# Cluster Stability Assessment Based on Theoretic Information Measures

Damaris Pascual[1], Filiberto Pla[2], and J. Salvador Sánchez[2]

[1] Center for Pattern Recognition and Data Mining, Universidad de Oriente,
Av. Patricio Lumumba s/n, Santiago de Cuba, 90500, Cuba
damaris@cerpamid.co.cu
[2] Dept. Llentguages i Sistemas Informátics, Universitat Jaume I, 12071, Castelló, Spain
{pla, sanchez}@lsi.uji.es

**Abstract.** Cluster validation to determine the right number of clusters is an important issue in clustering processes. In this work, a strategy to address the problem of cluster validation based on cluster stability properties is introduced. The stability index proposed is based on information measures taking into account the variation on some of these measures due to the variability in clustering solutions produced by different sample sets of the same problem. The experiments carried out on synthetic and real database show the effectiveness of the cluster stability index when the clustering algorithm is based on a data structure model adequate to the problem.

**Keywords:** cluster validation, stability indices, information theory.

## 1 Introduction

There exist in the literature different techniques aimed at discovering relationships among objects in a database. Clustering algorithms are one of these techniques used to infer properties of the data set, whose goal is to divide the data into groups so that objects of the same group are more similar than objects of different groups. One of the problems to be solved in a clustering process is the clustering model to be used, given the data set [2, 6]. Another important question is to assess the "natural" number of groups in a given data set, which is even more challenging when no clustering model is available.

There exist several approaches to address the problem of determining the number of clusters, which uses different validity indices. Most of these approaches exploit the idea of within-cluster variability and the "elbow" phenomenon [1, 5, 7, 8]. Other methods focus on the "elbow" phenomenon, proposing statistical measures, like the gap statistic [9]. A comprehensive survey of methods for estimating the number of clusters is given in [4].

An alternative approach to asses the "natural" number of clusters is the so-called stability behaviour of the resulting clustering with respect to variations in the data sample used. The stability of the clustering solutions is then assessed by defining a variability measure of the clustering solutions. Most of the works based on cluster

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 219–226, 2008.

stability try to assess this variability by measuring indices related to the ratios of objects or pair of objects that have not been included in the same partition by two different solutions of the clustering algorithm on two different samples [10], or using statistical tests on these variations [11], which usually need some parameter to be set to find the optimal solution for the number of clusters.

Other approaches on cluster stability are based on a transfer by prediction strategy [3], using the prediction made by a given classifier trained on the resulting partitions of clustering solutions from different data samples. Apart from the need of choosing a certain type of classifier, this method involves a considerable computational burden, due to the assessment of all possible permutations label assignments of the clustering solutions. To avoid the dependency of the index proposed with respect to the number of clusters, this has to be normalised with respect to the cost of a random predictor.

In the present work, we focus on the clustering stability approach to determining the "natural" number of clusters in a dataset, proposing a new strategy based on a cluster stability criterion based on information theory, modeling the partition of a data set as a noisy communication channel [12], exploiting the relationship of some information measures with a pattern recognition problem. The proposed algorithm for cluster stability assessment tries to avoid the drawbacks of the transfer by prediction strategy and the need of setting-up parameters. The method presented is also aimed at assessing clustering solutions from any clustering model and algorithm. An open question would be the assessment of the clustering model that better fits a given data set, but this question will not be addressed in the present work.

## 2 Measuring the Cluster Stability

### 2.1 Channel Communication as a Pattern Recognition Problem

Let $X$ be the random variable distributed as $p(x)$, representing the dataset in a d-dimensional space $(x_1,\ldots,x_d)$, and let $Y$ be the random variable distributed as $p(y)$ representing the $k$ class labels, $y\in\{y_1,\ldots,y_k\}$, of the objects in the database.

A classification process can be modeled as a noisy communication channel [12], where the channel transition probability distribution can be represented by the class likelihoods $p(x/y)$. The pattern recognition process is then represented by a set $W$ of $m$ possible messages $w\in\{w_1,\ldots,w_m\}$, using a mn-code $C^{(\mathrm{n})}=(m,n)$ made by sequences of $n$ label values $y^n$, whose code values are distributed as the class labels $p(y)$. When the sender sends a sequence $y^n$, the receiver sees on the other side of the channel the corresponding sequence $x^n$. The receiver then uses a decoding function $g(x^n)$:$X^n\rightarrow W$, making a guess about the message sent $g(x^n)=w$.

In a pattern recognition problem, the decoding function can be represented by the decision rule. If we use the Bayes decision rule, then, the decoding function becomes

$$y = g(x) = \underset{j=1,\ldots,k}{\arg\max}\left\{p(y_j / x)\right\} = \underset{j=1,\ldots,k}{\arg\max}\left\{p(x / y_j)p(y_j)\right\} \tag{1}$$

On the other hand, the channel capacity represented by $p(x/y)$, is defined as the supreme of its possible achievable rates. According to the Channel Capacity Theorem [12] the channel capacity is provided by

$$C = \max_{p(y)} I(X;Y) \tag{2}$$

Li [13] showed that the mutual information $I(X;Y)$ between the data distribution $p(x)$ and the class distribution $p(y)$ in a decision problem is related to the Bayes error $R$ of the decision problem when using the decision rule (1) as

$$\frac{1}{4(k-1)}\left(H(Y)-I(X;Y)\right)^2 \leq R \leq \frac{1}{2}\left(H(Y)-I(X;Y)\right) \tag{3}$$

where $H(X)$ and $H(Y)$ are the Shannon entropies of random variables $X$ and $Y$, respectively. Expression (3) provides a lower and an upper bound for the Bayes error. Therefore, if we can make an estimate of these information measures for a given pattern recognition problem, we could have an estimate of these Bayes error bounds.

### 2.2 Cluster Stability Assessment

The approach here proposed is based on measure the transfer by prediction variability by means of information measures, as a way of assessing cluster stability. Variability on prediction will convey a variability of the decision error, the proposed method will estimate this variability by means of assessing the variation of some of the information measures involving expression (3) and the decision rule (1).

Let two different data samples be extracted by a statistically independent process, $L_1$ and $L_2$, drawn from the unknown true distribution $p(x)$ representing the data. Let a clustering algorithm representing an optimization rule of a clustering model. For a given number of clusters $k$, the clustering algorithm will provide a partition solution of data sets $L_1$ and $L_2$, represented by distributions $p_1(y)$ and $p_2(y)$, respectively.

Let us assume that, for a given number of clusters k, the data partition provided by the clustering algorithm over the true data distribution is represented by $p_k(y)$. Therefore, if we fix $H_k(Y)$ in expression (3), the variation due to two different clustering solutions in the estimation of the Bayes error bounds in expression (3) could be assessed by estimating the variation in the mutual information due to the use of two different samples in the transfer by prediction between the corresponding two different clustering solutions $p_1(y)$ and $p_2(y)$.

In order to estimate the transfer by prediction variability, let us have a look to the mutual information measure $I_k(X;Y)$ between the true data distribution $X$ and the data partition of $k$ clusters provided by the given clustering algorithm $Y$

$$I_k(X;Y) = \int_x p(x) \sum_y p(y/x) \log \frac{p(y/x)}{p(y)}\, dx = \mathrm{E}_{p(x)}\left\{KL(p(y/x) \,\|\, p(y))\right\} \tag{4}$$

Previous expression can be interpreted as the expected value according to $p(x)$, of the Kullback-Leibler divergence between the data partition distribution $p(y)$ generated by the clustering algorithm in $k$ different classes, and the posterior probabilities $p(y/x)$ used in the decision rule (1). In order to make an estimation of $I_k$ given the sample data sets $L_1$ and $L_2$ , let us assume data set $L_1$ is used as an empirical estimate for the true data distribution $p(x) \approx p_1(x)$, and the data partition of the clustering algorithm on $L_1$ as an estimate for the data partition $p(y) \approx p_1(y)$. Therefore, the empirical data distribution of data set $L_1$ containing $N_1$ samples, can be expressed as

$$p(x) \approx p_1(x) = \frac{1}{N_1} \sum_{i=1}^{N_1} \delta(x, x_i) \tag{5}$$

Then, expression (3) becomes

$$I_k(X;Y) \approx \frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{y} p(y/x_{1i}) \log \frac{p(y/x_{1i})}{p(y)} \tag{6}$$

On the other hand, in order to make an estimate of the posteriors $p(y/x)$, if we have used data sample $L_1$ to estimate the true probability distribution and the data partition distribution, let us use data sample $L_1$ and the labeling made by the partition solution provided by the clustering algorithm, as the test set, and data sample $L_2$ , as a training set. Thus, the posteriors $p(y/x)$ can be estimated using training set $L_2$ and test set $L_1$ as

$$p_2(y/x_{1i}) = \delta\big(y_1(x_{1i}), y_2(NN_2(x_{1i}))\big) \tag{7}$$

Where $y_1(x_{1i})$ is the class label assigned to sample $x_{1i}$ of data set $L_1$ by the clustering algorithm, and $NN_2(x_{1i})$ is the nearest neighbor of $x_{1i}$ in the sample set $L_2$. Eventually, the estimate of $I_k(X;Y)$ using two independent data samples $L_1$ and $L_2$ , is expressed as

$$I_k(X;Y) \approx \hat{I}_k(X;Y) = \frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{j=1}^{k} p_2(y_j/x_{1i}) \log \frac{p_2(y_j/x_{1i})}{p_1(y_j)} \tag{8}$$

Since the estimate $\hat{I}_k$ varies for different sample sets, it is a random variable. Thus, according to the law of large numbers, when the number of measurements tends to infinity, the expected value of this variable tends to the true value. For a finite number of measurements $N$, the true value of this variable concentrates around its mean with variance $\sigma_k^2$ ,

$$\sigma_k^2 = \mathrm{E}(\hat{I}_k - \mathrm{E}(\hat{I}_k))^2 \tag{9}$$

The standard deviation $\sigma_k$ will represent the cluster stability index of the algorithm for k clusters. This cluster stability index is an estimate of the variation of the transfer by prediction between $N$ couples of clustering solutions, when partitioning every pair of independent sample sets into $k$ clusters using a certain clustering algorithm. Therefore, for a given clustering algorithm the correct number of clusters $k^*$ will be chosen as the number of clusters that minimize the stability index (9), that is,

$$\sigma_{k^*} = \min_k \sigma_k \tag{10}$$

## 3 Experimental Results

In order to show the performance of the cluster validity index proposed, three types of databases were used. Two types of synthetic databases, one of them consists of Gaussian clusters, and the other one consists of clusters of arbitrary shapes and sizes. A third group of data consists of real databases. On the other hand, three different algorithms of clustering are used in the experiments, the well known K-Means, the Gaussian mixture model using EM (Expectation Maximization) and H-Density [6]. These algorithms correspond to three different models, where K-means looks for compact clusters around a mean, the Gaussian mixture model looks for Gaussian-shape clusters and the H-Density is an agglomerative hierarchical algorithm based on data density estimates and a single link strategy for clusters of any shape.

For each database used in the experiments, the experimental set-up consists of splitting randomly the database into two equal size sample sets $L_1$ and $L_2$ . For $k$=2, …,10 number of clusters, each of the clustering algorithm is run on data samples $L_1$ and $L_2$ , and the cluster stability index (10) is assessed on 10 different realisations of $L_1$ and $L_2$ . Inverting the role of $L_1$ and $L_2$ in the estimate of mutual information (8) provides an estimate of the stability index over $N$=20 realisations.

### 3.1 Gaussian Databases

The first database consists of "Four Gaussians" in two dimensions, with little overlapping among them. The second one is formed by "Six Gaussians" with different degree of overlapping, and the third database has also six clusters, "Six Gaussians-II", but two of them are completely overlapped, with the same mean and different covariance. Table 1 shows the result of the cluster stability index proposed for the second and third database, for the three clustering algorithms. Notice how the index shows a minimum value for the right number of clusters (see Fig. 1) when using the right clustering model, in this case, the Gaussian mixture or the H-density algorithms. In the case of the "Six Gaussians-II" database, because the completely overlapped couple of Gaussians, the algorithms cannot distinguish them, although for the Gaussian mixture model, the 6 correct clusters are the second lowest stability index. k-Means is able to detect the four Gaussians (Fig. 1a), but fails on the other examples, due to the overlapping and the fact that clusters are not modelled as spherical Gaussian distributions, which would be more adequate for a K-Means model.

### 3.2 Arbitrary Shape Databases

This section presents the results obtained over three synthetic databases, which have clusters with different shapes, sizes and densities, and they are separated by zones of

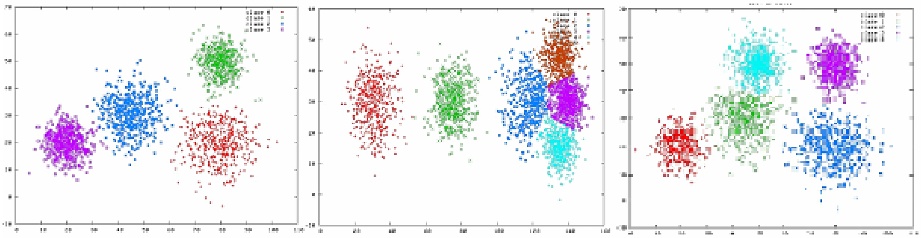

**Fig. 1.** Partitions selected by the best stability index using the EM-Gaussian Mixture algorithm: a) Four Gaussians b) Six Gaussians and c) Six Gaussians-II (two concentric)

**Table 1.** Stability indices for the “Six Gaussians” and “Six Gaussians-II” databases

| | Six Gaussians | | | Six Gaussians-II (two concentric) | | |
|---|---|---|---|---|---|---|
| Num. clusters | H-Density | K-Means | Gaussian Mixture | H-Density | K-Means | Gaussian Mixture |
| 2 | 0,027 | 0,00003 | 0,000002 | 0,024 | 0,017 | 0,0043 |
| 3 | 0,00008 | 0,056 | 0,0000084 | 0,01 | 0,0044 | 0,0037 |
| 4 | 0,002 | 0,0041 | 0,00002 | 0,00006 | 0,0042 | 0,0078 |
| 5 | 0,003 | 0,0023 | 0,002 | 0,00003 | 0,0001 | 0,00003 |
| 6 | 0,0000084 | 0,0021 | 0,000009 | | 0,0066 | 0,00029 |
| 7 | | 0,0025 | 0,005 | | 0,016 | 0,0065 |
| 8 | | 0,015 | 0,026 | | 0,0045 | 0,017 |
| 9 | | 0,017 | 0,034 | | 0,012 | 0,04 |
| 10 | | 0,009 | 0,01 | | | |

low density (Fig. 2). The first database consists of three concentric rings, the second one is a pair of half-rings and the third one is the DS1 database used in other works (see [6]). Table 2 shows the stability indices in the case of the concentric rings and DS1 database. In this case; the H-Density algorithm is able to provide the correct clusters if the right number of clusters is selected. The stability index proposed shows a minimum value in the right number of clusters for this clustering algorithm, because the clustering model is more adequate in the databases used (see Fig. 2).

**Table 2.** Stability indices for the “Three Concentric Rings” and DS1 databases

| | Three Concentric Rings | | | DS1 database | | |
|---|---|---|---|---|---|---|
| Num. clusters | H-Density | K-Means | Gaussian Mixture | H-Density | K-Means | Gaussian Mixture |
| 2 | 0,0014 | 0,00009 | 0,026 | 0,217 | 0,0000005 | 0,03 |
| 3 | 0,0011 | 0,00017 | 0,017 | 0,013 | 0,0001 | 0,008 |
| 4 | 0,0547 | 0,00006 | 0,0029 | 0,023 | 0,0015 | 0,0014 |
| 5 | 0,0356 | 0,0002 | 0,0045 | 0,023 | 0,000004 | 0,000015 |
| 6 | 0,021 | 0,0009 | 0,0048 | 0,0025 | 0,001 | 0,0018 |
| 7 | 0,0332 | 0,002 | 0,0035 | 0,013 | 0,001 | 0,000036 |
| 8 | 0,0249 | 0,001 | 0,006 | 0,042 | 0,001 | 0,000041 |
| 9 | | 0,004 | 0,007 | 0,015 | 0,0005 | 0,0011 |
| 10 | | 0,001 | 0,004 | 0,015 | 0,0006 | 0,0041 |

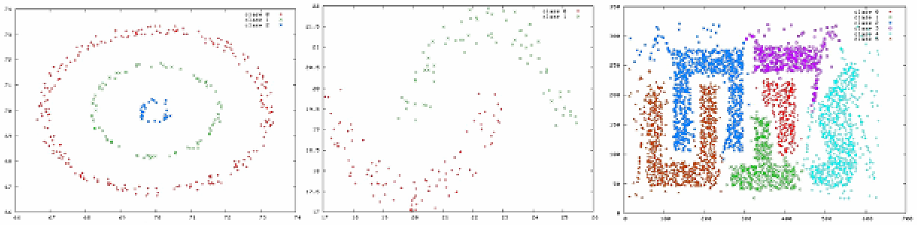

**Fig. 2.** Partitions selected by the best stability index using the H-Density algorithm for the a) Three concentric rings b) Two half rings and c) DS1 database [6]

### 3.3 Real Databases

The “House” database [6] represents the chromatic ab pixel values of the House colour image in the Lab space (Fig. 3b). Fig. 3 shows the result of the clustering selected by the best stability index (see Table 3) when using the H-Density algorithm. Notice the quality of the clustering selected by looking at the pixel labelling that provides the selected solution by the stability index (Fig. 3c).

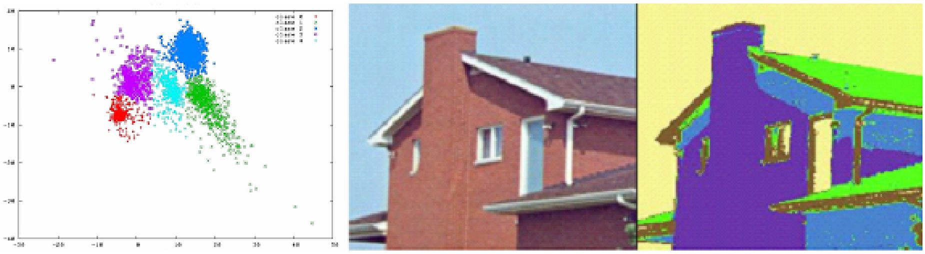

**Fig. 3.** (a) Clusters selected by the stability index when using H-Density algorithm in ab space. b) “House” image. c) Pixels labeled by H-Density algorithm result from (a).

The method was also tested in the IRIS database, which has 150 samples of three types of plants. Two of these classes are highly overlapped, and present a high difficulty for most of clustering algorithms. The other class is clearly separated from the other two ones. The H-Density algorithm is able to detect, among the levels of the hierarchy, the level of three classes with a high percent of correct classification (see [6] for details). In this case, the stability index introduced here has been able to detect the right number of classes, even in presence of high degree of overlapping of two of them. The other clustering algorithms do not provide the right clustering solution, because of the inadequate clustering model.

## 4 Conclusions

The problem of determining the optimal or “natural” number of groups in a database is an important issue for clustering processes, together with the selection of the adequate clustering model for each particular dataset. In this work, a method to select the “natural” number of clusters have been presented, based on cluster stability criterion inspired in an information theoretic approach to assess the variability of clustering solutions due to the different clustering partitions obtained from different data samples of the same problem.

The experiments carried out on several synthetic and real databases, using three different clustering models; show that, when the clustering algorithm is adequate to model the data structure, the stability index proposed can select the right solutions in a wide variety of synthetic and real examples with cluster structures of different shapes, sizes and overlapping degree.

As it has also been outlined, another issue beyond the scope of this work is the problem of selecting the adequate clustering model for a given problem, which can affect significantly the stability performance of the clustering algorithm used.

**Acknowledgments.** This work has been partially supported by Spanish Ministry of Science and Education under projects ESP2005-00724-C05-05, CSD2007-00018 and PET2005-0643 and project P1 1B2007-48 by Fundació Caixa-Castelló.

## References

1. Bouguessa, M., Wang, S., Sun, H.: An Objective approach to cluster validation. Pattern Recognition Letters 27, 1419–1430 (2006)
2. Ertoz, L., Steinbach, M., Kumar, V.: Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In: Third SIAM International Conference on data Mining (2003)
3. Lange, T., Braun, M.L., Buhmann, J.M.: Stability-Based Validation of Clustering Solutions. Neural Computation 16, 1299–1323 (2004)
4. Milligan, G.W., Cooper, M.C.: An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179 (1985)
5. Pal, N.R., Bezdek, J.C.: On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst. 3(3), 370–379 (1995)
6. Pascual, D., Pla, F., Sánchez, J.S.: Non Parametric Local Density-based Clustering for Multimodal Overlapping Distributions. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 671–678. Springer, Heidelberg (2006)
7. Sugar, C.: Techniques for clustering and classification with applications to medical problems. PhD Dissertation Stanford University, Stanford (1998)
8. Sugar, C., Lenert, L., Olshen, R.: An application of cluster analysis to health services research: empirically defined health states for depression from the sf-12. Technical Report Stanford University, Stanford (1999)
9. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. R. Statist Soc. B 63, Part 2, 411–423 (2001)
10. Ben-Hur, A., Guyon, I.: Detecting stable clusters using principal component analysis. In: Brownstein, M., Khodursky, A. (eds.) Methods in Molecular Biology, pp. 159–182. Humana press (2003)
11. Mufti, G.B., Bertrand, P., Moubarki, L.E.: Determining the number of groups from measures of cluster validity. In: ASMDA 2005, pp. 404–414 (2005)
12. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, Chichester (1991)
13. Li, J.: Divergence measures based on Shannon entropy. IEEE Trans. on Information Theory 37(1), 145–151 (1991)