NEW INFORMATION-BASED CLUSTERING METHOD USING RENYI'S ENTROPY AND FUZZY C-MEANS CLUSTERING

M. Aghagolzadeh[a], H. Soltanian-Zadeh[a, b], B. Araabi[a]

[a]*Department of Electrical and Computer Engineering, University of Tehran, Tehran 14395, Iran*
[b]*Radiology Image Analysis Lab., Henry Ford Health System, Detroit, MI 48202, USA*

Emails: m.golzadeh@ece.ut.ac.ir, hszadeh@ut.ac.ir, hamids@rad.hfh.edu, araabi@ut.ac.ir

Abstract: *This paper presents a new clustering method based on Renyi entropy. The proposed method maximizes entropy of clusters using between and within clusters entropies. It is a top-down multi-resolution method and uses the initial clusters found by Fuzzy C-Means. Applications of the proposed algorithm on the synthetic data are compared with those of C-Means and Gustafson-Kessel algorithms. Results show superiority of the proposed algorithm to these methods.*

Keywords: *Information theory, Renyi's entropy, Top-down hierarchical algorithms, clustering.*

1. Introduction

Clustering is an important tool for pattern recognition; it is an unsupervised approach for splitting data into its natural groups. Clustering has extensive applications in image segmentation and compression, machine learning, and remote sensing and data mining [1]-[3]. In recent years, several clustering methods based on artificial neural networks [4] and support vector machines [5] have been developed that are talented to identify clusters with any shape and without knowing the correct number of clusters. However, these methods are often very complex and necessitate perfect association.

Clustering depends on data structure and information theory is a useful tool for mining data structures. In this paper, a novel hierarchical algorithm is presented, which is based on Renyi's entropy and applied for clustering. The advantages of the proposed algorithm compared to Gokcay and Jenssen algorithms [6]-[9] are its higher speed and stability. The novelties of the proposed algorithm are: a) using an improved factor in detecting the worst cluster for clustering enhancement, and b) using Fuzzy C-mean clustering as the primary clustering for decreasing computational complexity and increasing stability.

In the next section, Renyi's entropy is introduced and the reason behind its usage is described. In Section 3, the proposed algorithm is presented. The method for finding the worst cluster and allocating data points to clusters with differential entropy and the initial clustering method and also a technique for finding the ultimate clustering are expressed in this section. In Section 4, the experimental results are presented. Conclusions are given in Section 5.

2. Renyi's Entropy

Renyi defined entropy with order α as [10]:

$$H_R(x) = \frac{1}{1-\alpha}\log\left(\int P^{\alpha}(x)dx\right) \qquad (1)$$

which becomes Shannon's entropy when $\alpha \to 1$. The main application of Renyi's entropy is when $\alpha = 2$, where it is also called quadratic entropy and can be written as [10]:

$$H_R\left(\forall x_i, x_j \in C\right) = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(x_i - x_j, 2\delta^2)\right) \qquad (2)$$

Therefore, the Renyi's quadratic entropy can be computed from the Gaussian functions summation based on the differential between data point pairs. The term inside phrase in the logarithm function of equation (2) is named information potential:

$$V\left(\forall x_i, x_j \in C_k\right) = \frac{1}{{N_k}^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} G(x_i - x_j, 2\delta^2) \qquad (3)$$

$$H(x) = -\log V(x) \qquad (4)$$

Since computing Renyi's entropy for data points of a dataset is very easy, it has been utilized as a criterion for clustering in the proposed algorithm.

3. Proposed Clustering Algorithm

The proposed algorithm has two main steps: a) finding the worst cluster among the existing clusters and splitting it into its constructing data points. b) Allocating the extracted data points from step (a) to the remaining clusters. These steps are repeated until an acceptable clustering is attained.

3.1 Finding Worst Cluster

In any iteration of the proposed top-down hierarchical algorithm, one of the clusters is eliminated and its data points are allocated to the remaining clusters. There are numerous methods for finding a cluster to be vanished, which one of them is selecting a cluster randomly from the clusters. Another technique is choosing a cluster that has the maximum inter-coordination (variance). Both of these methods are not capable to distinguish data structures and select the actual worst cluster. In the proposed algorithm, between clusters entropy is used for finding the worst cluster; this measure, first proposed by Gokcay et al [6], [7], is:

$$V(C_1,\ldots,C_K)=\frac{1}{\prod_{k=1}^{K} N_k}\sum_{i=1}^{N}\sum_{j=1}^{N} M(x_{ij})G(x_i-x_j,\sigma^2) \qquad (5)$$

$$H(C_1,\ldots,C_K)=-\log V(C_1,\ldots,C_K) \qquad (6)$$

Elements of matrix $M(x_{ij})$ are zero when $x_i, x_j \in C_k$ and one otherwise. One of the difficulties in utilizing equation (5) is selecting σ, this will be addressed in Section 3.e. As the clusters are moved away from each other, V will decrease and between cluster entropy, H will increase. The worst cluster that is the best candidate for being eliminated has the maximum amount of between cluster entropy.

For finding the worst cluster, between clusters entropy is computed in the absence of a desirable cluster for each of the data clusters. First, each cluster is eradicated and between clusters entropy is calculated for the rest of the clusters by (5). The worst cluster is established as the cluster that offers the maximum amount of entropy or minimum amount of information potential:

$$C_k=\min_k\{V(C_2,\ldots,C_K),\ldots,V(C_1,\ldots,C_{K-1})\} \qquad (7)$$

The most important problem of the above method is the clusters that are copious centralized with a low number of data points. These clusters merge with a low possibility with other clusters and usually reside as autonomous islands. In computing entropy in the absence of this kind of clusters, information potential is increased; so they are not detected as a worst cluster. For solving this problem, the number of data samples of a cluster is used as a multiplicative factor in the entropy calculation. Experiments show that this multiplication factor facilitates for enhanced clustering and prevents the algorithm from trapping in local minimum. So the improved version of (7) is given by multiplying the information potential by the number of clusters with power α:

$$C_k=\min_k\left\{N^{\alpha}(1)\times V(C_2,\ldots,C_K),\ldots,N^{\alpha}(K)\times V(C_1,\ldots,C_{K-1})\right\} (8)$$

Experiments show that the best results are achieved for $\alpha=2$. After finding the worst cluster, it is vanished and its data points label are removed.

3.2 Proposed Method for Allocating Free Data Points

After finding the worst cluster and removing the data point labels, each of the freed data points are independently allocated to one of the rest of the clusters by differential entropy.

3.2.1 Clustering Based on Differential Entropy

Any data point that is attached to a cluster would increase uncertainty or entropy of that cluster. When a data point is properly assigned to a cluster, its entropy will be increased less than if it is adjoined improperly to another cluster. This idea suggests a new method; a data point is allocated to a cluster that its entropy is minimally increased compared to the other clusters. The following equation finds this cluster.

$$C_k=\max_k\{(V(C_1+x_i)-V(C_1)),\ldots,(V(C_K+x_i)-V(C_K))\} \qquad (9)$$

which requires computation of the within cluster entropy.

3.2.2 Within Cluster Entropy

This criterion is similar to the between cluster entropy; the main difference is that within cluster entropy shows entropy among data pairs of a single cluster but between cluster entropy computes entropy among data pairs of different clusters. The within cluster entropy for data points of a cluster is defined by the following equation:

$$V(\forall x_i,x_j\in C_K)=\frac{1}{{N_k}^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}G(x_i-x_j,\sigma^2) \qquad (10)$$

$$H(C_k)=-\log V(C_k) \qquad (11)$$

For fast computation of within cluster entropy, the elements of the matrix G are computed for all pairs of data samples and saved in the beginning of the algorithm; then simply using general matrix operations, equation (10) is calculated. Within cluster entropy can be calculated from the complete matrix G:

$$V(\forall x_i,x_j\in C_K)=\frac{1}{{N_k}^2}\sum_{i=1}^{N}\sum_{j=1}^{N}M'(x_{ij})G(x_i-x_j,\sigma^2) (12)$$

Elements of matrix $M'(x_{ij})$ are one when $x_i, x_j \in C_k$ and zero otherwise. This fast method can be also utilized for calculating between cluster entropy.

3.2.3 Ordering of Free Data Points for Clustering

The order of choosing free data points of the eliminated cluster for a new clustering is important; randomly choosing them might make the clustering algorithm unstable. For this purpose, a method is needed for appropriate ordering of the free data points. One simple method is updating the changed cluster after the data point allocating process. This method stabilizes the clustering process but it might fall into the local minima. In the proposed algorithm, the arrangement of free data points is upon the nearest free data point to data points of the rest of the clusters. The changed cluster after allocating any data point will be updated and this operation will be repeated until the last free data point of the vanished cluster. This method decreases the probability of trapping in the local minima and the clustering will become stable. Figure 1 shows an iteration of the proposed algorithm.

3.3 Initial Clustering

In the proposed algorithm, Fuzzy C-means clustering is used as initial clustering. The advantage of this method is its faster execution compared to Jenssen et al’s method; it benefits from multi-resolution concept. This method guarantees the convergence and transfers data points to plenty of clusters, each with a few data points. The number of clusters in initial clustering depends on the number of data points and the final number of clusters expected. It is clear that increasing the number of dataset points or increasing the number of final clusters will increase the number of clusters in the primary clustering.

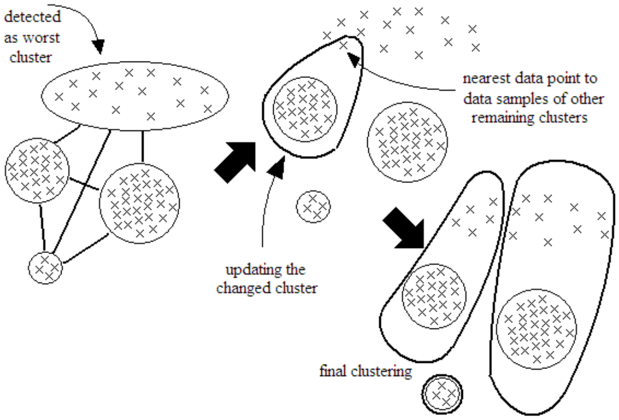

Figure 1: One step of the proposed algorithm and reduction of one cluster and allocating its data point to the remaining clusters.

3.4 Final Clustering

The proposed algorithm begins from a large number of clusters and in each step of the algorithm, one cluster is vanished and this operation is repeated until two clusters remain. If the clustering is stored in each iteration, then a hierarchical clustering from N primary clusters to two clusters is available. At last, the perfect clustering is selected from the stored clustering at each step.

It is difficult and sometimes impossible to determine the number of clusters accurately. There are several methods that can estimate the number of final clusters. One of them uses between scatter matrix. When the number of clusters reduces, the size of the clusters enlarges and the trace of between scatter matrix increases. When the rate of increasing diminishes, the process stops. Another efficient way, utilized by Jenssen et al [8], [9], uses between cluster entropy for finding the number of final clusters. This criterion increases when the number of cluster reduces; so the final clustering is chosen when a high disparity is seen from one step to the next.

3.5 Method for Choosing σ

One of the main issues of the proposed algorithm is choosing σ in equation (5). By choosing σ as a small quantity, a high attention is given to clustering of close data points and by selecting σ as a large value, an attention is given to clustering of far data points. Different data need different values for σ; unfortunately, there is not a particular method for choosing σ . A simple method for estimating σ is defined by the following equation [11].

$$\sigma = \min\left\{ \frac{\sqrt{Var(x\_Dimension)}\times 1.06}{\sqrt{N}}, \ldots, \frac{\sqrt{Var(z\_Dimension)}\times 1.06}{\sqrt{N}} \right\} \tag{13}$$

This equation sets σ equal to the minimum σ in the direction of one of the features.

4. Experimental Results

To evaluate the proposed algorithm and compare it with other algorithms, several experiments are done. In the following subsections, we present 3 illustrative examples.

4.1 Standard Synthetic Data

Figure 2 shows the clustering results for a standard dataset, where the clusters are not mass prototype clusters. Note that neither C-means nor Gustafson-Kessel algorithms [12] are able to detect shell prototype clusters, but the proposed algorithm is able to detect line and shell prototypes properly. To ensure how the final clustering is achieved, in figure 3 the quantity of between cluster entropy is plotted for all steps of the proposed algorithm on the dataset in Figure 2. Since there is a large variation in clustering between two and three clusters, the final number of clusters is set to two.

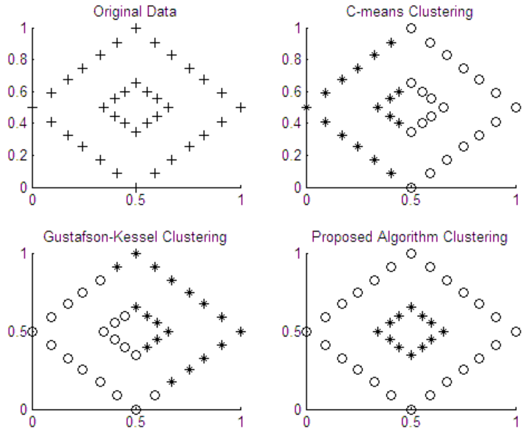

Figure 2: Clustering a dataset with two shell prototype clusters.

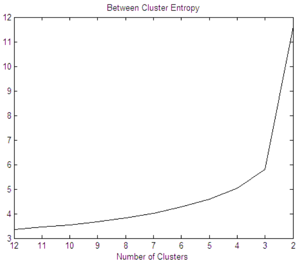

Figure 3: Between cluster entropy for the dataset presented in Figure 2.

4.2 Centralized Synthetic Data

Figure 4 shows clustering results for a dataset with a huge number of data samples and centralized clusters. Note that the proposed algorithm outperforms both of the C-means and Gustafson-Kessel algorithms.

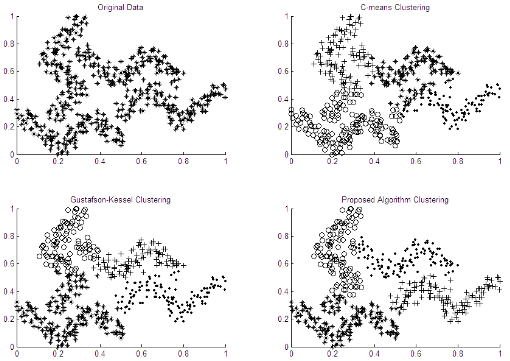

Figure 4: Clustering a dataset with four centralized clusters.

4.3 Synthetic Particular Data

Figure 5 shows a dataset with combined centralized and regionalized clusters. Note that the proposed algorithm is able to recognize regional clusters with any degree of distraction from centralized clusters by changing the value of σ. Also, note that it outperforms the other methods.

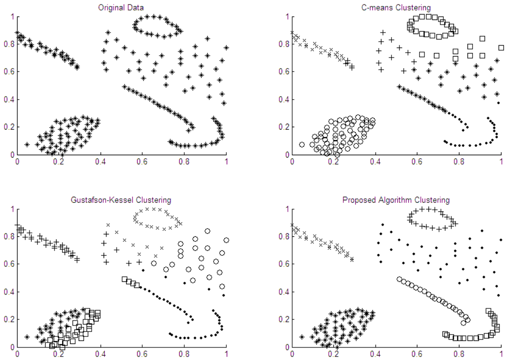

Figure 5: Clustering a dataset with five centralized and one regionalized clusters.

5. Conclusions

In this paper, a new top-down hierarchical method is proposed for data clustering based on information theory and Renyi's entropy. The proposed algorithm uses between cluster entropy and within cluster entropy for removing one cluster in each step. The proposed algorithm detects different structures of clusters (mass, shell, and linear clusters) and outperforms the C-means clustering and Gustafson-Kessel algorithm. The proposed algorithm is compared with recent information-based algorithms using Renyi's entropy, like Jenssen et al method and Gokcay et al method. The proposed method has a higher speed of execution and has solved the convergence problem. Experiments using the proposed algorithm are done on the synthetic datasets. The results show the effectiveness of the proposed algorithm. We show that this algorithm is able to cluster both highly distracted clusters and centralized clusters.

References

[1] H. Frigui, and R. Krishuapuram, A Robust Competitive Clustering Algorithm with Applications in Computer Vision, *IEEE Trans. on PAMI*, Vol. 21, No. 5, Page(s):450 – 465, 1999.

[2] C. Carpineto, and G. Romano, A Lattice Conceptual Clustering System and its Application to Browsing Retrieval, *Machine Learning*, Vol. 24, No. 2, Page(s): 95-122, 1996.

[3] D. Judd, P. McKinley, and A.K. Jain, Large-Scale Parallel Data Clustering, *IEEE Trans. on PAMI*, Vol. 20, No. 8, Page(s):871 – 876, 1998.

[4] T. Eltoft, and R. deFigueiredo, A New Neural Network for Cluster Detection and Labeling, *IEEE Trans. On Neural Networks*, Vol. 9, No. 5, Page(s):1021 – 1035, 1998.

[5] A. Ben Hur, D. Horn, H. Siegelman, and V. Vapnik, Support Vector Clustering, *Journal of Machine Learning Research 2*, Vol. 2, Page(s): 125-137, 2001.

[6] E. Gokcay, and J. C. Principe, Information Theoretic Clustering, *IEEE Tran. On PAMI*, Vol. 24, No. 2, Page(s):158 – 171, Feb. 2002.

[7] E. Gokcay, and J. C. Principe, A new clustering evaluation function using Renyi's Information Potential, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, 5-9 June 2000, Vol. 6, Page(s) 3490-3493.

[8] R. Jenssen, K.E. Hild, D. Erdogmus, J.C. Principe, and T. Eltoft, Clustering using Renyi's Entropy, *Proceedings of the International Joint Conference on Neural Networks*, 20-24 July 2003, Vol. 1, Page(s): 523-528.

[9] R. Jenssen, T. Eltoft, and J. C. Principe, Information Theoretic Clustering: A unifying review of three recent algorithms, *Proc. Nordic Int'l. Symposium on Signal Processing (NORSIG2004)*, Page(s): 292-295, Espoo, Finland, June 2004.

[10] A. Renyi, On Measures of Entropy and Information, Proceedings of the 4th Berkley Symposium on Mathematics of Statistics and Probability, Vol. 1, Page(s): 547–561, 1961.

[11] B.W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, 1986).

[12] S. Theodoridis and K. Koutroumbas, *Pattern Recognition* (Academic Press, 1999).