Determining the number of groups from measures of cluster stability

G. Bel Mufti[1], P. Bertrand[2] and L. El Moubarki[3]

[1] U.R. CEFI, ESSEC de Tunis, 4 rue Abou Zakaria El Hafsi, Montfleury, 1089 Tunis, Tunisie (e-mail: belmufti@yahoo.com)
[2] GET - ENST Bretagne, Dept. Lussi, CS 83818, 29238 BREST Cedex 3, France (e-mail: patrice.bertrand@enst-bretagne.fr)
[3] ISG de Tunis, 41 rue de la liberté, Cité Bouchoucha, Le Bardo, 2000 Tunis, Tunisie (e-mail: elmoubarki_lassad@yahoo.fr)

Abstract. An important line of inquiry in cluster validation involves measuring the stability of a partition with respect to perturbations of the data set. Several authors have recently suggested that the 'correct' number of clusters in a partition can be determined simply by examining the partition stability measures for different values of numbers of clusters. In this paper, we consider the clustering stability measures that were recently proposed in [Bertrand and Bel Mufti, 2005], and we present experiments that compare the method for predicting the number of clusters that is derived from these stability measures with two of the most successful methods reported in recent surveys.
Keywords: Cluster Stability, Monte Carlo Test, Cluster Isolation and Cluster Cohesion, Loevinger's measure, Number of clusters of a partition.

1 Introduction

A major challenge in cluster analysis is the validation of clusters resulting from cluster analysis algorithms. One relevant approach involves defining an index measuring the adequacy of a cluster structure to the data set and establishing how likely a given value of the index is under some null model formalizing 'no cluster structure', e.g., [Bailey and Dubes, 1982], [Jain and Dubes, 1988], [Gordon, 1994], [Milligan, 1996] and [Gordon, 1999]. Another type of approach is concerned with the estimation of the stability of clustering results. Informally speaking, cluster stability holds when membership of the clusters is not affected by small changes in the data set [Cheng and Milligan, 1996]. Several recent approaches, see for example [Tibshirani *et al.*, 2001], [Levine and Domany, 2001], [Ben-Hur *et al.*, 2002] and [Bertrand and Bel Mufti, 2005], suggest that cluster stability is a valuable way to determine the number of clusters of any partitioning of the data. Such a stability based approach aims to identify those values of the number of clusters (or any other parameter of the clustering method) for which local maxima of stability are reached.

The main contribution of this paper is to compare this stability based approach with two of the most (classical) successful methods of predicting

the number of clusters. In what follows, we restrict our attention to the measures of cluster stability that were introduced by Bertrand and Bel Mufti [Bertrand and Bel Mufti, 2005]. In section 2, a summarized description of the cluster validation method introduced by Bertrand and Bel Mufti [Bertrand and Bel Mufti, 2005] is presented. This method involves the definition of stability measures both of the partition and of its clusters. Each stability measure is defined as Loevinger's measure of a rule quality, that is assessed by a probability significance which is approximated by comparing the value of the measure with values that would be obtained under a null model that specifies the absence of cluster stability. In section 3, we compare three methods for determining the number of clusters of any partitioning of a data set, on the basis of their experimental results obtained for the partitioning of two data sets. The first method is the stability based approach that is briefly mentioned here above and that is specified by the stability measures of Bertrand and Bel Mufti [Bertrand and Bel Mufti, 2005]. The other two methods are classical methods performing the best for estimating the number of clusters, according to the survey of Milligan and Cooper [Milligan and Cooper, 1985].

2 The cluster stability measures proposed by Bertrand and Bel Mufti (2005)

In this section, we briefly describe the stability based method of cluster validation that was recently introduced by Bertrand and Bel Mufti, and we refer the reader to [Bertrand and Bel Mufti, 2005] for more details.

We will denote as $\mathcal{X}$ an arbitrary data set of n objects to be clustered, and as P_k any generic k-way partitioning algorithm. The partition obtained by running P_k on the data set $\mathcal{X}$ will be denoted by $\mathcal{P}$, in other words $\mathcal{P} = \mathrm{P}_k(\mathcal{X})$. The validation method proposed in [Bertrand and Bel Mufti, 2005] is designed to estimate the stability of both the partition $\mathcal{P}$ and its clusters, with regards to both cluster isolation and cluster cohesion criteria. The perturbed data sets are (random) samples of the population $\mathcal{X}$. If all partitions into k clusters obtained from running algorithm P_k on different samples of $\mathcal{X}$ are close in structure to partition $\mathcal{P}$, then $\mathcal{P}$ can be deemed to be stable. In order to guarantee that each cluster of $\mathcal{P}$ is still represented in each random sample of $\mathcal{X}$, we use a sampling procedure, called *proportionate stratified sampling*. More precisely, given any cluster A of $\mathcal{P}$ and denoting by n_A the size of A, and by f some sampling ratio, this sampling procedure involves selecting randomly and without replacement n'_A elements in each cluster of $\mathcal{P}$, where n'_A is the integer value obtained by rounding down fn_A to the nearest integer. On the basis of experimental results presented in [Bertrand and Bel Mufti, 2005] and recommendations given in [Levine and Domany, 2001] and [Ben-Hur *et al.*, 2002], the value of f has to be chosen in the interval $[0.7, 0.9]$.

Let us focus on the single criterion of cluster isolation. Informally speaking, there is much evidence that any cluster of $\mathcal{P}$, say A, is isolated whenever the following rule holds for any sample $\mathcal{X}'$ of $\mathcal{X}$:

(R) *Isolation rule of A.* If two objects of $\mathcal{X}'$ are not clustered together by partition $\{A, \mathcal{X} \setminus A\}$, then they are not in the same cluster of $\mathrm{P}_k(\mathcal{X}')$.

Any measure of rule quality can assess the rule (R). However, due to its specific properties and its simplicity of interpretation (see [Lenca *et al.*, 2003]), Loevinger's measure ([Loevinger, 1947]) is preferred to other measures of rule quality. Loevinger's measure of rule $E \Rightarrow F$, is defined as the expression $1 - P(E \cap \neg F)/P(E)P(\neg F)$. Denoting by $t(A,\, \mathcal{X}')$ Loevinger's measure of the quality of rule (R), we obtain:

$$t(A,\, \mathcal{X}') = 1 - \frac{n'(n'-1)m_{(\mathcal{X}';\, A,\, \overline{A})}}{2n'_A(n'-n'_A)\; m_{(\mathcal{X}')}}\,, \tag{1}$$

where $m_{(\mathcal{X}')}$ is the number of pairs of objects that are clustered together by $\mathrm{P}_k(\mathcal{X}')$, and where $m_{(\mathcal{X}';\, A,\, \overline{A})}$ is the number of pairs of sampled objects that are in the same cluster of $\mathrm{P}_k(\mathcal{X}')$ and for which exactly one of the two objects belongs to A. Taking into account only the criterion of cluster isolation, the stability measure of cluster A is defined simply as the average, denoted here by $\overline{t}_N(A)$, of the values $t(A,\, \mathcal{X}'_i)$ obtained for a large number N of samples $\mathcal{X}'_i$ $(i = 1, \ldots, N)$:

$$\overline{t}_N(A) = \frac{1}{N} \sum_{i=1}^{N} t(A,\, \mathcal{X}'_i). \tag{2}$$

It should be noted that $\overline{t}_N(A)$ is an (unbiaised) estimation of the expected value of the random variable $t(A, \mathcal{X}')$, when $\mathcal{X}'$ is considered as a random sample. This leads us to select a value of N large enough so that both the central limit theorem holds and the length of the approximate standard 95%-confidence interval is less than some maximal desired length l.

Several other stability measures were similarly defined in order to assess other characteristics of any cluster, *i.e.*, its isolation with respect to another cluster, its cohesion and its validity. In addition, the same three characteristics (isolation, cohesion and validity) of any partition were defined. Furthermore, it was proved that each stability measure of any partition that concerns isolation (resp. cohesion) is a weighted mean of the stability measures of all its clusters with respect to the criterion of isolation (resp. cohesion).

One important issue concerns the interpretation of the order of magnitude of the observed values of stability measures. This is a general problem in cluster validation: Jain and Dubes ([Jain and Dubes, 1988] p.144) noted that it is easy to propose indices of cluster validity, but that it is very difficult to fix thresholds on such indices that define when the index is large or small enough to be 'unusual'. The difficulty is solved by following the general

procedure presented by Jain and Dubes [Jain and Dubes, 1988] (see also [Gordon, 1994]), since it seems reasonable to specify the absence of cluster stability by the absence of clustering structure:

Step 1. Define a null model $\mathcal{M}_0$ that specifies the null hypothesis H_0 of absence of cluster stability for the data set under investigation; in the case of a data set that can be represented by n points of an euclidean space, an example of such a null model is the uniform distribution of n points in the convex hull of the data set.

Step 2. Estimate the probability significance of the observed value of the stability measure under the null hypothesis H_0. Since the analytic expression of the distribution of the stability measure under the null model $\mathcal{M}_0$ is usually unknown, this step generally involves performing a Monte Carlo test: a large number, say M, of data sets are simulated according to the model $\mathcal{M}_0$, and each of them is partitioned and the corresponding value of stability measure is computed. The probability significance is then estimated on the basis of these M values of the stability measure.

For example, the value $\overline{t}_N(A) = 0.899$ is an indication of high stability if and only if its estimated probability significance value under H_0 is less than 5%.

3 Experimental comparison with two methods for determining the number of clusters

As previously mentioned in section 1, a method for determining the 'optimal' number of clusters in a partitioning of a data set can easily be derived from the stability measure of a partition introduced in [Bertrand and Bel Mufti, 2005]: a k-clusters partition is considered as meaningful if the value of the partitional stability measure is a local maximum when k varies. In what follows, this partitional stability measure will be denoted as $BB(k)$, when k is the number of clusters of the partition. The information provided by the stability index $BB(k)$ can be refined by considering its probability significance under the null hypothesis H_0, and also by taking into account the stability measures (concerning isolation and cohesion) of each cluster in the partition together, with their probability significances.

Otherwise, many indices that measure the adequation between the partition and the data set were proposed to determine the number of clusters. According to the survey of Milligan and Cooper [Milligan and Cooper, 1985], the index of Calinski and Harabasz [Calinski and Harabasz, 1974] and the index of Krzanowski and Lai [Krzanowski and Lai, 1985] are among the indices that perform the best (see also Tibshirani et al [Tibshirani *et al.*, 2001]

for another experimental comparison). The index of Calinski and Harabasz is defined by:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{3}$$

where k denotes the number of clusters, and $B(k)$ and $W(k)$ denote the between and within cluster sums of squares of the partition, respectively. An optimal number of clusters is then defined as a value of k that maximizes $CH(k)$. The index of Krzanowski and Lai is defined by:

$$KL(k) = |\frac{DIFF(k)}{DIFF(k+1)}|, \tag{4}$$

where:

$$DIFF(k) = (k-1)^{2/p}W(k-1) - (k)^{2/p}W(k), \tag{5}$$

and p denotes the number of features in the data set. A value of k is optimal if it maximizes $KL(k)$.

The rest of the section is devoted to the comparison of the performance of the three indices BB, CH and KL on the basis of results obtained for two data sets: an artificial data set and the well known Iris data set.

3.1 An artificial data set

We consider the artificial data set that is represented in Figure 1. This data set is a 200 point sample of a mixture of four normal distributions.

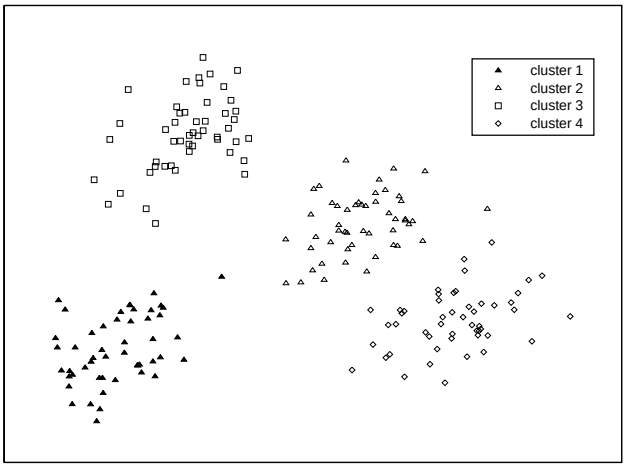

Fig. 1. Artificial data set structured into four clusters.

Each cluster is indeed a 50 point sample of one of the four normal distributions, and except for one point, the four clusters are easily identified by looking at Figure 1. The four normal distributions are centered respectively at $\mu_1 = (-1.5, -.5)$, $\mu_2 = (3, 2)$, $\mu_3 = (0, 4)$ and $\mu_4 = (4.5, 0)$ and have the same variance-covariance matrix $V = .5I$, where I denotes the identity matrix.

This data set was partitioned using the batch K-means method and the stability measures were computed with the ratio sampling $f = 0.8$. The values of the three indices are given in Table 1 for $k \in \{2, 3, 4, 5, 6\}$. The probability significances under (H_0) (p-values) suggest that the 4-partition is the most significant.

	Number of clusters (k)				
Index	2	3	4	5	6
$CH(k)$	145	414	580 *	494	446
$KL(k)$	.26	3.36	3.89	1.39	5.95 *
$BB(k)$	.779	.958	.992 *	.914	.816
Prob. sign. of $BB(k)$ (%)	$48 - 61$	$2.4 - 6.8$	$0 - 1$	$0 - 4.5$	$2.5 - 9.2$

Table 1. Values of the three indices for partitions of the artificial data. According to each index (row), a symbol (*) indicates the optimal numbers of clusters.

Table 2 contains all the cluster stability measures concerning the 4-partition. These values indicate that the four clusters are stable: all the stability measures are high and assessed as being significant under H_0 by low p-values.
Each stability measure in Table 3 was computed with a precision of at least 0.01, which required running the batch K-means method on $N = 140$ samples of the artificial data set. The slight lack of isolation of cluster 2 (.984), just like the slight lack of cohesion of cluster 1 (.980), suggests the presence of an outlier between these two clusters (see Fig.1). The partition into 4 clusters is also identified as optimal by the index CH, but the index KL suggests that $k = 6$ is the optimal number of clusters.

Table 3 presents the stability measures of the 5-partition. Note that with a p-value that is less than 4.5% at a 97.5%-approximate coverage probability, the global validity of the partition into 5 clusters can be deemed as significant. Each of these stability measures were computed with a precision of at least 0.02, and $N = 1500$ samples were necessary in order to obtain this precision. It turns out that the clusters 1, 2 and 3 (which coincide with clusters 3, 4

		Isolation		*Cohesion*		*Validity*	
			%		%		%
Cluster	**1**	.990	$0-1$	.980	$0-5$	.986	$0-1$
	2	.984	$0-1$	.992	$0-2$	.987	$0-1$
	3	1.	$0-1$	1.	$0-1$	1.	$0-1$
	4	.994	$0-1$	.996	$0-2$	.995	$0-1$
Partition		.992	$0-1$	.992	$0-1$	.992	$0-1$

Table 2. Stability measures for the 4-partition (prec. 0.01), and their p-values (%).

and 2 respectively in Figure 1) are clearly stable, for all cluster characteristics except the cohesion of cluster 3. Clusters 4 and 5 (obtained by splitting the cluster 1 of Figure 1 into two clusters) are assessed by low stability values (*i.e.*, .716 and .777) and by high p-values (*i.e.*, in the intervals $34-50\%$ and $22-39\%$). Therefore, their existence is clearly dubious. Stability measures for partial isolation between clusters were also computed: the extremely weak stability measure for partial isolation between cluster 4 and cluster 5 (*i.e.*, -.999) suggests that the split represents more a dissection than a real cluster structure involving separate and homogeneous clusters.

		Isolation		*Cohesion*		*Validity*	
			%		%		%
Cluster	**1**	.993	$0-1$	.939	$0-1$	.973	$0-1$
	2	.993	$0-1$	.936	$0-1$	.972	$0-1$
	3	.989	$0-5$	.873	$1-13$	.945	$0-8$
	4	.696	$32-49$	.798	$48-65$	.716	$34-50$
	5	.727	$29-47$	.980	$1-9$	.777	$22-39$
Partition		.915	$0-4.5$	.913	$0-1$	.914	$0-4.5$

Table 3. Stability measures (prec. 0.01) of the 5-partition, and their p-values (%).

3.2 Iris data

The famous Iris data set reports four characteristics of 3 species namely the iris setosa, versicolor and virginica. Each class contains 50 instances. One class (namely, the virginica) is linearly separable from the others, but the latter are not linearly separable from each other. Iris data were partitioned using the batch K-means method, taking into account only the two variables petal length and width. As in the previous subsection, we have set the value of the ratio sampling f to 0.8.

		Number of clusters (k)			
Index		2	3	4	5
$CH(k)$		756	1211	1266	1358^*
$KL(k)$		4.83	6.01^*	1.3	1.12
$BB(k)$		$.992^*$	.959	.881	.900
Prob. signif. of $BB(k)$	(%)	$.3 - 3.4$	$6.7 - 11.9$	> 34	$5.2 - 9.4$

Table 4. Values of the indices on Iris data partitions. According to each index (row), a symbol (*) indicates an optimal number of clusters.

Table 4 shows the values of the three indices used for choosing the optimal number of clusters on Iris data. The 2-partition with a p-value between .3 and 3.4% is the most stable partition according to the index BB, followed by the 5-partition and the 3-partition with p-values in the intervals $5.2 - 9.4\%$ and $6.7 - 11.9\%$, respectively. Even if the p-values of the last two partitions do not differ significantly, the large p-values of the stability measures of two clusters of the 5-partition (*i.e.*, in the intervals $39 - 53\%$ and $52 - 65\%$) raise doubts about the validity of this partition (see also [Bertrand and Bel Mufti, 2005]). The stability measure BB is the only one to identify the trivial partition in two clusters, and the KL index identifies the 3-partition as the optimal one. Choosing the 5-partition, the index CH is the worst performer on the Iris data set.

4 Conclusion

The results presented in this paper confirm that measuring cluster stability can be a valuable approach to determine the 'correct' number of clusters of any partition. A real advantage of this general approach is that it does not require selecting or using any measure of adequation between the data set and the partition examined.

It can be noticed that the p-values for assessing the measures of cluster stability may be decisive when estimating the stability of clusters. For example, the p-values of Table 1 show that the stability value .915, which assesses the stability of the 5-partition, is statistically more significant under the null hypothesis of absence of structure, than the stability value .958 which assesses the stability of the 3-partition. In addition, an advantage of the stability based approach that is proposed in [Bertrand and Bel Mufti, 2005] is that a careful interpretation of the p-values of the stability measures enables one to identify not only a pertinent partition but also several sources of variation in the partitional stability, such as individual cluster isolation and cohesion.

References

[Bailey and Dubes, 1982]T. A. Bailey and R. Dubes. Cluster validity profiles. *Pattern Reconition* 15, 61–83, 1982.

[Ben-Hur *et al.*, 2002]A. Ben-Hur, A. Elisseeff and I. Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing* 7, 6–17, 2002.

[Bertrand and Bel Mufti, 2005]P. Bertrand and G. Bel Mufti. Loevinger's measures of rule quality for assessing cluster stability. *Computational Statistics and Data Analysis*, 2005, to appear.

[Calinski and Harabasz, 1974]R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics* 3, 1–27, 1974.

[Cheng and Milligan, 1996]R. Cheng and G. W. Milligan. Measuring the influence of individual data points in a cluster analysis. *J. Classification* 13, 315–335, 1996.

[Gordon, 1994]A. D. Gordon. Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis* 18, 561–581, 1994.

[Gordon, 1999]A. D. Gordon. *Classification*. Chapman & Hall, 1999.

[Jain and Dubes, 1988]A. K. Jain and R. Dubes. *Algorithms for clustering data.* Prentice-Hall, Englewood Cliffs, NJ, 1988.

[Krzanowski and Lai, 1985]W. J. Krzanowski and Y. T. Lai. A criterion for determing the number of groups in a data set using sum-of-squares clustering. *Biometrics* 44, 23–34, 1985.

[Lenca *et al.*, 2003]P. Lenca, P. Meyer, B. Vaillant and S. Lallich. Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information (Entreposage et Fouille de données)*, 1, 123–134, 2003.

[Levine and Domany, 2001]E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Comput.* 13, 2573–2593, 2001.

[Loevinger, 1947]J. Loevinger. A systemic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61 (4), 1947.

[Milligan and Cooper, 1985]G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50, 159–179, 1985.

[Milligan, 1996]G. W. Milligan. Clustering validation: results and implications for applied analyses. In P. Arabie, L. J. Hubert and G. De Soete, editors,

Clustering and Classification. Word Scientific Publ., River Edge, NJ, pp. 341-375, 1996.

[Tibshirani *et al.*, 2001]R. Tibshirani, G. Walther, D. Botstein and P. Brown. Cluster validation by prediction strength. *Stanford Technical Report, Department of Statistics, Stanford University, USA*, 2001.