Implementation of some cluster validity methods for fuzzy cluster analysis

Erind Bedalli
Department of Mathematics and Informatics,
University of Elbasan
Computer Engineering Department
Epoka University
Tirana, Albania
ebedalli@epoka.edu.al

Ilia Ninka
Department of Informatics
University of Tirana
Tirana, Albania
ilia.ninka@yahoo.com

***Abstract*—Cluster analysis is an important tool in the exploration of large collections of data, revealing patterns and significant correlations in the data. The fuzzy approach to the clustering problem enhances the modeling capability as the results are expressed in soft clusters (instead of crisp clusters), where the data points may have partial memberships in several clusters. In this paper we will discuss about the most used fuzzy cluster analysis techniques and we will address an important issue: finding the optimal number of clusters. This problem is known as the cluster validity problem and is one of the most challenging aspects of fuzzy and classical cluster analysis. We will describe several methods and we will combine and compare them on several synthetic data sets.**

Keywords— fuzzy cluster analysis; cluster validity; validation measures;

I. INTRODUCTION

Clustering is an important research area in a variety of disciplines like business, pattern recognition, machine learning, biology, cognitive sciences etc [1, 2]. In business applications, clustering assists the market specialists and policy makers in discovering target groups, in characterization of customers, in new product profiling etc. In biology, it is utilized in grouping genes with similar behaviors, in finding patterns inside populations of livings etc. In pattern recognition and machine learning, besides the useful information it provides itself, clustering frequently serves as a pre-processing stage for several other algorithms [3]. The central idea about clustering is the distribution of the data points into groups (clusters) such that the data points inside the same cluster are more similar to each other than to the data points in other clusters.

Many clustering algorithms have been developed and they may be categorized into several sub-categories [1]. Firstly we would distinguish between crisp and fuzzy clustering. In the crisp clustering the data points are distributed in clusters where each data point belong to exactly one of the clusters, while in the fuzzy clustering the data points have partial memberships into several clusters. Another categorization would be between partition and hierarchical clustering. The partition clustering gives as result a single partition of the data while the hierarchical clustering generates a tree of clusters where the data points are distributed into smaller clusters at each level of the hierarchy. Furthermore partition clustering techniques may be categorized into other sub-categories like: square error clustering, graph theoretic clustering, mixture resolving clustering etc.

One of the most important problems which aims to complete the cluster analysis is the evaluation of the quality of the obtained clusters. So our crucial question is: are the obtained clusters optimal? This problem is known as the cluster validity problem. There are several methods for assessing the fuzzy clusters validity; some of them rely merely on the characteristics of the fuzzy membership values to assess the clusters, while others rely also on the structure of data [1, 2]. In this paper we will discuss some of the most important methods for cluster validity evaluation like: the partition coefficient, the partition index, the partition entropy, the partition index, the separation index, the Xie-Beni index, the Fukuyama-Sugeno index etc. We will implement them on several synthetic data and also perform some combinations of them.

II. FUZZY CLUSTER ANALYSIS

The fuzzy clustering algorithms provide more flexibility and a richer semantics of the data compared to the classical (hard) clustering algorithms as they allow partial membership (gradual membership) of the data points in the clusters. So a data point may be an element of several clusters in the same time with different membership values.

The result of a fuzzy clustering algorithm is typically represented as a matrix $M = [\mu_{i,j}]$ of dimensions $c \ x \ N$. Here $\mu_{i,j}$ denotes the membership in the $i-th$ cluster of the $j-th$ data point. The membership values in the matrix satisfy [2]:

$0 \leq \mu_{i,j} \leq 1$ for every $i \in \{1,2,\dots,c\}$ and $j \in \{1,2,\dots,N\}$

$\sum_{i=1}^{c} \mu_{i,j} = 1$ for every $j \in \{1,2,\dots,N\}$

$0 < \sum_{j=1}^{N} \mu_{i,j} < N$ for every $i \in \{1,2,\dots,c\}$

As we see from the second condition, it is required that the total sum of the memberships of each data point into the obtained clusters must be equal to one. And as we see from the

third condition it required that the total sum of the memberships for each cluster must be greater than zero (so each cluster must have at least one element whose membership value in that cluster is greater than zero) and smaller than N.

The most fundamental algorithm of fuzzy cluster analysis is the fuzzy C-means algorithm. This algorithm operates in a supervised way distributing the data points into pre-defined clusters with partial memberships. This algorithm takes as inputs:

a. The number of clusters: c, (s. t. $1 < c < n$).

b. The distance metrics (dissimilarity metrics).

c. The fuzzy exponent φ, (such that φ > 1).

d. The scale of precision ε (for example ε = 0.001 would give good approximations).

There are several options for the distance metrics which evaluates the distance among the data points. Considering two data points $U = \{u_1, u_2, \dots, u_n\}$ and $V = \{v_1, v_2, \dots, v_n\}$ the distance between them according to some of the most widely used distance metrics is expressed as:

a. The Euclidian distance

$$d(U,V) = \sqrt{\sum_{k=1}^{n}(u_k - v_k)^2}$$

b. The Manhattan distance:

$$d(U,V) = \sum_{k=1}^{n}|u_k - v_k|$$

c. The Max distance:

$$d(U,V) = \max_{k\in\{1,2,\dots,n\}}|u_k - v_k|$$

d. The Minkowski distance:

$$d(U,V) = \sqrt[m]{\sum_{k=1}^{n}(u_k - v_k)^m}$$

As it can be noticed, Euclidian, Manhattan and Max distances are particular cases of the Minkowski distance respectively for $m = 2, m = 1$ and $m = \infty$.

e. The Pearson correlation distance:

$$d(U,V) = \frac{\sum(u\cdot v) - n\bar{U}\bar{V}}{\sqrt{(\sum u^2 - nU^2)(\sum v^2 - nV^2)}}$$

Two other important algorithms, the Gustafson-Kessel and Gath-Geva algorithms are developed as extensions of the fuzzy c-means algorithm employing adaptive-distance metrics to distinguish clusters with various shapes and orientations.

The fuzzy C-means algorithm may be briefly described by the given pseudo-code [1]:

1. Choose k random data points as the initial centers of the clusters ($M = \{\mu_{ik}\} = M^{(0)}$).
2. Assign i =1.
3. Evaluate the new centers $C = C^{(i)}$ using $M^{(i)-1}$ (evaluated in the previous iteration if i>1, or the initial $M^{(0)}$ if i = 1).
4. Evaluate $M = M^{(i)}$ using $C^{(i)}$ found in the previous step.
5. Compare $M^{(i)}$ to $M^{(i-1)}$. If $||M^{(i)} - M^{(i-1)}|| < e$, then terminate; else increment i and jump to 3.

III. CLUSTER VALIDATION

At the initial moment of the clustering procedure, the data points do not have labels what would point out the desired result. So we are lacking references to check and assess our obtained results. Most validation methods rely on two important characteristics: compactness and separation of the fuzzy clusters. Compactness is a quantity that evaluates the variation of the data within the same cluster. On the other hand, separation is a quantity that describes the structures among the different clusters. The primary goal of all the validation methods is to decrease the compactness and to increase the separation of the clustering. As there are several ways to express the compactness and the separation of the clusters, we will have limitations in picking some general way that would optimally describe these two quantities.

There are three main approaches for the cluster validation problem:

The first approach assumes that we have a validity measure to assess the entire fuzzy partition that we will obtain. Some of the most important validity measures are described in the next section of this paper. We guess some maximal reasonable value for the number of the clusters, let us denote it c_{max}. Then for every natural value from 2 to c_{max} we execute our fuzzy clustering algorithm. For each value we calculate the validity measure separately and at the end of this procedure we have found the number of clusters that would optimize our validity measure.

The second approach assumes that we have a validity measure to assess each cluster (separately). We guess some maximal reasonable value for the number of the clusters, let us denote it c_{max}. Then our fuzzy algorithm is executed for the value c_{max}. We compare the obtained clusters based on the validity function that we are provided. The similar clusters are merged into larger cluster, thus removing the non-appropriate clusters. This procedure is repeated in an iterative way until no non-appropriate clusters are remained.

The third approach is initialized with a large number of clusters and then it proceeds in an iterative way merging the clusters which are very similar to each other according to some condition specified in advance. This procedure is known as compatible cluster merging.

As it may be noticed in the description of the three approaches, the problem of cluster validation is not only a challenging one but also a computationally expensive one.

IV. VALIDITY MEASURES

In this section we will describe some of the most widely used validity measures. It is strongly recommended to utilize more than one of these in implementations as they do not have the capability of being decisive in all cases [1,2].

A. The partition coefficient

It is a quantity that expresses the amount of shared regions among the clusters. It is calculated as:

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{i,j}\right)^2$$

with $\mu_{i,j}$ denoting the membership in the $i-th$ cluster of the $j-th$ data point. The value satisfies the inequality $\frac{1}{c} \le PC(c) \le 1$. The disadvantages of this method are that it monotonically decreases with c and there is no direct relation to some property of the data [1,4,5]. The optimal value of c is the value that maximizes the partition coefficient.

B. The partition entropy

It is a quantity that expresses the amount of fuzziness in the clusters. It is calculated as:

$$PE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{i,j}\ln {\mu_{i,j}}^2$$

The value satisfies the inequality $0 \le PE(c) \le log_2 c$. The optimal value of c is the value that minimizes the partition entropy. Bezdek has also proved the inequality $0 \le 1 - PC(c) \le PE(c)$ for probabilistic cluster partitions [3,4].

C. The modified partition index

The previous indexes show monotonic behavior as the number of the cluster increases. A modification in the way we calculate this new index improves this behavior. The index is calculated as:

$$MPC(c) = 1 - \frac{c}{c-1}\left(1 - PC(c)\right)$$

As it may be easily noticed, the value of the index satisfy the inequality: $0 \le MPC(c) \le 1$. The optimal value of c is the value that maximizes the modified partition index [3].

D. The partition index

It is a quantity that expresses the ratio of the total compactness over the total separation of the clusters. It is calculated as:

$$PI(c) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}\left(\mu_{i,j}\right)^m\left\|x_j - v_i\right\|^2}{\sum_{j=1}^{N}\mu_{i,j}\sum_{j=1}^{c}\left\|v_j - v_i\right\|^2}$$

The partition index is appropriate to be applied when we are comparing fuzzy partitions consisting of the same number of clusters. The optimal value of c is the value that maximizes the partition index [1,5,6].

E. The separation index

It is a quantity that expresses the ratio of the total compactness over the smallest distance separation of the clusters. It is calculated as:

$$SI(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{i,j}\right)^m\left\|x_j - v_i\right\|^2}{N\min_{i,j}\left\|v_j - v_i\right\|^2}$$

The partition index is appropriate to be applied when we are comparing fuzzy partitions consisting of the same number of clusters. The optimal value of c is the value that maximizes the separation index. [1,4,5]

F. The Xie-Beni's index

It is calculated as:

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{i,j}\right)^m\left\|x_j - v_i\right\|^2}{N\min_{i,j}\left\|x_j - v_i\right\|^2}$$

At a first appearance the Xie-Beni's index seems almost the same with the separation index, but practically the small change in the evaluation of the denominator of the expression gives a significant improvement. This handles two drawbacks of the previous methods: they don't consider all the parameters (e.g. V is not taken into consideration) and they do not completely utilize the value of X too. The optimal value of c is the value that minimizes the index.[1,5,6]

G. The Fukuyama-Sugeno index

It is calculated as:

$$FS(c) = \sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{i,j}\right)^m\left(x_j - a_i\right)^2 - \sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{i,j}\right)^m\left(a_i - \bar{a}\right)^2$$

where $\bar{a} = \left(\sum_{i=1}^{c} a_i\right)/c$. The optimal value of c is the value that maximizes the index [3].

H. The Dunn's index

It is calculated as:

$$DI(c) = \min_{i\in c}\left\{\min_{j\in c, j\neq c}\left\{\frac{\min_{x\in C_i, y\in C_j,} d(x,y)}{\max_{k\in C}\{\max_{x,y\in C} d(x,y)\}}\right\}\right\}$$

Initially this index has been devised for distinguishing well-separated clusters. The main disadvantage of the Dunn's index is scalability. When the values of c and N increase the computation of the index becomes hardly feasible [1,2].

I. The alternative Dunn's index

It is calculated as:

$$ADI(c) = \min_{i \in c} \left\{ \min_{j \in c, j \neq c} \left\{ \frac{\min_{x \in C_i, y \in C_j,} |d(y, v_j) - d(x_i, v_j)|}{\max_{k \in C} \{ \max_{x,y \in C} d(x,y) \}} \right\} \right\}$$

As it seen it is a variation of the original Dunn's index. The primary goal of this variation is to improve the computational complexity. So instead of using directly the distance $d(x,y)$ we use a smaller quantity $|d(y,v_j) - d(x_i,v_j)|$ It is obvious that the second quantity is smaller than the first one based on the triangular inequality [1,2].

J. The fuzzy hypervolume

It is one of the most frequently used validity measures. It can be considered as the volume of the fuzzy clusters and it is calculated as:

$$\Upsilon(c) = \sum_{i=1}^{c} \det(F_i)$$

So it is expressed as the sum of the determinant of the F_i matrices, where $F_i$represents the matrix [1,2,3]:

$$F_i = \frac{\sum_{j=1}^{N} \mu_{ij}^m (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{N} \mu_{ij}^m}$$

V. IMPLEMENTATION AND EXPERIMENTAL RESULTS

We have implemented the given validity measures on several synthetic data sets according to the first approach mentioned in the third section of this paper. So we have executed the fuzzy clustering algorithm several times with values of c starting from 2 up to some estimated large value of c, in our case this value was 10.

The obtained results for the data set SYNTH_1 are given in the tables 1 and 2 with the optimal values of c, according to each index being highlighted.

TABLE 1. Experimental results for data set SYNTH_1 (p1)

c	$PC(c)$	$PE(c)$	$MPC(c)$	$PI(c)$	$SI(c)$
2	0.72845	0.23655	0.608722	1.2165	0.0154
3	0.68715	0.3184	0.596856	1.3291	0.0169
4	0.6775	0.3394	0.604286	0.9346	0.0173
5	0.6711	0.3655	0.599775	0.5293	0.0127
6	0.66455	0.39395	0.59746	0.4755	0.0104
7	0.65695	0.41275	0.588875	0.3826	0.0099
8	0.65375	0.4223	0.570001	0.3881	0.0094
9	0.64165	0.4541	0.530725	0.4265	0.0102
10	0.64785	0.4510	0.4569	0.3568	0.0092

TABLE 2. Experimental results data set SYNTH_1 (p2)

c	$XB(c)$	$FS(c)$	$DI(c)$	$ADI(c)$	$FHV(c)$
2	21.5271	0.2201	0.1131	0.0024	0.8474
3	7.7585	0.8954	0.0311	0.0019	1.0251
4	5.3847	0.6210	0.0188	0.0012	1.9141
5	6.2213	0.0782	0.0225	0.0008	2.2108
6	7.8739	-0.5841	0.0265	0.0006	2.0034
7	9.3617	-1.2018	0.0179	0.0005	1.7105
8	11.1568	-1.8841	0.0199	0.0001	1.3912
9	14.7286	-2.6810	0.0270	0.0002	1.0154
10	16.8604	-3.7541	0.0192	0.0003	0.9454

The obtained results for the data set SYNTH_2 are given in the tables III and IV with the optimal values of c, according to each index being highlighted.

TABLE 3. Experimental results for data set SYNTH_2 (p1)

c	$PC(c)$	$PE(c)$	$MPC(c)$	$PI(c)$	$SI(c)$
2	0.9713	0.3171	0.8725	1.7213	0.0161
3	0.9162	0.5905	0.8743	1.8482	0.0197
4	0.9033	0.6692	0.8711	1.9265	0.0170
5	0.8948	0.7721	0.8685	1.6121	0.0104
6	0.8861	0.8904	0.8633	1.3377	0.0094
7	0.8759	0.9721	0.8553	0.9057	0.0106
8	0.8717	1.0147	0.8533	0.8137	0.0101
9	0.8555	1.1618	0.8375	0.8686	0.0109
10	0.8638	1.1471	0.8487	0.7683	0.0099

TABLE 4. Experimental results data set SYNTH_1 (p2)

c	$XB(c)$	$FS(c)$	$DI(c)$	$ADI(c)$	$FHV(c)$
2	11.1959	0.5214	0.1304	0.00594	1.5153
3	12.7646	0.7412	0.0358	0.00226	2.3875
4	14.4532	0.8954	0.0217	0.00178	2.9123
5	8.7432	0.6542	0.0260	0.00096	2.7997
6	10.9261	0.2879	0.0305	0.00081	2.7034
7	12.4278	-0.4219	0.0207	0.00064	2.6895
8	14.8412	-1.3510	0.0229	0.00052	2.2985

9	15.206	-3.0854	0.0312	0.00044	1.7730
10	15.5387	-4.1576	0.0221	0.00030	1.6674

The optimal number of clusters for the data sets according to each method may be summarized in the following table:

TABLE 5. Summary of the validation procedure with the optimal number of clusters for each case.

Validation measure	SYNTH_1	SYNTH_2
$PC(c)$	2	2
$PE(c)$	2	2
$MPC(c)$	2	3
$PI(c)$	3	4
$SI(c)$	4	2
$XB(c)$	4	5
$FS(c)$	3	4
$DI(c)$	2	2
$ADI(c)$	2	2
$FHV(c)$	5	3

VI. CONCLUSIONS

Fuzzy clustering is an unsupervised form of learning where no initial information about the data set is provided. During the fuzzy clustering procedure the data points are distributed into clusters where the patterns inside the same cluster are similar to each other and different from the patterns in the other clusters. In comparison to hard clustering, fuzzy clustering provides more flexibility as the results are obtained in soft clusters where one data point may belong simultaneously to different clusters (with different membership degrees).

The cluster validity problem is a crucial problem for the fuzzy clustering analysis. So we are interested in finding the optimal number of clusters into which the data will be decomposed. In this paper we have given an overview of the most widely used validity measures and also we have made a brief comparison of these validity measures. None of these measures is capable of being decisive by itself, so it is recommended to apply more than one method to obtain optimal results.

We have implemented these methods on two large synthetic data sets and we have observed that different validation measures may suggest different optimal number of clusters.

REFERENCES

[1] J.Abonji, B.Feil, “Cluster analysis for data mining and system identification,”, Birkhauser 2007, pp. 17-45.

[2] J. Valente De Oliviera, W. Pedrycz “Advances in fuzzy clustering and its applications”, John Wiley and Sons 2007, pp 53-66, 99-116

[3] Kwo-Lang Wu, Miin-Shen Yang, “A cluster validity index for fuzzy clustering”, Pattern Recognition Letters, Volume 26, Issue 9, July 2005

[4] M.Halkidi, Y. Batistakis, M. Vazirgiannis “On Clustering Validation Techniques”, Journal of Intelligent Information Systems, Volume 17, Issue 2-3, December 2001

[5] W.Wang, Y.Zhang, “On fuzzy cluster validity indices”, Fuzzy Sets and Systems, Volume 158, Issue 19, October 2007.

[6] M. Rezaee, B. Lelieveld, J.Reiber “A new cluster validity index for the fuzzy c-mean”, Pattern Recognition Letters, Volume 19, Issue 3-4, March 1998