Information-Based Clustering Using Renyi's Entropy and Scatter Matrices

M. Aghagolzadeh*, H. Soltanian-Zadeh*, **, B. Araabi*

*Department of Electrical and Computer Engineering, University of Tehran, Tehran 14395, Iran
**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

In this paper a new method based on the information theory and Renyi's definition of entropy is proposed. The main idea of the proposed method is detecting clusters in a dataset such that the entropy of the clusters is maximized. For this purpose, four main criteria, namely, between cluster entropy, within cluster entropy, between scatter matrix and within scatter matrix, and a top-down hierarchical clustering method are used. The initial clustering is done using Fuzzy C-Means method. Applications of the proposed algorithm on synthetic data are compared with those of C-Means, Gustafson-Kessel algorithm, and recently proposed algorithms using information theory and Renyi's entropy. This comparison shows superiority of the proposed algorithm to the mentioned methods.

Keywords: Between cluster entropy, between scatter matrix, fuzzy C-means clustering, Renyi's entropy, top-down hierarchical algorithm, within cluster entropy, within scatter matrix.

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 [1], machine learning [2], and remote sensing and data mining [3]. Clustering has been generally used when labeling data by a human operator is costly and subject to error. The purpose of clustering is labeling unlabeled data so that the data in a labeled group have the highest similarity among themselves and the highest dissimilarity with data in other groups.

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 is dependent on data structures and information theory is a useful tool for mining data structures.

Information theory was first introduced by Shannon in 1948 [6]. However, its practical difficulties in approximating the probability density function, has limited its usage in clustering methods. The majority of clustering algorithms utilize the minimum variance criteria. Among these, C-means clustering, expectation maximization, splitting and merging, fuzzy C-means clustering, and ART neural networks [7] can be mentioned. Information theory was first used by Watanabe et al for clustering [8]. Information theory unchallenged supremacy in clustering stems from its ability to extract data structures further than the 2nd order statistics (variance). The major problem in employing information theory in clustering algorithms is making some unrealistic assumptions. This problem has been solved by Renyi's entropy, which estimates entropy pointwise without any distribution assumption. Renyi's entropy has been used in recent years in algorithms proposed by Gokcay et al [9, 10] and Jenssen et al [11, 12], which have generated superior results in clustering data with complex structures.

In this paper, a novel hierarchical algorithm has been developed based on Renyi's entropy for clustering. The advantages of the proposed algorithm compared to Gokcay and Jenssen algorithms [9-12] are its higher speed and stability. The proposed algorithm begins from a large number of clusters and in a hierarchical approach, in each iteration first it finds the worst (the most improper) cluster by introducing a new measurement using between clusters entropy and within scatter matrix, and then eliminates that cluster. Then by measuring within cluster entropy, each data point of the vanished cluster is allocated to a remaining cluster. This loop is repeated until between scatter matrix, calculated after each iteration, stops decreasing. The efficiency of the proposed algorithm in clustering complex structures is compared with ordinary algorithms such as C-means clustering and Gustafson-Kessel algorithm (a generalized version of fuzzy C-means clustering) and also with recent information-based clustering algorithms such as algorithms proposed by Jenssen et al and Gokcay et al.

The novelties of the proposed algorithm are: a) using within scatter matrix with an improved factor of between cluster entropy in detecting the worst cluster, b) using between scatter matrix to find an acceptable clustering, and c) 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 2, the proposed algorithm is presented. A method for finding the worst cluster and allocating data points to clusters with differential entropy, the initial clustering method, and also a technique for finding the ultimate clustering are expressed in this section. In Section 3, the experimental results are presented.

1. Renyi's entropy

For computing entropy, first the probability density function of data samples must be estimated. For data with N data samples, the probability density function can be estimated using Parzen's window estimator through Gaussian functions:

$$P(x)=\frac{1}{N}\sum_{i=1}^{N}G\left(x-x_i,\sigma^2\right) \qquad (1)$$

where the Gaussian function G is defined as:

$$G\left(x-x_i,\sigma^2\right)=\frac{1}{(2\pi\sigma)^{\frac{d}{2}}}\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) \qquad (2)$$

In (2), *d* is the vector length of the data point's feature. Renyi defined a new concept for entropy with order α as [13]:

$$H_R(x)=\frac{1}{1-\alpha}\log\left(\int P^{\alpha}(x)dx\right) \qquad (3)$$

Renyi's entropy 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 [13]:

$$H_R(x)=-\log\left(\int P^2(x)dx\right) \qquad (4)$$

Combining (1) and (4) results in:

$$H_R(x)=-\log\left(\int\left(\frac{1}{N}\sum_{i=1}^{N}G\left(x-x_i,\sigma^2\right)\right)\left(\frac{1}{N}\sum_{j=1}^{N}G\left(x-x_j,\sigma^2\right)\right)dx\right)$$
$$=-\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int G(x-x_i,\sigma^2)G(x-x_j,\sigma^2)dx\right) \qquad (5)$$

It is clear that the convolution of two Gaussian functions is also a Gaussian function; so equation (5) can be altered into a simple form:

$$H_R\left(\forall x_i,x_j\in C\right)=-\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}G(x_i-x_j,2\sigma^2)\right) (6)$$

Therefore, the Renyi's quadratic entropy can be simply computed from the summation of Gaussian functions based on the differential between every data point pairs. The term inside bracket in equation (6) is named information potential:

$$V\left(\forall x_i,x_j\in C_k\right)=\frac{1}{{N_k}^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}G(x_i-x_j,2\sigma^2) \qquad (7)$$

$$H(x)=-\log V(x) \qquad (8)$$

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

2. Proposed clustering algorithm

The proposed hierarchical algorithm has two main steps: a) finding the worst cluster among the existing clusters, b) allocating the worse cluster data points to the remaining clusters. These steps are repeated until an acceptable clustering is attained. In the following, different parts of the proposed algorithm are described in more details.

2.1. Finding the 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; 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. The proposed algorithm uses between clusters entropy for finding the worst cluster; this measure, first proposed by Gokcay et al [9], [10], is:

$$V(C_1,\ldots,C_K)=\frac{1}{\prod_{k=1}^{K}N_k}\sum_{i=1}^{N}\sum_{j=1}^{N}M\left(x_{ij}\right)G\left(x_i-x_j,\sigma^2\right) \qquad (9)$$

$$H(C_1,\ldots,C_K)=-\log V(C_1,\ldots,C_K) \qquad (10)$$

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 (9) is selecting σ, this will be addressed in Section 2.5. As the clusters are moved away from each other, V will decrease and between cluster entropy, H will increase.

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 (9). The worst cluster is established as the cluster that offers the maximum amount of entropy or minimum amount of information potential as shown in the following equations.

$$C_k = \min_k \{V(C_2,\dots,C_K),\dots,V(C_1,\dots,C_{K-1})\} \quad (11)$$

Or

$$C_k = \max_k \{H(C_2,\dots,C_K),\dots,H(C_1,\dots,C_{K-1})\} \quad (12)$$

Equations (11) and (12) are equivalent; (12) can be deduced from (11) considering (10). The most important problem of the above method is the clusters that are copious centralized with a low number of data points. It is unlikely that these clusters merge with the other clusters and usually remain as autonomous islands. In computing entropy in the absence of this kind of clusters, information potential increases; 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 enhanced clustering and prevents the algorithm from trapping in local minima. So, an improved version of (12) is obtained by multiplying the information potential by the number of clusters with power α :

$$C_k = \min_k \{N^\alpha(1)\times V(C_2,\dots,C_K),\dots,N^\alpha(K)\times V(C_1,\dots,C_{K-1})\} \quad (13)$$

By transforming (13) to an equivalent equation based on between cluster entropy, (14) is resulted:

$$C_k = \max\{H(C_2,\dots,\underset{k}{C_K}) - \alpha\log N(1),\dots,$$
$$\text{(14)}\ H(C_1,\dots,C_{K-1}) - \alpha\log N(K)\}$$

Experiments show that the best results are achieved for $\alpha = 2$. Although the above equation solves the problem of selecting small centralized clusters with a low number of data points, it prevents selecting large spread clusters with a high number of data points. There are two ways for overcoming this problem:

1) Increasing the initial number of clusters by splitting these large clusters to some smaller clusters. This method is efficient but it will increase the computational complexity.

2) Using within scatter matrix criteria in (13) so that the chance of choosing a large spread cluster as the worst cluster increases.

$$C_k = \min_k \{N^\alpha(1)\times V(C_2,\dots,C_K) - \beta.S(C_1),\dots,$$
$$\text{(15)}\ N^\alpha(K)\times V(C_1,\dots,C_{K-1}) - \beta.S(C_K)\}$$

In (15), best results are achieved by choosing $\beta = 0.3$.

2.2. Proposed method for allocating worst cluster data points

After finding the worst cluster, each of its data points are allocated to one of the remaining clusters by differential entropy. This method was first used by Jenssen et al [9, 10] for information-based clustering.

2.2.1. 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 increase less than if it is assigned 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.

$$C_k = \min_k \{(H(C_1 + x_i) - H(C_1)),\dots,(H(C_K + x_i) - H(C_K))\} \quad (14)$$

Or

$$C_k = \max_k \{(V(C_1 + x_i) - V(C_1)),\dots,(V(C_K + x_i) - V(C_K))\} \quad (15)$$

Equations (14) and (15) require computation of the within cluster entropy represented by $H(C_K)$.

2.2.2. Within cluster entropy

This criterion is similar to the between cluster entropy; 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:

$$V(\forall x_i, x_j \in C_K) = \frac{1}{{N_k}^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} G(x_i - x_j, \sigma^2) \quad (16)$$

$$H(C_k) = -\log V(C_k) \quad (17)$$

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 (16) is calculated. Within cluster entropy can be calculated from the complete matrix G:

$$V(\forall x_i, x_j \in C_K) = \frac{1}{{N_k}^2}\sum_{i=1}^{N}\sum_{j=1}^{N} M'(x_{ij}) G(x_i - x_j, \sigma^2) \quad (18)$$

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.

2.2.3. Ordering of free data points for clustering

The order of assigning data points of the eliminated cluster to the remaining clusters is important; a random order may make the clustering algorithm unstable. Thus, an appropriate ordering method is needed. A simple method is updating the changed cluster after a data point is allocated. This method stabilizes the clustering process but may push it into a local minimum. In the proposed algorithm, the data points are ordered based on the distances of the free data points to the data points of the remaining clusters. The changed cluster after allocating a data point is updated and this operation is repeated until the last data point of the vanished cluster is classified. This stabilizes the clustering process and decreases the probability of trapping in a local minimum. Figure 1 shows an iteration of the proposed algorithm.

2.3. Initial Clustering

Final clustering highly depends on the initial clustering. In doing the initial clustering two points are considered: stability assurance and computational complexity decreasing. The initial clustering can be done with numerous methods. For example, initial clustering can be done by arbitrary clustering of 20 percent of the data points and then starting the clustering procedure (Jenssen et al [11]) or randomly clustering all data points. The problem of the former method is its high computational complexity and needs to cluster 80 percent of remaining data points to the primary clusters and then begins the iterative clustering. The problem of the later method is the possibility of final clustering instability due to utilizing random 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 dataset points and the final number of clusters expected.

2.4. 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 action 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. In the proposed algorithm, between scatter matrix is used to find the ultimate number of clusters. When the number of clusters reduces, the size of the clusters enlarges and the trace of between scatter matrix decreases. When the rate of decreasing diminishes, the process stops.

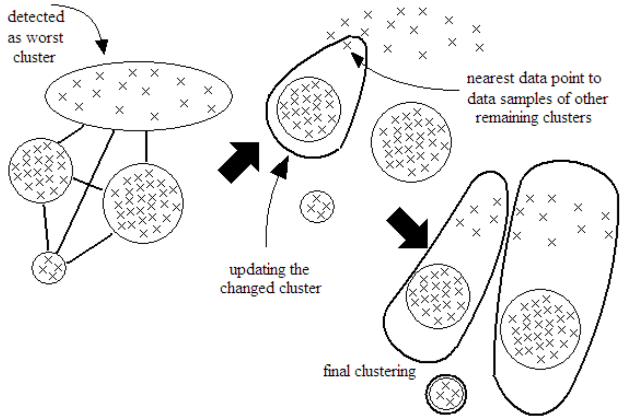
detected as worst cluster
nearest data point to data samples of other remaining clusters
updating the changed cluster
final clustering

Fig. 1 One step of the proposed algorithm.

2.5. Method for choosing σ

One of the main issues of the proposed algorithm is choosing σ in equation (9). 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. A simple method for estimating σ is defined by the following equation [14].

$$\sigma = \min\left\{\frac{\sqrt{Var(x\_Dimension)}\times 1.06}{\sqrt{N}},\ldots,\frac{\sqrt{Var(z\_Dimension)}\times 1.06}{\sqrt{N}}\right\} \tag{19}$$

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

3. Experiments Results

For evaluating the efficiency of the proposed algorithm for clustering, some experiments are done on the synthetic data.

3.1. Standard synthetic data

Figures 2 and 3 show clustering's result on some standard datasets. Since ordinary algorithms like C-means [15] and Gustafson-Kessel [15] are designed for mass clusters, they are not able to correctly cluster shell and linear clusters. Gustafson-Kessel algorithm is an improved version of Fuzzy C-means clustering [15] and can detect elliptic shape mass clusters.

In Figures 2 and 3 the upper left picture shows the data samples and the upper right picture shows clustering by C-means clustering and the bottom left picture shows clustering by Gustafson-Kessel algorithm and the bottom right picture shows the proposed clustering algorithm. As it can be seen, the proposed algorithm is able to detect line and shell prototypes properly. Each cluster is shown using a different symbol.

To ensure how the final clustering is achieved, in Figure 4 the quantity of between scatter matrix is plotted for all steps of the proposed algorithm on the dataset presented in Figure 3. Since the decreasing rate in between scatter matrix flattens between two and three clusters, the final number of clusters is set to three clusters.

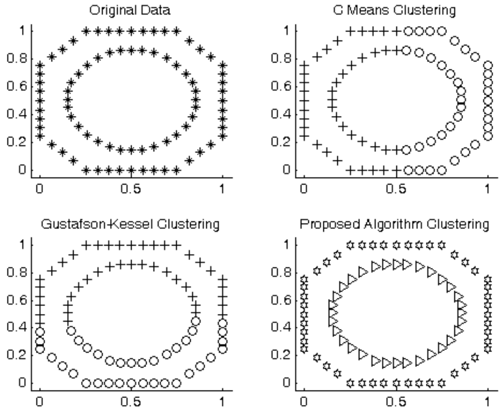

Fig. 2 Clustering a dataset with two shell clusters.

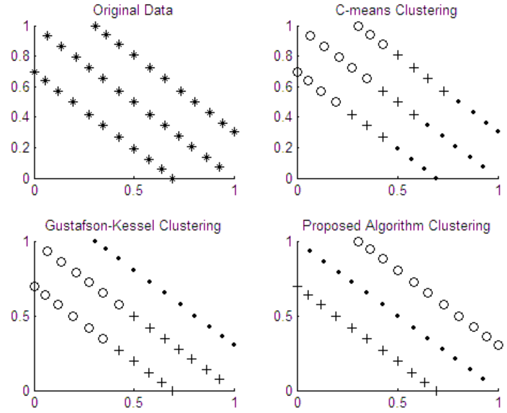

Fig. 3 Clustering a dataset with three linear clusters.

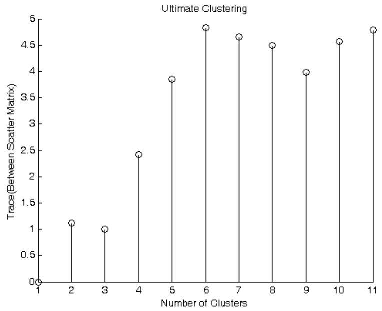

Fig. 4 Between cluster entropy for the dataset presented in Fig. 3.

3.2. Centralized synthetic data

Figures 5 and 6 show clustering for datasets with a huge number of data samples and with centralized clusters (mass clusters). Clustering results for the proposed algorithm are compared with C-means clustering and Gustafson-Kessel algorithm. Figure 5 resembles chromosomes and shows the proposed algorithm successful clustering. Figure 7 shows the between scatter matrix for all of the steps of the proposed algorithm on the dataset of Figure 6. From Figure 7, it can be realized that four clusters is an appropriate number of clusters.

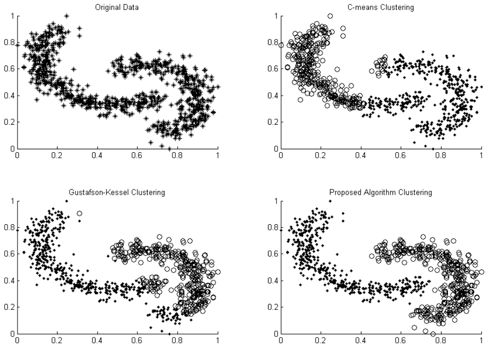

Fig. 5 Clustering a dataset with two centralized clusters.

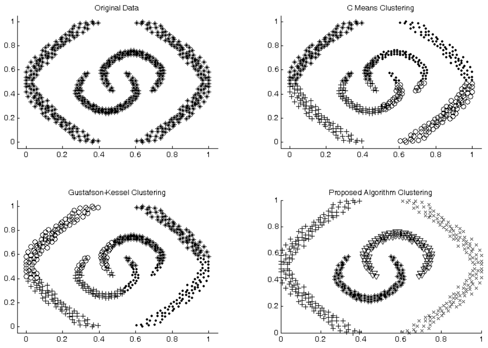

Fig. 6 Clustering a dataset with four centralized clusters.

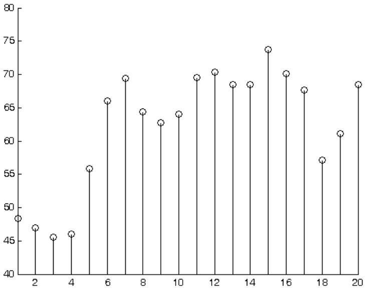

Fig. 7 Between cluster entropy for the dataset presented in Fig. 6.

Conclusion

In this paper, a new top-down hierarchical method is proposed for data clustering based on information theory and Renyi's entropy. In each iteration, the proposed algorithm uses between cluster entropy and within scatter matrix to eliminate a cluster named worst cluster. It is shown that the proposed algorithm detects different structures of clusters (mass, shell and linear clusters) and works better than 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 method.

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, Vol. 2, Page(s): 125-137, 2001.

[6] C. E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, Vol. 27, July, October, 1948.

[7] G. A. Carpenter, and S. Grossberg, "A Massively Parallel Architecture for Self-Organizing Neural Pattern Recognition Machine," Computer Vision, Graphics, and Image Processing, Vol. 37, 1987.

[8] S. Watanabe, Pattern recognition: human and mechanical (New York: John Wiley & Sons, 1985).

[9] E. Gokcay, and J. C. Principe, "Information Theoretic Clustering," IEEE Transaction on PAMI, Vol. 24, No. 2, Page(s):158 – 171, February 2002.

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

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

[12] 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.

[13] 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.

[14] B. W. Silverman, Density estimation for statistics and data analysis (Chapman and Hall, 1986).

[15] S. Theodoridis and K. Koutroumbas, Pattern recognition (Academic Press, 1999).