Finding the Number of Clusters in a Dataset Using an Information Theoretic Hierarchical Algorithm

M. Aghagolzadeh, H. Soltanian-Zadeh, B. N. Araabi
Control and Intelligent Processing Center of Excellence
Department of Electrical and Computer Engineering
University of Tehran, Tehran 14395-515, Iran
m.golzadeh@ece.ut.ac.ir, hszadeh@ut.ac.ir, araabi@ut.ac.ir

A. Aghagolzadeh
Faculty of Electrical and Computer Engineering
University of Tabriz
Tabriz, Iran
aghagol@tabrizu.ac.ir

Abstract—One of the most challenging problems of clustering is detecting the exact number of clusters in a dataset. Most of the previous methods, presented to solve this problem, estimate the number of clusters with model based algorithms, which are not able to detect all types of clusters and also face a problem in detecting coupled clusters in a dataset. In this paper we propose a new method for finding the number of clusters in a dataset utilizing information theory and a top-down hierarchical clustering algorithm. The algorithm starts from a large number of clusters and reduces one cluster in any iteration and then allocates its data points to the remaining clusters. Finally, by measuring Information Potential, the exact number of clusters in a desired dataset is detected. Our method shows high capability and stability in detecting the number of clusters even in complex datasets, as it is computational efficient too. We show the effectiveness of the proposed method by experimenting on several artificial and real datasets and comparing its results with two recently developed methods for finding the number of clusters in a dataset. The comparisons show superiority of the proposed method

I. INTRODUCTION

Dataset classification depends on the methods used for classifying, which differs in the similarity and dissimilarity measurement tool; consequently there is no concept as a single correct classification. There have been investigations and attempts to define the optimal classification and the optimal number of clusters, known as cluster analysis (CA). CA seeks to identify a set of groups which minimize within-group variations and maximize between-group variations.

The first step in any CA algorithm is to establish a similarity measurement. In the proposed method, an information theoretic similarity measurement is used based on the Renyi's definition of entropy [1]. Different kinds of entropy are measured after each step in a proposed hierarchical algorithm similar to the agglomerative clustering but with main differences in its structure. This algorithm, called top-down hierarchical algorithm, does not combine exactly two clusters to generate a new one; instead, a cluster which is detected as the most improper cluster by the entropy measurements is blown up and its elements are individually allocated to the remaining clusters. Therefore a dendrogram like the one in divisive or agglomerative hierarchical clustering can not be drawn. At last, the final number of clusters is selected upon high variations in Information Potential measured at each step. To reduce computational complexity in a hierarchical algorithm, especially in clustering huge datasets, the proposed algorithm applies fuzzy C-means clustering as the initial clustering to create a dataset with multiple tiny clusters.

This research was supported by Iranian Telecommunication Research Center (ITRC), Tehran, Iran.

II. THE TOP-DOWN HIERARCHICAL ALGORITHM

The proposed algorithm has two main steps: a) finding the most improper cluster, known as worst cluster, among the existing clusters and splitting it into its constructing data points. b) Allocating the freed data points of the worst cluster independently to the remaining clusters.

A. Finding the worst cluster

In the proposed method an estimation of entropy is utilized to detect the worst cluster. Various kinds of entropy can be defined between data points of a dataset like Total Dataset Entropy (TDE), Between Cluster Entropy (BCE), Within Cluster Entropy (WCE) and Between Cluster Entropy in Absence of a Cluster (BCEAC). Any of the above entropies are a summation of entropy calculated between each pairs of data points. Information potential between pairs of data points shown by $I_{i,j}$ is computed by the following.

$$I_{i,j} = \beta \times G\left(x_i - x_j, \sigma^2\right) \tag{1}$$

Where β is a constant value for all information datums and limits the summation of information datums between an [0 1] interval and therefore all entropy values will be positive. TDE is constant for a dataset and is not altered by changing the cluster labels or eliminating any cluster.

$$TDE = -\log\left(\sum_{i=1}^{N}\sum_{j=1}^{N} I_{i,j}\right) \tag{2}$$

BCE is the entropy between any pairs of data points from different clusters, first proposed by Gockay et al [2]. It is clear that this entropy is not changed until the cluster labels are altered, therefore it is not changed during any step.

$$BCE = -\log\left(\sum_{i:\mathbf{x}_i \in C_k} \sum_{\substack{j:\mathbf{x}_j \in C_l \\ l \neq k}} I_{i,j} \right) \quad (3)$$

Equal to the numbers of clusters, there are quantities for WCE and BCEAC in a dataset, so if a dataset has K clusters at a step, K quantities for WCE and K quantities for BCEAC are defined by the following.

$$WCE_k = -\log\left(\sum_{i:\mathbf{x}_i \in C_l} \sum_{j:\mathbf{x}_j \in C_l} I_{i,j} \right) \quad (4)$$

$$BCEAC_k = -\log\left(\sum_{i:\mathbf{x}_i \in C_f} \sum_{\substack{j:\mathbf{x}_j \in C_l \\ f \neq l,\ f,l \neq k}} I_{i,j} \right) \quad (5)$$

Two main relationships exist between the above identified entropies which are shown by (6) and (7).

$$TDE = \sum_{k=1}^{K} WCE_k + BCE \quad (6)$$

$$BCE = BCEAC_k - \log\left(\sum_{i:\mathbf{x}_i \in C_f} \sum_{\substack{j:\mathbf{x}_j \in C_k \\ f \neq k}} I_{i,j} \right) \quad (7)$$

The effect of different clusters on BCE is not equal. Since the best clustering is achieved when clusters have the maximum between dissimilarity, the more the BCE is increased, the better the clustering is. The worst cluster increases the BCE less than other clusters. To detect this cluster, BCE is computed in the absence of any cluster (BCEAC). Because BCE is constant during any iteration, it can be simply shown from (7) that the cluster with maximum BCEAC has the minimum affect on BCE (logarithmic term of (7)). Therefore the cluster with maximum BCEAC is detected as the worst cluster. After finding the worst cluster, it is vanished and its data points are freed by removing labels.

B. Allocating the freed data points

In the proposed hierarchical clustering algorithm, a freed data point is allocated based on the minimum Euclidian distance to one of the remaining clusters. The order of choosing free data points of the eliminated cluster for allocation 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. In the proposed algorithm, the arrangement of the free data points is based on the nearest free data point to data points of the rest of the clusters. This means that freed data points are allocated in a low to high "minimum Euclidean distance" order. The changed cluster after allocating any data point will be updated and this process will be repeated until the last freed data point is allocated. This method decreases the probability of trapping into local minima and stables the clustering.

C. Initial Clustering

One of the important factors of the proposed algorithm is its initial clustering, which benefits from multi-resolution concept, and the final clustering highly depends on it. In applying the initial clustering two points are considered: stability assurance and computational complexity decreasing. In the proposed algorithm, fuzzy C-means clustering is used as initial clustering. This method guarantees the convergence and transfers data points to plenty of clusters, each with few data points.

III. FINDING THE EXACT NUMBER OF CLUSTERS

Several methods for finding the number of clusters were developed for a specific problem or special type of clusters, for example Gaussian-distributed mass clusters. These methods are often model based which estimate model parameters, and also they are not able to detect the real number of clusters in datasets containing complex clusters and clusters that are coupled or near together. To extract data structures further than the second order statistics information measurement is utilized in the proposed method.

Final clustering is chosen when the within cluster similarity is maximized and between cluster similarity is minimized. As we know decreasing between cluster similarity is the same as increasing BCE, therefore in final clustering BCE is minimized. Because of the logarithmic relation between Information Potential and entropy, minimizing BCE is equal to maximizing information potential and vice versa. Therefore, for finding the final clustering, Information Potential quantities are utilized. In this method Between Information Potential in Absence of a cluster (BIPAC) and Within Information Potential (WIP) are utilized. Experiments show that BIPAC of the worst cluster multiplied by $G = N_{WC}^2$, where N_{WC} is the number of data points in the worst cluster and WIP of the worst cluster multiplied by $G = N_{WC}^2$ are useful tools to find the number of clusters in a dataset. Therefore the following equations can be written for modified BIPAC and modified WIP.

$$MBIPAC = N_{WC}^2 \left(\sum_{i:\mathbf{x}_i \in C_f} \sum_{\substack{j:\mathbf{x}_j \in C_l \\ f \neq l \quad f,l \neq WC}} I_{i,j} \right) \quad (8)$$

$$MWIP = N_{WC}^2 \left(\sum_{i:\mathbf{x}_i \in C_{WC}} \sum_{j:\mathbf{x}_j \in C_{WC}} I_{i,j} \right) \quad (9)$$

To show how our method uses the modified BIPAC and the modified WIP for finding the final clustering, a dataset containing four distinctive Gaussian distributed clusters is illustrated in Fig 1.a. Fig 1.b plots the modified BIPAC and the modified WIP based on the number of clusters, where the algorithms start clustering the desired dataset from a high number of clusters and ends to two clusters. By this example it can be seen that the final number of clusters is chosen four clusters where the modified BIPAC suddenly decreases, and the modified WIP has a sudden increase. Final clustering is detected where the two sudden variations in the modified BIPAC and the modified WIP happen in a same location. Based on these ideas, a function named CA function is proposed to find the final number of clusters.

$$CA_c = \frac{MWIP_c}{MWIP_{c+1}} \left[2.MBIPAC_c - MBIPAC_{c+1} - MBIPAC_{c-1} \right]$$

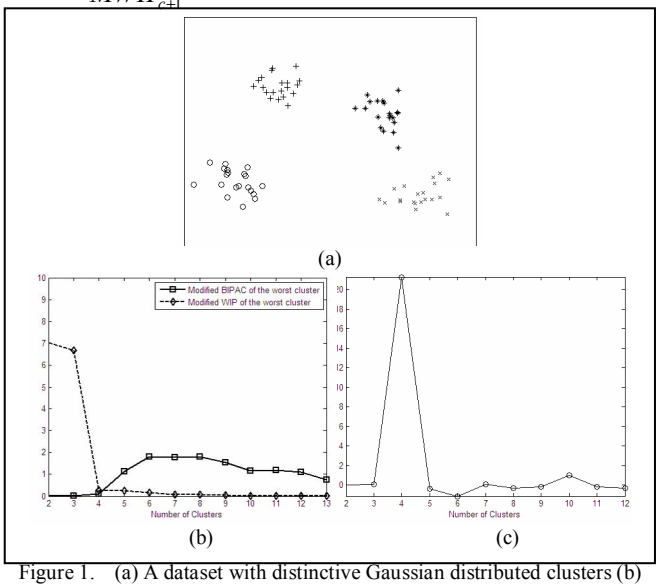

Figure 1. (a) A dataset with distinctive Gaussian distributed clusters (b) Modified BIPAC and Modified WIP (c) CA function.

Fig 1.c plots the CA function for the desired dataset. The maximum of CA function decides the final number of clusters. To show the ability of our method in estimating the number of clusters in a dataset containing complicated structures, a dataset is illustrated in Fig 2.a and the modified BIPAC and the modified WIP are drawn in Fig 2.b. Fig 2.c shows the CA function and as it can be seen, four clusters is selected as the number of clusters in this dataset.

IV. EXPERIMENT RESULTS

For evaluating the efficiency of the proposed method for finding the exact number of clusters in a dataset, some experiments are done on artificial datasets and the results are compared with two recently proposed methods for finding the number of clusters. The first method proposed by Sugar and James [3] is a non-hierarchical method for CA, developed on the definition of rate distortion theory. In [3], a comparison is done with five non-hierarchical CA algorithms which show superiority of this algorithm. This method applies C-means clustering on a dataset for $K = 1,2,\ldots,N$ clusters and computes the Mahalabonis distance between data points and any cluster prototypes by the following.

$$d_K = \frac{1}{p} \min_{C_1,\ldots,C_K} E\left[(X - C_x)^T (X - C_x) \right] \quad (11)$$

In (11), p is the dataset dimension and C_x is the center or prototype of cluster x. The method proposed by Sugar and James estimates the number of clusters by the following:

$$N = \max_K \left\{ d_K^{-\frac{p}{2}} - d_{K-1}^{-\frac{p}{2}} \right\} \quad (12)$$

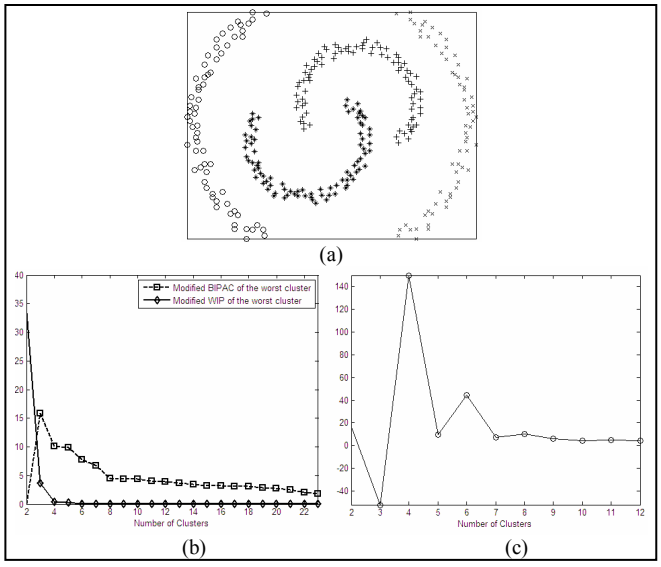

Figure 2. (a) A dataset with complicated structure (b) Modified BIPAC and Modified WIP (c) CA function.

The second method is proposed by Jenssen et al [4] which is a hierarchical clustering algorithm that uses maximum differential BCE to find the number of clusters in a dataset by the following:

$$k = \max_k \{ BCE_2 - BCE_1, \ldots, BCE_k - BCE_{k-1}, \ldots \} \quad (13)$$

To evaluate the ability to detect the exact number of clusters in a dataset including closed clusters, an artificial dataset is used which contains nine Gaussian-distributed mass clusters where all data points are normalized between an [0 1] interval. The clusters in this dataset are localized so that a method could simply mistake by selecting three clusters as the number of clusters instead of nine exact clusters. The variances of the Gaussian distributed clusters in the three datasets shown in Fig 3 are set to 0.02, 0.04 and 0.06, respectively. Table 1 shows the estimated number of clusters by the proposed method compared with the Sugar and James method [3] and Jenssen et al [4] method.

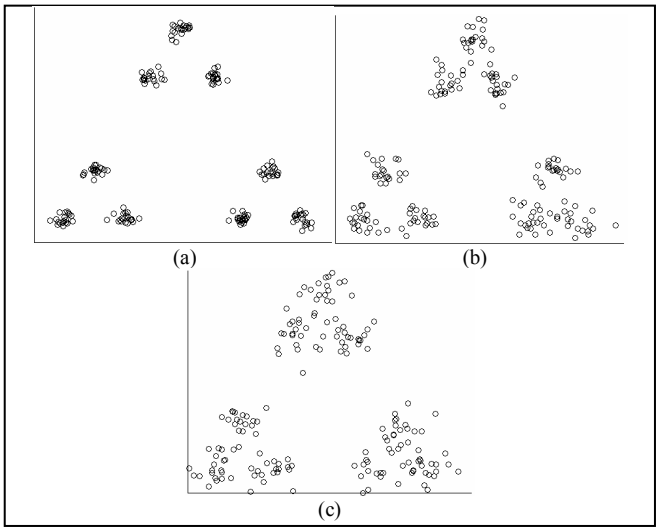

Figure 3. Datasets with nine Gaussian-distributed mass clusters, each with (a) variance = 0.02, (b) variance = 0.04, (c) variance = 0.06

TABLE I. ESTIMATED NUMBER OF CLUSTERS FOR DATASETS OF FIG 3

Simulation Dataset	Method	Estimated Clusters
Gaussian distributed mass clusters with var = 0.02	Sugar and James [3]	(Y=1,1.5) 9 Clusters
	Jenssen et al [4]	3 Clusters
	Proposed method	9 Clusters
Gaussian distributed mass clusters with var = 0.04	Sugar and James [3]	(Y=1,1.5) 3 Clusters
	Jenssen et al [4]	3 Clusters
	Proposed method	9 Clusters
Gaussian distributed mass clusters with var = 0.06	Sugar and James [3]	(Y=1,1.5) 3 Clusters
	Jenssen et al [4]	3 Clusters
	Proposed method	9 Clusters

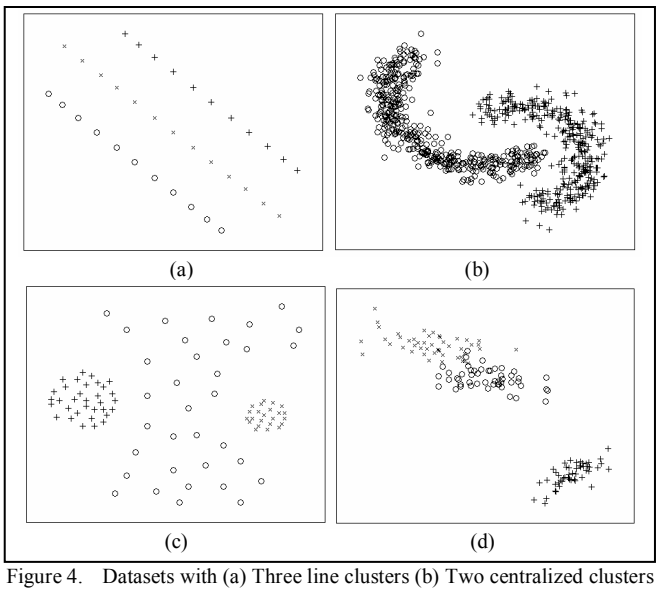

Figure 4. Datasets with (a) Three line clusters (b) Two centralized clusters (c) Two centralized and one regionalized clusters (d) Iris Dataset.

Fig 4 shows a wide variety of datasets and the estimated number of detected clusters by the proposed method is compared with the other CA methods in Table 2.

V. CONCLUSION

A new method to detect the exact number of clusters in a dataset based on information theory and hierarchical clustering was presented. This method eliminates one cluster called the worst cluster at each step and allocates its data point to the residual clusters and finds the final clustering based on the information measurements done on each step. Dispute algorithms which uses Gaussian functions in the Parzen window estimator for density estimation [2, 4], the proposed method has a good stability encountered to the kernel variations and this makes it much less sensitive to the kernel size and needless to the kernel selection methods. The ability to detect the real number of clusters, compared to the recently proposed CA methods shows the superiority and the effectiveness of this method.

TABLE II. ESTIMATED CLUSTERS FOR A VARIETY KIND OF DATASETS

Simulation Dataset	Method	Estimated Clusters
Dataset of Fig 1.a two dimensional dataset	Sugar and James [3]	4 Clusters
	Jensen et al [4]	4 Clusters
	Proposed method	4 Clusters
Dataset of Fig 2.a two dimensional dataset	Sugar and James [3]	(Y=1) 2 Clusters (Y=2) 2 or 3 Clusters
	Jensen et al [4]	4 Clusters
	Proposed method	4 Clusters
Dataset of Fig 4.a two dimensional dataset	Sugar and James [3]	(Y=1,1.5,2) 2 Clusters
	Jensen et al [4]	4 or 5 clusters
	Proposed method	3 Clusters
Dataset of Fig 4.b two dimensional dataset	Sugar and James [3]	(Y=1,1.5,2) 2 Clusters
	Jensen et al [4]	2 Clusters
	Proposed method	2 Clusters
Dataset of Fig 4.c two dimensional dataset	Sugar and James [3]	(Y=1,1.5,2) 2 Clusters
	Jensen et al [4]	5 Clusters
	Proposed method	3 Clusters
Iris Dataset (Fig 4.d) four dimensional dataset	Sugar and James [3]	(Y=3) 2 or 3Clusters (Y=2) 2 clusters
	Jensen et al [4]	2 Clusters
	Proposed method	3 Clusters

REFERENCES

[1] 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.

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

[3] C. A. Sugar and G. M. James, "Finding the Number of clusters in a Dataset: An Information-Theoretic Approach," Journal of the American Statistical Association, Vol. 98, No. 463, Page(s):750 – 763, September 2003.

[4] 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

[5] E. Anderson, "The Irises of the Gaspe peninsula," Bulletin of the American Iris Society 59, 2-5 (1935).

[6] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annuals of Eugenics 7, 179-188 (1936).