. 2. Also note that WLAC is less sensitive to its well setting of parameters than LAC to its well setting of parameter.

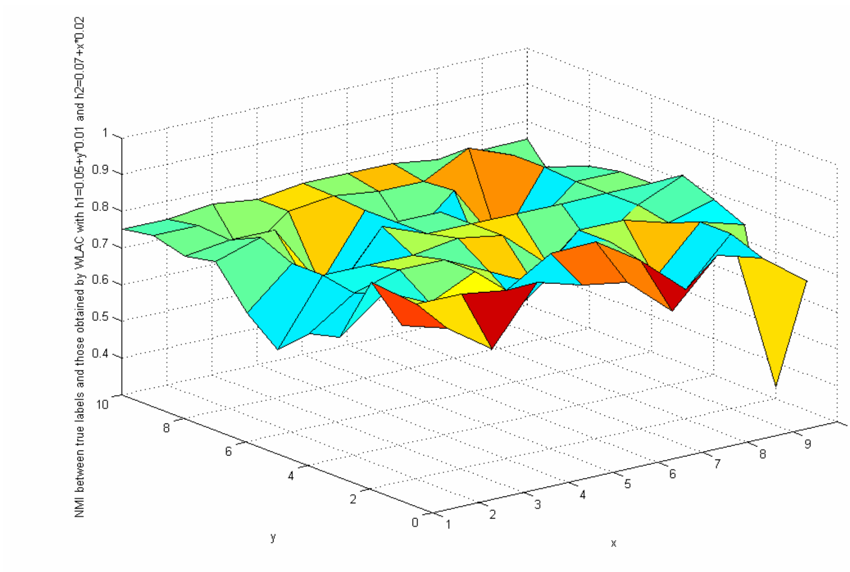

Fig. 2. NMI between real labels of Iris dataset and those of obtained by WLAC algorithm using different values of parameters h_1 and h_2

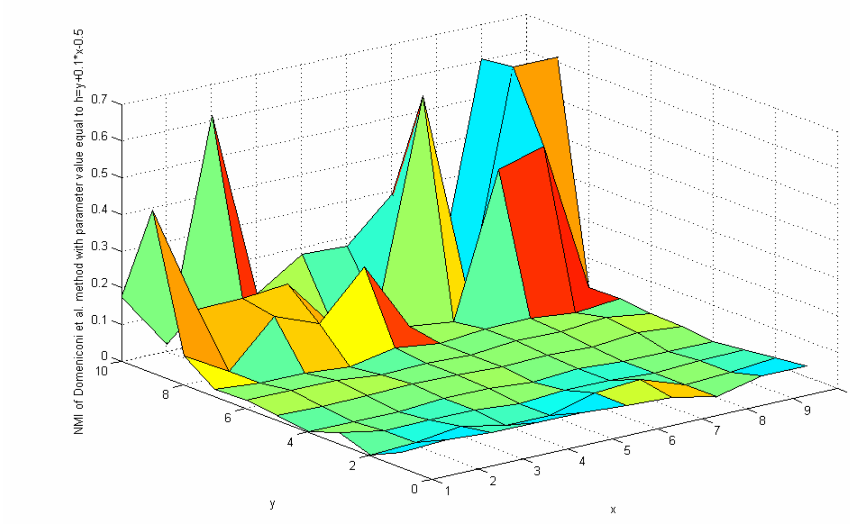

Fig. 3. NMI between real labels of Iris dataset and those of obtained by LAC algorithm using different values of parameters h (h=y+0.1*x-0.5)

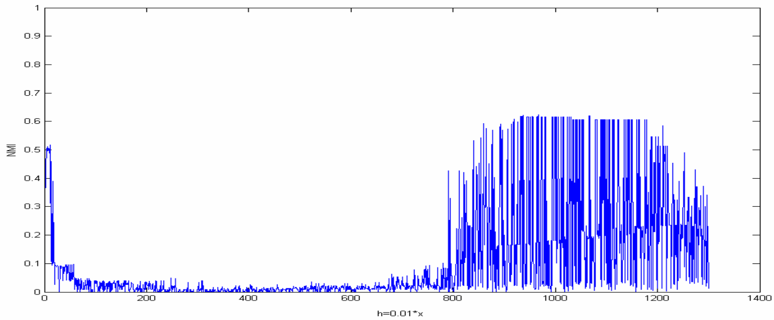

Fig. 4. NMI between real labels of Iris dataset and those of obtained by LAC algorithm using different values of parameters h (h=0.01*x)

6 Conclusion and Future Work

This paper proposes a new metric in clustering which simultaneously considers the feature weighting and cluster weighting. It is also solved in algebraic mathematic so as to obtain the minima. A new algorithm based on k-means is presented to handle the amenities added to k-means metric, i.e. Sum of Square Errors (SSE). The proposed method have two parameters which must be appropriately set to obtain a well output partitioning. Tuning these parameters can be an open problem as future work which we are working on it.

The only drawback of the proposed algorithm is its sensitivity to its two parameters. But well setting of its parameters leads to its perfect superiority to most of well-known clustering algorithms. So for future work, it can be studied how to overcome the problem of automatic setting of its two parameters.

References

1. Blum, A., Rivest, R.: Training a 3-node neural networks is NP-complete. Neural Networks 5, 117–127 (1992)
2. Chang, J.W., Jin, D.S.: A new cell-based clustering method for large, high-dimensional data in data mining applications. In: Proceedings of the ACM Symposium on Applied Computing, pp. 503–507. ACM Press, New York (2002)
3. Cheng, C.H., Fu, A.W., Zhang, Y.: Entropy-based subspace clustering for mining numerical data. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 84–93. ACM Press, New York (1999)
4. Domeniconi, C., Al-Razgan, M.: Weighted cluster ensembles: Methods and analysis. TKDD 2(4) (2009)
5. Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M., Papadopoulos, D.: Locally adaptive metrics for clustering high dimensional data. Data Min. Knowl. Discov. 14(1), 63–97 (2007)

6. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19(9), 1090–1099 (2003)
7. Faceli, K., Marcilio, C.P., Souto, d.: Multi-objective Clustering Ensemble. In: Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS 2006) (2006)
8. Friedman, J.H., Meulman, J.J.: Clustering objects on subsets of variables. Journal of the Royal Statistical Society, Series B 66, 815–849 (2004)
9. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
10. Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
11. Liu, B., Xia, Y., Yu, P.S.: Clustering through decision tree construction. In: Proceedings of the Ninth International Conference on Information and Knowledge Management, pp. 20–29. ACM Press, New York (2000)
12. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
13. Miller, R., Yang, Y.: Association rules over interval data. In: Proc. ACM SIGMOD International Conf. on Management of Data, pp. 452–461 (1997)
14. Mirzaei, A., Rahmati, M., Ahmadi, M.: A new method for hierarchical clustering combination. Intelligent Data Analysis 12(6), 549–571 (2008)
15. Newman, C.B.D.J., Hettich, S., Merz, C.: UCI repository of machine learning databases (1998), `http://www.ics.uci.edu/~mlearn/MLSummary.html`
16. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: a review. ACM SIGKDD Explorations Newsletter 6(1), 90–105 (2004)
17. Srikant, R., Agrawal, R.: Mining Quantitative Association Rules in Large Relational Tables. In: Proc. of the ACM SIGMOD Conference on Management of Data, Montreal, Canada (1996)