Cluster Validity Methods : Part I

Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis
Department of Informatics,
Athens University of Economics & Business
Email: {mhalk, yannis, mvazirg}@aueb.gr

ABSTRACT

Clustering is an unsupervised process since there are no predefined classes and no examples that would indicate grouping properties in the data set. The majority of the clustering algorithms behave differently depending on the features of the data set and the initial assumptions for defining groups. Therefore, in most applications the resulting clustering scheme requires some sort of evaluation as regards its validity. Evaluating and assessing the results of a clustering algorithm is the main subject of *cluster validity*.

In this paper we present a review of the clustering validity and methods. More specifically, Part I of the paper discusses the cluster validity approaches based on *external* and *internal* criteria.

1. INTRODUCTION

Clustering is one of the most useful tasks in the data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data [4]. Thus, the main concern in the clustering process is to reveal the organization of patterns into "sensible" groups, which allow us to discover similarities and differences, as well as to derive useful inferences about them [6].

In the literature a wide variety of algorithms have been proposed for different applications and sizes of data sets[7, 8]. The application of an algorithm to a data set aims at, assuming that the data set offers such a clustering tendency, discovering its inherent partitions. However, the clustering process is perceived as an unsupervised process, since there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data [1]. Then, the various clustering algorithms are based on some assumptions in order to define a partitioning of a data set. As a consequence, they may behave in a different way depending on:

i) the features of the data set (geometry and density distribution of clusters) and
ii) the input parameters values.

It is obvious that a problem we face in clustering is to decide the optimal number of clusters that fits a data set. In most algorithms' experimental evaluations 2D-data sets are used in order that the reader is able to visually verify the validity of the results (i.e., how well the clustering algorithm discovered the clusters of the data set). It is clear that visualization of the data set is a crucial verification of the clustering results. In the case of large multidimensional data sets (e.g. more than three dimensions) effective visualization of the data set would be difficult. Moreover the perception of clusters using available visualization tools is a difficult task for humans that are not accustomed to higher dimensional spaces.

For instance, assume the data set in Figure 1a. It clearly consists of three clusters. However, if we consider a clustering algorithm (e.g. K-Means) with certain input parameter values (in the case of K-Means [9] the number of clusters) so as to partition the data set in four clusters, the result of clustering process would be the clustering scheme presented in Figure 1b. In our example the clustering algorithm (K-Means) found the best partitioning into four clusters. However, this is not the optimal partitioning for the considered data set. We define, here, the term "optimal" clustering scheme as the outcome of running a clustering algorithm (i.e., a partitioning) that best fits the inherent partitions of the data set. Similarly, Figure 2 presents the behaviour of the algorithm DBSCAN [3] under the consideration of different input parameter values. DBSCAN achieves to partition the data set optimally into three clusters (see Figure 2a) only under the consideration of the suitable input parameters' values (i.e., Eps=2, Nps=4). Using different input parameter values, it fails to find the optimal partitioning of the dataset (e.g. Figure 2b).

As a consequence, if the clustering algorithm parameters are assigned an improper value, the clustering method results in a partitioning scheme that is not optimal for the specific data set leading to wrong decisions. The problems of deciding the number of clusters better fitting a data set as well as the evaluation of the clustering results has been subject of many research efforts [2, 5, 10, 11, 12].

In the sequel, we discuss the fundamental concepts of clustering validity. Further more, in Part I of the paper we present the external and internal validity criteria while the relative ones will be discussed in the forthcoming Part II of the paper.

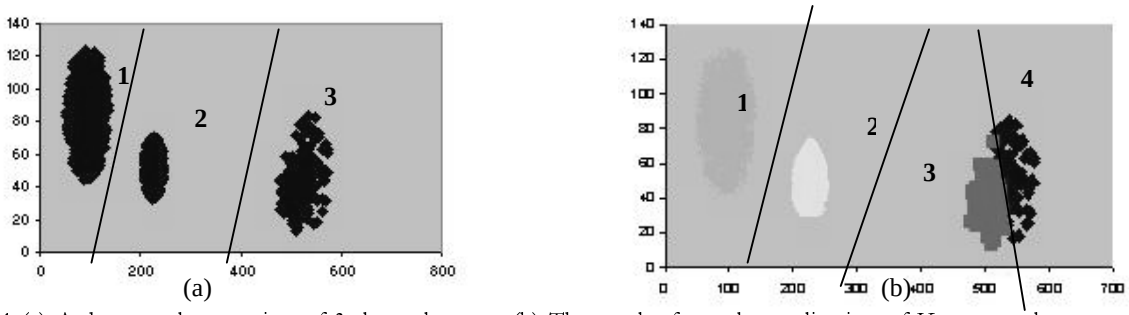
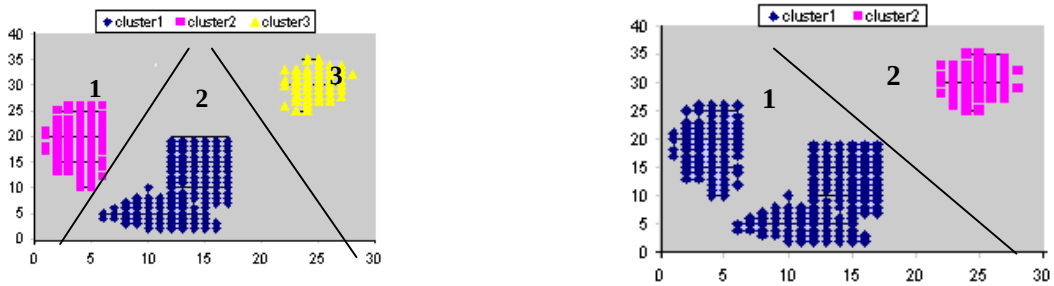


Figure 1 (a) A data set that consists of 3 three clusters, (b) The results from the application of K-means when we ask four clusters



(a) Eps=2, Nps=4 (b) Eps=6, Nps=4

Figure 2 The different partitions resulting from running DBSCAN with different input parameter values.

2. CLUSTER VALIDITY FUNDAMENTAL CONCEPTS

The procedure of evaluating the results of a clustering algorithm is known under the term *cluster validity*. In general terms, there are three approaches to investigate cluster validity [11]. The first is based on *external criteria*. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set. The second approach is based on *internal criteria*. In this case the clustering results are evaluated in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix). The third approach of clustering validity is based on *relative criteria*. Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different input parameter values.

The two first approaches are based on statistical tests and their major drawback is their high computational cost. Moreover, the indices related to these approaches aim at measuring the degree to which a data set confirms an a-priori specified scheme. On the other hand, the third approach aims at finding the best clustering scheme that a clustering algorithm can define under certain assumptions and parameters.

In the rest section, we present the fundamental criteria and representative indices for the first two approaches.

3. EXTERNAL AND INTERNAL VALIDITY INDICES.

In this section, we discuss methods suitable for quantitative evaluation of the clustering results, known as cluster validity methods. However, these methods give an indication of the quality of the resulting partitioning and thus they can only be considered as a tool at the disposal of the experts in order to evaluate the clustering results.

The cluster validity approaches based on external and internal criteria rely on statistical hypothesis testing. In the following section, an introduction to the fundamental concepts of hypothesis testing in cluster validity is presented.

3.1 Hypothesis Testing in Cluster Validity

In cluster validity the basic idea is to test whether the points of a data set are randomly structured or not. This analysis is based on the *Null Hypothesis*, known as *Ho*, expressed as a statement of random structure of a dataset, let X. To test this hypothesis we are based on statistical tests, which lead to a computationally complex procedure. In the sequel Monte Carlo techniques are used as a solution to high computational problems [11].

3.1.1 How Monte Carlo is used in Cluster Validity

The goal of using Monte Carlo techniques is the computation of the probability density function (*pdf*) of the validity indices. They rely on simulating the process of estimating the *pdf* of a validity index using a sufficient number of computer-generated data. First, a large amount of synthetic data sets is generated by a normal distribution. For each one of these synthetic data sets, called X_i, the value of the defined index, denoted q_i, is computed. Then based on the respective values of q_i for each of the data sets X_i, we create a scatter-plot. This scatter-plot is an approximation of the probability density function of the index. In Figure 3 we see the three possible cases of probability density function's shape of an index q. There are three different possible shapes depending on the

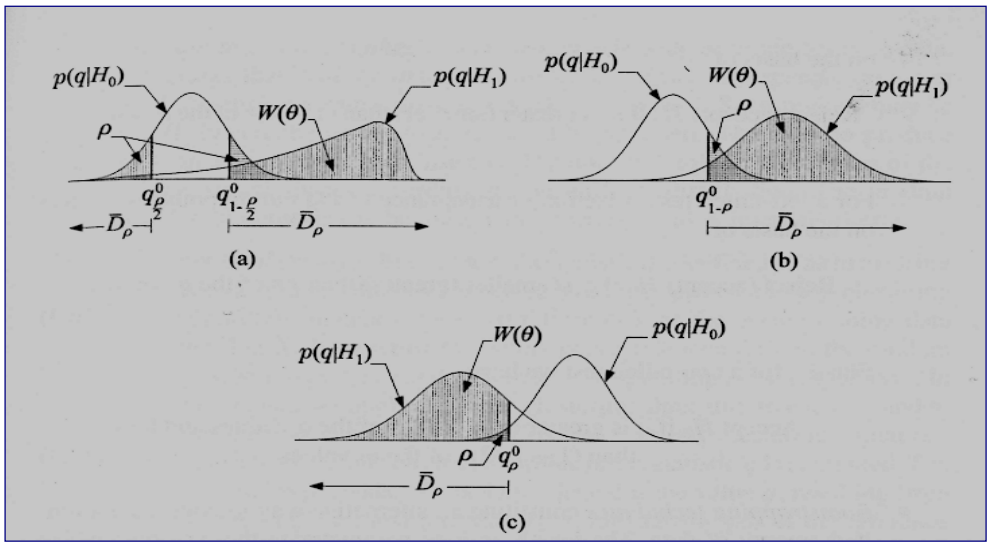


Figure 3. Confidence interval for (a) two-tailed index, (b) right-tailed index, (c) left-tailed index, where q^0_ρ is the ρ proportion of q under hypothesis H_0. [11]

critical interval $\overline{D_\rho}$, corresponding to *significant level* ρ (statistic constant). As we can see the probability density function of a statistic index q, under *Ho*, has a single maximum and the $\overline{D_\rho}$ region is either a half line, or a union of two half lines.

Assuming that the scatter-plot has been generated using r values of the index q, called q_i, in order to accept or reject the *Null Hypothesis Ho* we examine the following conditions [11]:

```
if the shape is right-tailed
  if (q's value of our data set, is
  greater than  (1-ρ)· r of qi values)
  then
   Reject Ho
  else
   Accept Ho
  endif
else if the shape is left-tailed
  if (q's value for our data set, is
  smaller than ρ· r of qi values) then
     Reject  Ho
   else
    Accept Ho
  endif
else if the shape is two-tailed
  if (q is greater than (ρ/2)·r number
  of qi values and smaller than (1-
  ρ/2)·r of qi values)
     Accept Ho
  endif
 endif
```

3.2 External Criteria

Based on the external criteria we can work in two different ways. Firstly, we can evaluate the resulting clustering structure **C**, by comparing it to an independent partition of the data ***P*** built according to our intuition about the clustering structure of the data set. Secondly, we can compare the proximity matrix P to the partition ***P***.

3.2.1 Comparison of C with partition P

Consider $\boldsymbol{C}=\{C_1 \ldots C_m\}$ is a clustering structure of a data set X and $\boldsymbol{P}=\{P_1 \ldots P_s\}$ is a defined partition of the data. We refer to a pair of points $(\mathbf{x_v}, \mathbf{x_u})$ from the data set using the following terms:

- **SS**: if both points belong to the same cluster of the clustering structure **C** and to the same group of partition ***P***.
- **SD**: if points belong to the same cluster of **C** and to different groups of ***P***.
- **DS**: if points belong to different clusters of **C** and to the same group of ***P***.
- **DD**: if both points belong to different clusters of **C** and to different groups of ***P***.

Assuming now that **a**, **b**, **c** and **d** are the number of SS, SD, DS and DD pairs respectively, then a + b + c + d = **M** which is the maximum number of all pairs in the data set (meaning, M=N(N-1)/2 where N is the total number of points in the data set).
Now we can define the following indices to measure the degree of similarity between ***C*** and ***P***:

1. *Rand Statistic*: R = (a + d) / M,

2. *Jaccard Coefficient*: J = a / (a + b + c),

The above two indices range between 0 and 1, and are maximized when m=s. Another index is the:

3. *Folkes and Mallows index*:

$$FM = a / \sqrt{m_1 m_2} = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}} \qquad (1)$$

where $m_1 = a / (a + b)$, $m_2 = a / (a + c)$.

For the previous three indices it has been proven that high values of indices indicate great similarity between **C** and ***P***. The higher the values of these indices are the more similar ***C*** and ***P*** are. Other indices are:

4. *Huberts Γ statistic*:

$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i, j)\, Y(i, j) \qquad (2)$$

High values of this index indicate a strong similarity between X and Y.

5. *Normalized Γ statistic*:

$$\hat{\Gamma} = \frac{[(1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i, j) - \mu_X)(Y(i, j) - \mu_Y)]}{\sigma_X \sigma_Y} \qquad (3)$$

where X(i, j) and Y(i, j) are the (i, j) element of the matrices X, Y respectively that we have to compare. Also μ_x, μ_y, σ_x, σ_y are the respective means and variances of X, Y matrices. This index takes values between –1 and 1.

All these statistics have right-tailed probability density functions, under the random hypothesis. In order to use these indices in statistical tests we must know their respective probability density function under the Null Hypothesis, *Ho*, which is the hypothesis of random structure of our data set. This means that using statistical tests, if we accept the Null Hypothesis then our data are randomly distributed. However, the computation of the probability density function of these indices is computational expensive. A solution to this problem is to use Monte Carlo techniques. The procedure is as follows:

For i = 1 to r

Generate a data set X_i with N vectors (points) in the area of X (i.e., having the same dimension with those of the data set X).

Assign each vector $y_{j,\,i}$ of X_i to the group that $x_j \in X$ belongs, according to the partition ***P***.

Run the same clustering algorithm used to produce structure C, for each X_i, and let C_i the resulting clustering structure.

Compute $q(C_i)$ value of the defined index q for ***P*** and C_i.

End For

Create scatter-plot of the r validity index values , $q(C_i)$ (that computed into the for loop).

After having plotted the approximation of the probability density function of the defined statistic index, its value, let q, is compared to the $q(C_i)$ values, let q_i. The indices R, J, FM, Γ defined previously are used as the q index mentioned in the above procedure.

Example: Assume a data set, X, containing 100 three-dimensional vectors (points). The points of X form four clusters of 25 points each. Each cluster is generated by a normal distribution. The covariance matrices of these distributions are all equal to 0.2I, where I is the 3x3 identity matrix. The mean vectors for the four distributions are $[0.2, 0.2, 0.2]^T$, $[0.5, 0.2, 0.8]^T$, $[0.5, 0.8, 0.2]^T$, and $[0.8, 0.8, 0.8]^T$. We independently group data set X in four groups according to the partition ***P*** for which the first 25 vectors (points) belong to the first group P_1, the next 25 belong to the second group P_2, the next 25 belong to the third group P_3 and the last 25 vectors belong to the fourth group P_4. The K-Means clustering algorithm is run for k = 4 clusters. Assuming that C is the resulting clustering structure the values of the indices for the clustering C and the partition ***P*** are computed. Thus we get R = 0.91, J = 0.68, FM = 0.81 and Γ = 0.75. Then the steps described above are followed in order to define the probability density function of these four statistics. We generate 100 data sets X_i, i = 1,…, 100, and each one of them consists of 100 random vectors (in 3 dimensions) using the uniform distribution. According to the partition ***P*** defined earlier for each X_i the first 25 of its vectors are assigned to P_1 and the second, third and forth groups of 25 vectors to P_2, P_3 and P_4 respectively. Then K-Means is run i-times, one time for each X_i, so as to define the respective clustering structures of datasets, denoted C_i. For each of them the values of the indices R_i, J_i, FM_i, Γ_i, i=1, …,100 are computed. Considering the significance level $\varrho = 0.05$ and these values are compared to the R, J, FM and Γ values corresponding to X. Then the null hypothesis is accepted or rejected whether $(1-\varrho)\cdot r = (1 - 0.05)\ 100 = 95$ values of R_i, J_i, FM_i, Γ_i are greater or smaller than the corresponding values of R, J, FM, Γ. In our case the R_i, J_i, FM_i, Γ_i values are all smaller than the corresponding values of R, J, FM, and Γ, which lead us to the conclusion that the null hypothesis Ho is rejected. Something that we were expecting because of the predefined clustering structure of data set X.

3.2.2 Comparison of P with partition **P**

Let *P* is the proximity matrix of a data set X and **P** is its partitioning. Partition **P** can be considered as a mapping

$$g: X \rightarrow \{1 \ldots n_c\}.$$

where n_c is the number of clusters.
Assuming the matrix Y defined as:

$$Y(i, j) = \begin{cases} 1, & \text{if } g(x_i) \neq g(x_j) \\ 0, & \text{otherwise} \end{cases}, \; i, j = 1 \ldots N.$$

Γ (or normalized Γ) statistic can be computed using the proximity matrix P and the matrix Y. Based on the index value, we may have an indication of the two matrices' similarity.

To proceed with the evaluation procedure we use the Monte Carlo techniques as mentioned above. In the *"Generate"* step of the procedure the corresponding mappings g_i is generated for every generated X_i data set. So in the *"Compute"* step the matrix Y_i is computed for each X_i in order to find the Γ_i corresponding statistic index.

3.3 Internal Criteria.

Using this approach of cluster validity the goal is to evaluate the clustering result of an algorithm using only quantities and features inherited from the dataset. There are two cases in which we apply internal criteria of cluster validity depending on the clustering structure: a) hierarchy of clustering schemes, and b) single clustering scheme.

3.3.1 Validating hierarchy of clustering schemes.

A matrix called cophenetic matrix, P_c, can represent the hierarchy diagram that is produced by a hierarchical algorithm. The element $P_c(i, j)$ of cophenetic matrix represents the proximity level at which the two vectors x_i and x_j are found in the same cluster for the first time. We may define a statistical index to measure the degree of similarity between P_c and P (proximity matrix) matrices. This index is called *Cophenetic Correlation Coefficient* and defined as:

$$CPCC = \frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}\, c_{ij} - \mu_P \mu_C}{\sqrt{\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 - \mu_P^2\right]\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_C^2\right]}} \quad (4)$$

where $M=N\cdot(N-1)/2$ and N is the number of points in a dataset. Also, μ_P and μ_C are the means of matrices P and P_c respectively, and are defined in the equation (5):

$$\mu_P = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P(i,j), \mu_C = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P_c(i,j) \quad (5)$$

Moreover, d_{ij}, c_{ij} are the (i, j) elements of P and P_c matrices respectively. The CPCC are between -1 and 1. A value of the index close to 1 is an indication of a significant similarity between the two matrices. The procedure of the Monte Carlo techniques described above is also used in this case of validation.

3.3.2 Validating a single clustering scheme

The goal here is to find the degree of match between a given clustering scheme C, consisting of n_c clusters, and the proximity matrix P. The defined index for this approach is Hubert's Γ statistic (or normalized Γ statistic). An additional matrix for the computation of the index is used, that is

$$Y(i,j) = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ belong to different clusters} \\ 0, & \text{otherwise} \end{cases}$$

where i, j = 1,…, N.

The application of Monte Carlo techniques is also here the way to test the random hypothesis in a given data set.

4. CONCLUSIONS AND TRENDS IN CLUSTERING VALIDITY

Cluster validity is one of the most important issues in cluster analysis related to the inherent features of the data set under concern. It aims at the evaluation of clustering results and the selection of the scheme that best fits the underlying data.

The majority of algorithms are based on certain criteria in order to define the clusters in which a data set can be partitioned. Since clustering is an unsupervised process and there is no a-priori indication for the actual number of clusters presented in a data set, there is need of clustering results' validation. We presented a survey of the most known validity criteria available in literature, classified in three categories: external, internal, and relative. In Part I of the paper we discuss representative validity indices of internal and external criteria while in Part II we will present validity approaches based on relative criteria along with sample experimental evaluation.

The validity assessment approaches proposed in the literature works better when the clusters are mostly compact. However, there are applications where we have to handle arbitrary shaped clusters (e.g. spatial data, medicine, biology). In this case the traditional validity criteria (variance, density and its continuity, separation) are not any more sufficient.

There is a need for developing quality measures that assess the quality of the partitioning taking into account: i. intra-cluster quality, ii. inter-cluster separation and iii. geometry of the clusters, using sets of representative points, or even multidimensional curves rather than a single representative point.

Also another challenge is addressing the issue of an integrated data mining results quality assessment model. The fundamental concepts and criteria for a global data mining validity checking process have to be introduced and integrated to define a quality model.

ACKNOWLEDGEMENTS

This work was supported by the General Secretariat for Research and Technology through the PENED ("99EΔ 85") project.

REFERENCES

[1] Michael J. A. Berry, Gordon Linoff . *Data Mining Techniques For marketing, Sales and Customer Support*. John Willey & Sons, Inc, 1996.

[2] Dave, R. N. . "Validating fuzzy partitions obtained through c-shells clustering", *Pattern Recognition Letters*, Vol .10, pp613-623, 1996.

[3] Ester, M., Kriegel, H-P., Sander, J., Xu, X.. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of 2nd Int. Conf. On Knowledge Discovery and Data Mining*, Portland, pp. 226-23, 1996.

[4] Fayyad, M. U., Piatesky-Shapiro, G., Smuth P., Uthurusamy, R.. *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996

[5] Gath I., Geva A.B. "Unsupervised optimal fuzzy clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 11(7), 1989.

[6] Guha, S, Rastogi, R., Shim K.. "ROCK: A Robust Clustering Algorithm for Categorical Attributes", *Published in the Proceedings of the IEEE Conference on Data Engineering*, 1999.

[7] Han, J., Kamber, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[8] Jain, A.K., Murty, M.N., Flyn, P.J.. "Data Clustering: A Review", *ACM Computing Surveys*, Vol.31, No3, 1999.

[9] MacQueen, J.B. "Some Methods for Classification and Analysis of Multivariate Observations", *In Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability*, Volume I: Statistics, pp281-297, 1967.

[10] Rezaee, R, Lelieveldt, B.P.F., Reiber, J.H.C. "A new cluster validity index for the fuzzy c-mean", *Pattern Recognition Letters*, 19, pp. 237-246, 1998.

[11] Theodoridis, S., Koutroubas, K.. *Pattern recognition*, Academic Press, 1999.

[12] Xie, X. L, Beni, G.. "A Validity measure for Fuzzy Clustering", *IEEE Transactions on Pattern Analysis and machine Intelligence*, Vol.13, No4, 1991.