A HYBRID APPROACH FOR SOLVING THE CLUSTER VALIDITY PROBLEM

Chong, A.[1], Gedeon, T.D.[1], Koczy, L.T.[1,2]

[1]School of Information Technology
Murdoch University
South Street, Murdoch, WA, 6150 Australia

[2]Department of Telecom & Telematics
Budapest University of Technology and Economics

Abstract: A hybrid approach for solving the cluster validity problem is proposed. The proposed technique uses a cluster validity index in conjunction with a merging index to find the optimal number of clusters in a set of data. The technique does not place any restriction on the fuzzy clustering technique and the cluster validity index used. This paper examines the use of Fuzzy c-Means clustering and the validity index proposed by Fukuyama and Sugeno. Experiments are carried out to verify the effectiveness of the proposed technique. It is shown in this paper that the proposed technique is more reliable than Fukuyama and Sugeno's original validity index.

1 Introduction

Fuzzy clustering plays an important role in a wide range of areas including image processing and fuzzy system modelling. Given a set of data, clustering techniques partition the data into several groups such that the degree of association is strong within one group and weak for data in different groups. Classical clustering techniques result in crisp partitions where each data can belong to only one partition. Fuzzy clustering extends this idea to allow data to belong to more than one group. The resulting partition is therefore a fuzzy partition. Each cluster is associated with a membership function that expresses the degree to which individual data belongs to the cluster.

Of all fuzzy clustering methods, Fuzzy c-Means clustering (FCMC) remains predominant in the literature [1]. The algorithm for FCMC will be discussed in section 2.0. The FCMC algorithm relies on the user to specify the number of clusters present in the set of data to be clustered. Therefore, the use of FCMC is not suitable in situations where the number of clusters in the dataset is not known in advance. Since the introduction of FCMC, a significant amount of work has been carried out on finding the optimal number of clusters in a set of data. The problem is known as the cluster validity problem. The number of clusters in a dataset is often determined by means of a criterion, known as the cluster validity criterion. One of the most widely used criteria was proposed by Fukuyama and Sugeno. [2] (see section 3.0). The criterion is formulated using a function of the within and between group variances. As the number of clusters increases, the function exhibits a monotonic behavior. The optimal number of clusters is found when a local minimal is reached.

FCMC, when used in conjunction with Fukuyama and Sugeno's validity index (FS) works well if the clusters in the given set of data are distinct from one another. In certain situations, when there exists clusters that are close to one another, the optimal number of clusters produced by the FS index can be unreasonable. Typically, an unreasonably large number of clusters is obtained. Also, in some concrete cases, a dataset with only one cluster may be encountered. Whereas to perform clustering such a dataset is trivial, to acknowledge the fact that the dataset has one individual cluster is necessary. The use of a cluster validity criterion is inappropriate in such cases. This is explained as follows. Given a cluster validity index formulated by some function f, the optimal number of cluster c is found by solving:

$$\min_{2\le c\le n-1} f(c) \qquad \text{Eqn 1.1}$$

where n is the number of data. From eqn 1, we observe that $c = 1$ is never possible. Therefore, most of the cluster validity indices will fail if there is only one cluster in the dataset.

To tackle the above problems, we propose a hybrid approach in finding the optimal number of clusters c. The technique has two steps. In the first step, a cluster validity index is used to estimate a rough estimation of the optimal number of clusters. The number is later refined in the second step by means of a merging index. This paper is organized as follows. Section 2 discusses the Fuzzy c-Means Clustering algorithm. Section 3 presents an overview of the FS validity index and discusses the problems with the index. The proposed technique is explained in sections 4 and 5. Section 6 reports on the experiments and results. The conclusion is presented in section 7.

2 Fuzzy c-Means Clustering

Given a set of data, Fuzzy c-Means clustering (FCMC) performs clustering by iteratively searching for a set of fuzzy partitions and the associated cluster centers that represent the structure of the data as best as possible. The FCMC algorithm relies on the user to specify the number of clusters present in the set of data to be clustered. Given the number of clusters c, FCMC

0-7803-7503-3/02/$17.00 ©2002 IEEE

partitions the data $X = \{x_1, x_2, \ldots, x_n\}$ into c fuzzy partitions by minimizing the within group sum of squared error objective function as follows (eqn 2.1).

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(U_{ik})^m \| x_k - v_i \|^2, \quad 1 \le m \le \infty$$

Eqn 2.1

where $J_m(U,V)$ is the sum of squared error for the set of fuzzy clusters represented by the membership matrix U, and the associated set of cluster centers V. $\|.\|$ is some inner product-induced norm. In the formula, $\|x_k - v_i\|^2$ represents the distance between the data x_k and the cluster center v_i. The squared error is used as a performance index that measures the weighted sum of distances between cluster centers and elements in the corresponding fuzzy clusters. The number m governs the influence of membership grades in the performance index. The partition becomes fuzzier with increasing m and it is proven that the FCMC algorithm converges for any $m \in (1,\infty)$ [1]. The necessary conditions for eqn 2.1 to reach its minimum are

$$U_{ik} = \left(\sum_{j=1}^{c} \left(\frac{\| x_k - v_i \|}{\| x_k - v_j \|} \right)^{2/(m-1)} \right)^{-1} \qquad \forall i, \forall k$$

Eqn 2.2

And

$$v_i = \frac{\sum_{k=1}^{n} (U_{ik})^m x_k}{\sum_{k=1}^{n} (U_{ik})^m}$$

Eqn 2.3

In each iteration of the FCMC algorithm, matrix U is computed using eqn 2.2 and the associated cluster centers are computed as eqn 2.3. This is followed by computing the square error in eqn 2.1. The algorithm stops when either the error is below a certain tolerance value or its improvement over the previous iteration is below a certain threshold.

3 Number of Clusters Determination with the SC Technique

Fukuyama and Sugeno [2] proposed the following cluster validity index (FS) for choosing the number of clusters for fuzzy c-means clustering:

$$S(c) = \sum_{k=1}^{n}\sum_{i=1}^{c}(U_{ik})^m (\| x_k - v_i \|^2 - \| v_i - \bar{x} \|^2) 2 < c < n$$

Eqn 3.1

where n is the number of data points to be clustered; c is the number of clusters; x_k is the k^{th} data, $\bar{x}$ is the average of data; v_i is the i^{th} cluster center; U_{ik} is the membership degree of the k^{th} data with respect to the i^{th} cluster and m is the fuzzy exponent as described in section 2.0. The terms $\| x_k - v_i \|$ and $\| vi - \bar{x} \|$ represent the variance in each cluster and variance between clusters respectively. Therefore, the optimal number of clusters is found by minimizing the distance between data to the corresponding cluster center and maximizes the distance between different cluster centers.

Theoretically, the number of cluster c is found by solving eqn 1.1. However, it is computationally intensive to perform Fuzzy c-Means clustering with $c = 2 \ldots n-1$. Fortunately, eqn 3.1 exhibits a monotonic behavior as c increases [2]. It is often sufficient to determine c so that $S(c)$ reaches a local minimum as c increases. This monotonic behavior distinguishes the FS index from most of the cluster validity indexes in the literature.

The FS index works well when the clusters are highly separable (e.g. non-overlapping clusters). When cluster centers are sufficiently close to one another, the between group variance plays a less important role in eqn 3.1. This is because the term $\| vi - \bar{x} \|$ is always significantly small. This can lead to the index favoring more clusters than desired. Extensive experiments have confirmed the potential of FS Index to produce a number greater than the desired number of clusters when cluster centers are close to one another. Similar results are also reported in [3].

4 Merging Index

The idea of a merging index comes from the hierarchical fuzzy clustering techniques (HFC) literature. HFC often pursue clustering with a bottom-up approach. Two or more data points are merged to form clusters based on some criterion. Subtractive clustering [4] is one of the well-known techniques that uses such a bottom-up approach. The algorithm is unfortunately sensitive to input parameters. In this paper, we adapt the criterion in [4] to form a merging index.

Consider merging two cluster centers at a time, v_i and v_j. Figure 4.1a – c shows the situations where the two centers should be merged whereas figure 4.1d – e shows two centers that should not be merged. The decision on whether to merge the centers can be made based on the middle point $v_m = (v_i+v_j)/2$. The following observation can be obtained from figure 4.1:

1. When two centers are in the same cluster, v_m is always located in an area denser (more data points located within the area) than either v_i or v_j
2. When two centers are in different clusters, v_m is likely to be located in the least dense area.

Following the observations, the decision on whether to merge vi and vj can be made based on the density of the area where v_i, v_j and v_m are located respectively. The density of the area where each cluster center v is situated can be computed as:

$$P(v) = \sum_{j=1}^{n} e^{-4\left\| (v - x_j) / \frac{(v_i - v_j)}{2} \right\|^2}$$

Eqn 4.1

where x_j is the j^{th} data. Using the terminology in [4], the term $(v_i-v_j)/2$ can be called the 'radius of influence'. It defines an 'area of influence' of the cluster. When examining a cluster center, points whose distance from

the center exceeds the radius of influence carry a small weight. In other words, Eqn 4.1 yields a large value only when the cluster center is located in a very dense area where the distance of points from the center is within the radius of influence. Figure 4.2 shows the 'area of influence' for three cluster centers. If $p(v_m)$ is smaller than both $p(v_i)$ and $p(v_j)$, then the centers stay un-merged. Otherwise, they should be merged.

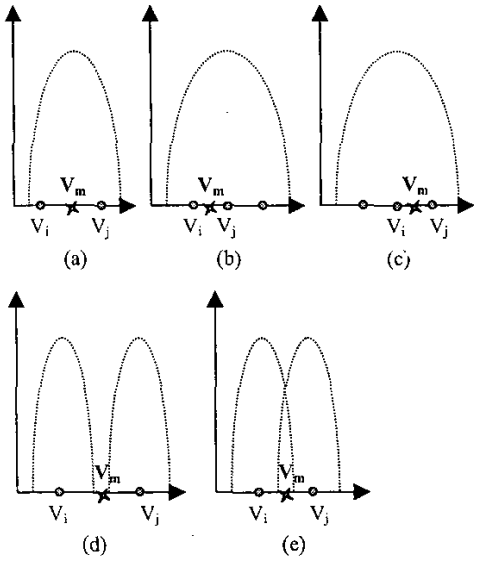

Figure 4.1. The merging of cluster center V_i and V_j

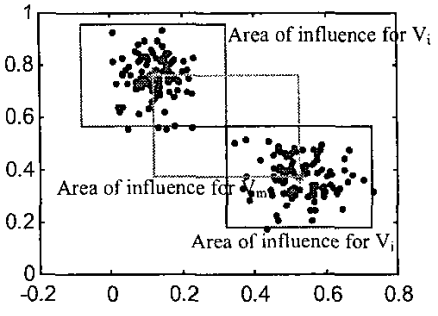

Figure 4.2. Area of influence for cluster center V_i, V_j and V_m

5 Hybrid Cluster Validity

In this section, a hybrid cluster validity approach is proposed. The technique involves the following two steps:

1. Compute a rough estimate c of the optimal number of clusters using a proper validity index. Based on experimentation results, we strongly recommend the use of the FS index because it seldom reports a number smaller than the desired number of clusters. However, we remark that the proposed technique does not place any restriction on the validity index used in this step.

2. Let $C = \{c_1, c_2, \ldots, c_n\}$ be a set of n cluster centers. Find a pair $<c_p, c_q> \in C$ that should be merged based on the technique discussed in section 4. Let c_i^d be the projection of c_i to dimension d, the pair $<c_p, c_q>$ must be an element of the set C_0:

$$c_0 = \left\{ < c_p, c_q > \middle| \begin{matrix} c_p^d > c_m^d \text{ or } c_q^d < c_m^d \text{ if } c_p^d < c_q^d \\ c_q^d > c_m^d \text{ or } c_p^d < c_m^d \text{ otherwise} \end{matrix} \right\}$$

$$c_p, c_q, c_m \in C; p \neq q \neq m; \forall d, m$$

Eqn 5.1

In other words, there should not be any other cluster center c_m that comes between c_p and c_q in the multi-dimensional space.

3. If the pair $<c_p, c_q>$ is successfully found, decrease the number of clusters by one and perform fuzzy clustering on the data. Repeat step 2 until no more clusters can be merged.
Note on implementation: For efficiency, the sets of cluster centers produced in each iteration of step 1 can be stored. After merging a pair of cluster, the program can revert to the previous set of cluster centers without performing FCMC again.

6 Experimental Results and Discussion

Experiments have been carried out to verify the effectiveness of the proposed technique using artificially generated data. In each experiment, a program is used to generate the two-dimensional data randomly based on a pre-defined number of normal clusters. The proposed technique is applied to the data. The location, size, and number of points in the clusters are all randomly generated. The details of the experiments are summarized in table 5.1.

No.	Goal	Number of datasets	Number of clusters
1	Examine the use of the propose technique in data that consist of only one cluster	100	1
2	Verify the effectiveness of the technique for data with distinct clusters	36	3
3	Verify the effectiveness of the technique when overlapping clusters (non-distinct) exist in the data	100	5

Table 5.1 Experiments

Experiment One

In this experiment, the data used has only one cluster. The number of clusters reported by the Fukuyama and Sugeno index (FS) ranges from 4 to 15. The proposed technique reported the correct number of clusters (one) for 77 out of the 100 sets of data. For the rest, the proposed technique reported 2 as the optimal number of clusters.

Experiment Two

The data used in this experiment has 3 clusters whose locations are reasonably distinct from one another. Overlapping among the clusters are kept minimal. Initially, 100 datasets were generated by random cluster generator. Each dataset was visually inspected and only dataset with sufficiently distinct clusters were kept. After the filtering process, 36 dataset were retained. FS index is able to determine the right number of clusters for all datasets. In all cases, the merging index correctly leaves all clusters un-merged.

Experiment Three

In this experiment, 100 datasets were generated by the cluster generator with the number of clusters pre-defined as 5. Since the locations of the clusters were generated at random, it was expected that some clusters may overlap significantly. Therefore, depending on the extent to which the clusters overlap, the desired number of clusters is often less than the pre-defined number of clusters (i.e. if two clusters overlap a great deal, it is desirable to treat them as one cluster). The datasets generated went through a visual inspection process to determine the desired number of clusters. The proposed technique reported the optimal number of clusters matching the desired number for 84 out of the 100 datasets. Although the number did not match the desired number for the other 16 datasets, the result was fairly close to the desired number. The FS index, on the other hand, was able to report the optimal number matching the desired number for only 2 datasets. The output of the FS index ranges from 4 to 15, which was far from the desired number for the majority cases. Table 5.2 shows the summary of the experimental results.

Exp	Desired number of cluster	Number reported by FS		Number reported by proposed technique	
		Correct	Range	Correct	Range
1	1	0	4 – 15	77	1 – 2
2	3	100	100	100	100
3	1 – 4	2	4 – 15	84	1 – 4

Table 5.2 Experimental Results Summary

7 Conclusion

A hybrid approach for the cluster validity problem has been proposed. Through extensive experimentation, the effectiveness of the approach has been validated. It is shown in the experiments that the performance of the hybrid approach is significantly better than the original validity index proposed by Fukuyama and Sugeno.

References:

[1] Bezdek, J.C., *Pattern Reconition with Fuzzy Objective Function Algorithms*. 1981, New York: Plenum Press.

[2] Fukuyama, Y. and Sugeno, M. *A new method of choosing the number of clusters for fuzzy c-means method.* in *Proceedings of the 5th Fuzzy System Symposium*. 1989.

[3] Yang, M.S. and Wu, K.L. *A New Validity Index For Fuzzy Clustering.* in *The 10th IEEE International Conference on Fuzzy Systems*. 2001. Melbourne, Australia.

[4] Chiu, S.L., *Fuzzy Model Identification Based on Cluster Estimation.* J. Intelligent and Fuzzy Systems, 1994. 2: p. 267-278.