Probabilistic Approach for Comparing Partitions

Osvaldo Silva[1], H. Bacelar-Nicolau[2], Fernando C. Nicolau[3], and Áurea Sousa[4]

[1] Dep. of Math., CES-UA, University of Azores, Ponta Delgada, Azores, Portugal
(Email: osilva@uac.pt)
[2] Faculty of Psychology, LEAD; ISAMB, CEA; University of Lisbon, Lisboa, Portugal
(Email: hbacelar@fp.ul.pt)
[3]FCT, Department of Mathematics, New University of Lisbon, Monte da Caparica, Portugal
(Email: fernandonicolau@netcabo.pt)
[4] Dep. of Math., CEEAplA, University of Azores, Ponta Delgada, Azores, Portugal
(Email: aurea@uac.pt)

Abstract: The comparison of two partitions in Cluster Analysis can be performed using various classical coefficients (or indexes) in the context of three approaches (based, respectively, on the count of pairs, on the pairing of the classes and on the variation of information). However, different indexes usually highlight different peculiarities of the partitions to compare. Moreover, these coefficients may have different variation ranges or they do not vary in the predicted interval, but rather only in one of their subintervals. Furthermore, there is a great diversity of validation techniques capable of assisting in the choice of the best partitioning of the elements to be classified, but in general each one tends to favour a certain kind of algorithm. Thus, it is useful to find ways to compare the results obtained using different approaches. In order to assist this assessment, a probabilistic approach to comparing partitions is presented and exemplified. This approach, based on the *VL (Validity Linkage) Similarity*, has the advantage, among others, of standardizing the measurement scales in a unique probabilistic scale. In this work, the partitions obtained from the agglomerative hierarchical cluster analysis of a dataset in the field of teaching are evaluated using classical and probabilistic (of *VL* type) indexes, and the obtained results are compared.

Keywords: Hierarchical cluster analysis, comparing partitions, affinity coefficient, *VL* methodology

1 Introduction

The Cluster Analysis aims to identify groups (classes or clusters) of entities (individuals, objects, etc.), that are relatively homogeneous and well separated, based on similarities or dissimilarities between them.
There are multiple indexes for comparing partitions, which complicates the decision-making, given that different indexes generally evaluate different peculiarities of the partitions to compare. Moreover, there is a great diversity of validation techniques capable of assisting in the choice of the best partitioning of the elements to be classified, but in general each one of them tends to favour a certain kind of algorithm.

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

Thus, it is imperative to find ways to compare the results obtained using different approaches.
In Section 2 the indexes for the comparison of partitions are introduced using classical coefficients. Section 3 is dedicated to the comparison of partitions using probabilistic coefficients of the *VL* type. In Section 4, we compare the results obtained with the implementation of the two approaches, classical and probabilistic, to a real data set, under a wider validation work in Cluster Analysis, using resampling methods. Finally, Section 5 presents the main conclusions.

2 Coefficients for comparison of partitions pairs

The comparison of two partitions in Cluster Analysis can be performed using various indexes or classical coefficients in the context of three approaches (based respectively on the count of pairs, on the pairing of the classes and on the variation of information). However, each of these coefficients assumes a certain value, depending on its analytic expression, and some have different variation ranges or they do not vary in the predicted interval, but rather only in one of its subintervals. In order for these coefficients to be more easily comparable, one should keep in mind their intrinsic characteristics, categorizing them into groups with similar characteristics.
To compare two partitions, P and P', of one same dataset of n cardinal based on the count of pairs, one can begin by constructing an associated 2×2 contingency table, as Table 1.

Table 1. Contingency table based on the count of pairs

	Partition P'	
Partition P	a	b
	c	d

Table 1 mentions the pairs of elements that exist in the two partitions, where "a" is the number of pairs of elements that are in the same classes in both partitions, "b" is the number of pairs of elements that belong to the same classes in a partition P but to different classes in the other partition (P'), "c" is the number of pairs belonging to different classes in the P partition and to the same classes in the P' partition and "d" is the number of pairs of elements belonging to different classes in both partitions. The total number of pairs of objects is $a + b + c + d = n \times (n-1) / 2$.
Silva (2012) contains a list of indexes for the comparison of binary data, which are functions of the four values of Table 1 and are also used for comparing partitions. In this list, the indexes are subdivided into similarity coefficients that consider the joint absence "d", similarity coefficients that do not take into account the joint absence "d" and other coefficients of association. For each of the coefficients the respective formula

is shown, as well as the symbol with which it is usually designated, its variation range and its author(s).

These indexes should be evaluated relatively to common properties, and can be sensitive to the number of classes in the partitions. Some of the indexes (for example, Hubbert and Rand) tend to have high values in the case of partitions with more classes, others in the case of partitions with a small number of classes (e.g. Jaccard). The adjusted Rand index has none of these undesirable characteristics (Milligan and Cooper, 1985; Jain and Dubes, 1988), which is why this is one of the indexes pertaining to the methodology used in this work. The standardized Ochiai index (a particular case of the affinity coefficient, e. g., Bacelar-Nicolau, 1985), has also been used with good results in the context of partitions comparison (Silva, 2004; Silva, 2012).

As noted above, the evaluation of the partitions comparison indexes based on the count of pairs must take into account the scale of variation and the relation that can be established between the various indexes from their mathematical expressions. Several studies of classification and comparison of these coefficients have been proposed by many authors since Sneath and Sokal (1963). Sibson (1972) made the grouping of coefficients into monotonic classes, establishing an equivalence relation in the set of comparison coefficients for binary data. Bacelar-Nicolau (1980, 1987) determined "*distributionally equivalent*" classes of coefficients, a concept that we will use in this work, as mentioned in the next section.

3 Comparison of pairs of partitions using probabilistic coefficients

Lerman (1970) proposed the use of a similarity coefficient of probabilistic nature between binary variables, which he then expanded to proximity coefficients between structures of the same type (Lerman, 1973, 1981). Bacelar-Nicolau (e.g., 1980, 1987) conducted a distributional study of the comparison coefficients for binary data, having verified and proved the distributional equivalence of a broad class of coefficients, under the assumption of fixed margins of the 2×2 contingency table associated with each pair of elements of the set to be classified. For other coefficients as well as in the hypothesis of free margins, although distributional exact equivalence does not occur, we can find classes of equivalent coefficients with respect to their asymptotic distribution, and take always, as information associated with a coefficient, its limit function of distribution (Bacelar-Nicolau, 1980, 1987; Lerman, 1981), which is a probabilistic similarity coefficient γ on the scope of *VL* methodology. Thus, we have for a similarity coefficient, S:

$$\gamma = F_S(s) = Prob_{H_0}\left(S \leq s\right) \cong Prob_{H_0}(S^* \leq s^*) \cong \phi\left(s^*\right)$$

where H_0 is an adequate reference hypothesis, F_S is the distribution function of S, $S^* = (S - E(S))/\sigma_S$, s^* is a realization of S^*, ϕ is the distribution function of the standard normal distribution and $E(S)$ and σ_S are respectively the mean value and the standard deviation, usually asymptotic. The probabilistic coefficient takes values in [0,1] (follows the Uniform distribution (0, 1)), and is generally calculated asymptotically because the exact distribution function may not be known. The *VL* coefficient was later extended to other types of data and to mixtures of different types of data (e.g. Bacelar-Nicolau, 1988, Nicolau, 1983; Nicolau and Bacelar-Nicolau, 1998; Bacelar-Nicolau *et al*, 2009, 2010).

The approach to comparing partitions, using probabilistic coefficients of the *VL* type, is based on studies of the comparison coefficients for binary data by Bacelar-Nicolau and proceeds as follows:

i) We start with a similarity index, S, for comparing two partitions, *P* and *P'*, based on the count of pairs of elements that exist in the two partitions.

ii) We calculate the value of $\gamma_{PP'}$ of the distribution function of the similarity index *S* used in point *s*, under the assumption of the considered reference:

$$\gamma_{PP'} = F_S(s) = Prob_{H_0}(S \le s) \cong Prob_{H_0}(S^* \le s^*) \cong \phi(s^*)$$

Two partitions, *P* and *P'*, will be considered the more consistent the larger is the value of $F_S(s)$, that is, the more unlikely is overcoming the *s* realization of *S* under the reference hypothesis.

As it has been pointed out by several authors (e.g., Lerman, 1973, 1981; Bacelar-Nicolau, 1980, 1987; Dubes and Jain, 1988), the different indexes do not show all values in [0, 1] and a proportion of the similarity between both partitions is assigned randomly. However, it is shown that the most used indexes are equivalent from the distributional point of view (Bacelar-Nicolau, 1980, 1987). The application of the *VL* methodology to these coefficients allows us to obtain comparison indexes of partitions that can be interpreted on a probabilistic scale. Thus, using a probabilistic coefficient we can choose only one classical coefficient in each (asymptotically) distributionally equivalent class of coefficients, in order to compare partitions.

4 Comparison of results obtained by classical and probabilistic approaches on a set of real data

The data (from a sample of 164 students) was obtained through a questionnaire containing twenty-two questions concerning attitudes/beliefs of students in the area of Social and Human Sciences regarding the subject of Statistics (Silva *et al*., 2007). Each

student selected one and only one of seven possible answers to each question (1 - *strongly disagree*, ..., 4 - *neither agree nor disagree*, ..., 7 - *strongly agree*).
An Agglomerative Hierarchical Cluster Analysis (*AHCA*), using the affinity coefficient (e.g., Bacelar-Nicolau, 1985) between variables and the probabilistic aggregation criteria *AVL*, *AV1* and *AVB* (e.g., Nicolau, 1983; Bacelar-Nicolau, 1985; Nicolau and Bacelar-Nicolau, 1998), was applied in order to obtain a typology of variables under study. The tables with the values of validation indexes to select the most significant partition, obtained from the initial data matrix, and the interpretation of the classes corresponding to this partition, in four classes, can be found in Silva et al. (2007). It has been noted that the most significant partition is the same for all three aggregation criteria.
The results were obtained for the case of evaluation and comparison of partitions using resampling methods. In the present study, we evaluate the most significant partition provided by the AHCA of the data, based on the affinity coefficient and on the aggregation criteria mentioned above. The procedure can briefly be described as follows: 1) from the original data 50 subsamples were generated, with a sampling rate defined *a priori* (80%), using simple random sampling; 2) the same model of AHCA was applied to data matrixes (subsamples) randomly generated by the Monte Carlo simulation method, to determine the partitions with the same number of classes presented by the most significant partition obtained from the original data; 3) this partition was compared to each of the partitions obtained in 2), based on the count of pairs, using each of the classical coefficients from the list in Silva (2012) or the associated *VL* probabilistic coefficient; statistics were also calculated regarding location and dispersion associated with each index, in order to analyze the respective behaviour.
Table 2 shows the values of summary statistics (measures of central tendency, dispersion and quantiles) for classical coefficients (Table 2-a) and probabilistic coefficients (Table 2-b) in the situation where joint absence "*d*" is not considered.
Silva (2012) also contains similar tables for the coefficients where the joint absence "*d*" is considered, as well as for other association coefficients.
It can be seen in Table 2-a) and in Silva (2012) that the most part of the classical comparison coefficients takes values in the interval [0,1]. However, the interval between the minimum and maximum values of the sampling distribution is very variable. The maximum value of the distribution reaches the upper limit 1 of the range in many of the coefficients, reaching a minimum often above 0.5 for the first two considered coefficients classes (similarity coefficients that consider the joint absence "d" and similarity coefficients that do not consider the joint absence "d"), but not for other association coefficients (Silva, 2012).

Table 2-a). Values of summary statistics related to classical coefficients that do not consider joint absence "*d*"

	S	J	O	CZ	K1	DW1	DW2	SS2	BB1	BB2	SO	JO	K2	FMG1
Min	.609	.432	.607	.603	.609	.547	.609	.276	.547	354	.299	1.219	.761	.544
Max	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	500	1.000	2.000	10.333	.938
Mean	.955	.908	.941	.941	.942	.934	.950	.870	.929	479	.891	1.884	2.491	.879
SD	.095	.177	.117	.119	.116	.138	.097	.238	.139	043	.222	.233	3.769	.118
Center	.804	.716	.803	.801	.804	.773	.804	.638	.773	427	.650	1.609	5.505	.741
.005	.609	.432	.607	.603	.609	.547	.609	.276	.547	354	.299	1.219	.761	.544
.01	.609	.432	.607	.603	.609	.547	.609	.276	.547	000	.299	1.219	.761	.544
.025	.673	.438	.609	.609	.610	.609	.673	.281	.609	379	.371	1.220	.761	.547
.05	.765	.513	.683	.678	.687	.609	.765	.345	.609	379	.371	1.374	.761	.620
.1	.765	.513	.683	.678	.687	.609	.765	.345	.609	379	.371	1.374	.770	.620
.25	.969	.912	.954	.954	.954	.969	.939	.838	.939	484	.938	1.908	1.054	.891
.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	500	1.000	2.000	2.491	.938
.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	500	1.000	2.000	5.798	.938
.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	500	1.000	2.000	10.333	.938
.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	500	1.000	2.000	10.333	.938
.975	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	500	1.000	2.000	10.333	.938
.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	500	1.000	2.000	10.333	.938
.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	500	1.000	2.000	10.333	.938

Similarly, measurements of location and dispersion of various classical coefficients show high variation. However, the sampling distributions of the associated probabilistic coefficients described in Table 2-b) feature ranges of similar magnitude with approximate minimum and maximum values.

Table 2-b). Values of summary statistics related to probabilistic coefficients that do not consider joint absence "d"

	S	J	O	CZ	K1	DW1	DW2	SS2	BB1	BB2	SO	JO	K2	FMG1
Min	.000	.004	.002	.002	.002	.003	.000	.006	.003	.002	.004	.374	.194	.002
Max	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691
Mean	.559	.557	.559	.559	.559	.561	.555	.554	.558	.560	.560	.438	.463	.559
SD	.233	.249	.242	.242	.242	.238	.249	.261	.247	.241	.242	.152	.305	.242
Center	.341	.351	.347	.347	.347	.343	.349	.357	.349	.346	.346	.687	.779	.347
.005	.000	.004	.002	.002	.002	.003	.000	.006	.003	.002	.004	.374	.194	.002
.01	.000	.004	.002	.002	.002	.003	.000	.006	.003	.000	.004	.374	.194	.002
.025	.002	.004	.002	.003	.002	.009	.002	.007	.011	.010	.009	.374	.194	.002
.05	.023	.013	.014	.013	.014	.009	.029	.014	.011	.010	.009	.374	.194	.014
.1	.023	.013	.014	.013	.014	.009	.029	.014	.011	.010	.009	.374	.195	.014
.25	.559	.509	.543	.544	.541	.600	.456	.447	.529	.550	.583	.374	.217	.540
.5	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	.374	.360	.691
.75	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	.479	.365	.691
.9	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	.530	.702	.691
.95	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691
.975	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691
.990	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691
.995	.682	.698	.691	.691	.691	.683	.697	.707	.696	.691	.688	1.000	.962	.691

Figure 1 illustrates the variation of mean values of some coefficients, both classic (taking values in the range [0,1]) and probabilistic. As it can be seen, the mean values of the classical indexes, considering the values obtained in 50 resamplings, vary greatly from index to index.

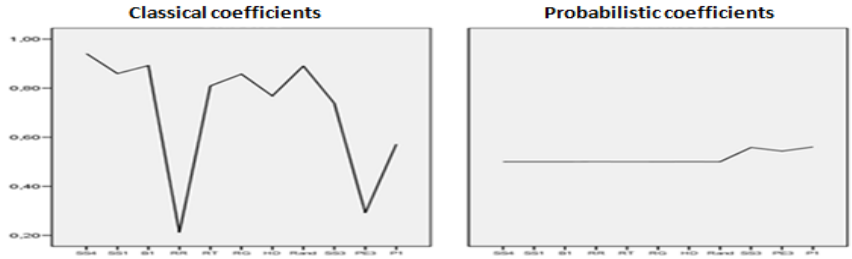

Fig. 1. Variation of means obtained for some classical and probabilistic coefficients

In the context of the *VL* approach it is found that, contrary to the respective basic indexes, the values obtained for means and other location measures of the sampling distribution of the probabilistic coefficient have been very close, as can be seen in Figures 1 and 2, as well as in Silva (2012).

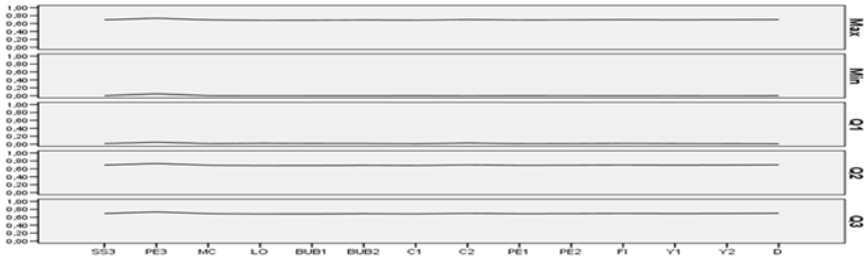

Fig. 2. Variation of the values of some summary statistics for the probabilistic *VL* coefficients associated with classical association coefficients.

These results are consistent with the theory that shows the property of (exact or asymptotic) distributional equivalence between comparison coefficients for binary data (Bacelar-Nicolau, 1980, 1987), mentioned in Section 3. The comparison of partitions using probabilistic coefficients of *VL* type is therefore a simpler and more robust approach than the comparison based on classical coefficients: instead of determining several of these indexes, we will choose a single index in each of the (exact or asymptotically) distributionally equivalent classes and use the *VL* probabilistic coefficient associated to it, which also has the advantage of standardizing the measurement scale on the same probabilistic scale. Finally, the variation ranges and

other statistics provided by the *VL* coefficient tables allow us evaluate the quality of the most significant partition provided by the three models of probabilistic classification. This conclusion is supported by an appropriate set of validation coefficients, which are not presented in this work.

5 Conclusions

In this work, we compare the performances of classical indexes with an associated probabilistic approach for the comparison of pairs of partitions. The described resampling methodology is part of a work on the evaluation of the stability of the obtained classifications in a AHCA.
Usual classical indexes show not to be a convenient choice since they may have distinct display ranges as well as quite different values for other statistics of location and dispersion or they do not take values in the entire variation interval but only in part of that interval. The probabilistic approach to the comparison of partitions using probabilistic coefficients of *VL* type has, among others, the advantage that all classical indexes used lead, exactly or asymptotically, to very close values of the probabilistic index (theoretically conduct to the same value, in the case of the reference hypothesis considered here) and in a probabilistic scale (0, 1). Thus, instead of determining various indexes the *VL* approach can be applied to any of the indexes belonging to a given class of distributionally equivalent indexes to carry out the comparison of partitions pairs with the same number of classes.

References

1. Bacelar-Nicolau, H., Contributions to the Study of Comparison Coefficients in Cluster Analysis, PhD Thesis (in Portuguese), Universidade de Lisboa (1980).
2. Bacelar-Nicolau, H., The Affinity Coefficient in Cluster Analysis, Methods of Operations Research, vol. 53, Martin J. Bekmann et al (ed.), Verlag Anton Hain, Munchen, pp. 507-512 (1985).
3. Bacelar-Nicolau, H., On the Distribution equivalence in Cluster Analysis, In Proceedings of the NATO ASI on Pattern Recognition Theory and Applications, Springer - Verlag, New York, pp. 73-79 (1987).
4. Bacelar-Nicolau, H., Two Probabilistic Models for Classification of Variables in Frequency Tables, In: Bock, H.-H. (Eds.), Classification and Related Methods of Data Analysis, North Holland, pp. 181-186 (1988).
5. Bacelar-Nicolau, H.; Nicolau, F.C.; Sousa, Á.; Bacelar-Nicolau, L., Measuring Similarity of Complex and Heterogeneous Data in Clustering of Large Data Sets, Biocybernetics and Biomedical Engineering, vol. 29, no. 2, pp. 9-18 (2009).

6. Bacelar-Nicolau, H.; Nicolau, F.C.; Sousa, Á.; Bacelar-Nicolau, L., Clustering Complex Heterogeneous Data Using a Probabilistic Approach, In *Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference* (SMTDA2010), 85-93 (2010) (electronic publication) .
7. Jain, A. K.; Dubes, R. C., *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ (1988).
8. Lerman, I. C., *Les Bases de la Classification Automatique*. Paris, Gauth.-Villars (1970).
9. Lerman, I. C., Étude Distributionnelle de Statistiques de Proximité entre Structures Algébriques Finies de Même Type – application à la Classification Automatique. In: *Cahiers du B.U.R.O.*, N°. 19, Paris (1973).
10. Lerman, I. C., Classification et Analyse Ordinale des Données, Dunod, Paris (1981).
11. 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).
12. Nicolau, F.C., Cluster Analysis and Distribution Function, Methods of Operations Research, vol. 45, 431-433 (1983).
13. Nicolau, F.C. and Bacelar-Nicolau, H., Some Trends in the Classification of Variables, In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., Baba, Y. (Eds.), Data Science, Classification, and Related Methods. Springer-Verlag, pp. 89-98 (1998).
14. Sibson, R., Multidimensional Scalling in Theory and Pratice. In: *Les Méthodes Mathématiques de l`Archéologie*, Centre d`Analyse Documentaire pour l`Archéologie, Marseille, 43-73 (1972).
15. Silva, A., Saporta, G. e Bacelar-Nicolau, H., *Missing Data and Imputation Methods in Partition of Variables*. Classification, Clustering and Data Mining Applications, Springer, 631-637 (2004).
16. Silva O.; Bacelar-Nicolau, H. e Nicolau, F., Utilização da Análise Classificatória para Avaliar as Atitudes/Crenças em relação à Estatística de Alunos da Área de Ciências Sociais e Humanas. In: *Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística*, (Ferrão, M. et al. Eds.) Edições S.P.E, 751-759 (2007).
17. Silva, O., Contributions to the Evaluation and Comparison of Partitions in Cluster Analysis, PhD Thesis (in Portuguese), Universidade dos Açores, Ponta Delgada (2012).
18. Sneath, P. H.; Sokal, R. R., *Principles of Numerical Taxonomy*. Freeman, San Francisco (1963).