Teaching and Learning Hierarchical Clustering Probabilistic Models for Categorical Data

Fernando Costa Nicolau,
Department of Mathematics and CMA-FCT, New University of Lisbon, Portugal
cladlead@fpce.ul.pt

Helena Bacelar-Nicolau,
Laboratory of Statistics and Data Analysis-FPCE and CEA,University of Lisbon, Portugal
hbacelar@fpce.ul.pt

1. On the VAL hierarchical clustering probabilistic models

Cluster analysis or classification concerns a set of multivariate methods and techniques for grouping data units (subjects, samples, subsets,…) or variables into clusters of similar elements. When we are dealing with data issued from psychology, education, economy and related areas we are often concerned with the classification of variables, searching for typologies or hierarchic typologies. Moreover in human sciences we frequently have to analyse and classify categorical data, either in an exploratory or in a confirmatory context. Exploratory analysis of such kind of data often has to be with extracting relevant (hidden) knowledge from large questionnaires and surveys, in the form of hierarchical structures, where non–homogenous data sometimes come out. Thus we need to develop and use robust and flexible classification methods and techniques. Furthermore we need to know about the quality or validity of the clustering results.

In this paper we refer to hierarchical clustering probabilistic models for the classification of variables, based on the affinity coefficient (e.g. Bacelar-Nicolau, 1980, 1985, 1988, 2000, 2002). After the general presentation included in the present section we just point out some features and proprieties of the probabilistic cluster approach concerning the measurement of the relationships within a set of categorical variables, which appear to be rather relevant to make students and/or users think about, from our teaching and training experience. The affinity coefficient was earlier introduced in the inferential statistics context by the pioneer work of K. Matusita, started in 1955.

The observed data can be represented in a bi-dimensional matrix, where rows describe data units and columns describe categorical variables (generalization to more complex data can be found for instance in Bock and Diday, 2000 and Bacelar-Nicolau, 2002). Empirical clustering models, are usually used to analyse such data, where one has in a first step to choose an appropriate proximity (similarity or dissimilarity) coefficient to measure the relationship between pairs of elements within the data set to classify, in a second step to define an aggregation criterion for merging similar clusters of elements and in a third step to use some way to assess the validity of the clustering results. However taking in account some natural reference hypothesis concerning the data knowledge may allow us to apply more appropriate probabilistic models: in a first step we compute normalized and/or (exactly or asymptotically) standardized similarity coefficients, in a second step we may apply probabilistic similarity coefficients that are measured in a probability scale, instead of simple/basic similarity coefficients and in a third step we select an aggregation (empirical or probabilistic) criterion. Therefore in such probabilistic models the step of assessing the validity is already (at least partially) included inside the probabilistic model. In this context a probabilistic coefficient is a measure that validates the corresponding linkage value assumed by a basic coefficient. Thus we often refer to such coefficients as “Validity Linkage” or VAL coefficients. Note that the probabilistic approach operates in an exploratory context that uses prior major knowledge on the data structure as a tool to extract knowledge on its clustering hierarchical structure (e.g. Bacelar-Nicolau, 1972, 1987, Lerman, 1970, 1981, Nicolau, 1972, 1985). An

aggregation criterion appear in this probabilistic approach often included into some parametric probabilistic formulae, such as Nicolau & Bacelar extension of the Lance & Williams adaptive formula for VAL similarity coefficients and their Nicolau extensions (Nicolau and Bacelar-Nicolau, 1996). Other validation studies particularly stand on internal validation (similarity coefficients comparing a classification with the original data set) and relative validation (similarity coefficients comparing several different classifications of the same data set) in this framework.

Hierarchical clustering probabilistic and non-probabilistic models have been teached, trained and discussed with different kinds of students and users, on different theoretical and applied perspectives. As always such kind of debate appear to be a source of enrichment for instance in what concerns understanding and assessment of concepts, techniques and methods used in confirmatory and/or in exploratory multivariate analysis. Applied research presently being developed on the VAL approach by our groups at the CMA-FCT, LEAD-FPCE and CEA-UL takes in account missing values effects, asymptotic convergence problems, associated confirmatory multivariate methods and software development (Silva, Bacelar-Nicolau and Saporta, 2001, Sousa Ferreira, Celeux and Bacelar-Nicolau, 1999, Sousa, Silva, Bacelar-Nicolau and Nicolau, 2002, Bacelar-Nicolau and Bacelar-Nicolau, 2001). All these aspects may be discussed with students of cluster analysis methods. Moreover part of this clustering probabilistic approach has been implemented in a software for teaching and training multivariate data analysis, and it is intended to pursue on this way too (Nicolau, Dias, Bacelar-Nicolau, 1998). Indeed (as cited by H.Bozdogan, University of Tennessee, USA, from the wall of NSF …) "research is to teaching as sin is to confession: without the one there is not much to say about the other".

2. Profiles and spherical coordinates

In this paper we just point out some features and proprieties of the VAL approach in the case of categorical data, which appear to be very appropriate to make students and/or users think about, from our teaching and training experience.

Let M(D,V) be a $(k \times p)$ data matrix, where D represents the set of data units and V is a set of p categorical variables. In many practical cases V_j (j=1,...,p) will be a vector of frequencies, a vector of scores or a vector of binary (presence/absence) values.

Let's consider the vector of positive frequencies case; other possibilities may appear as particular cases and extensions of this one. Thus V_j may be first represented by the k coordinates n_{ij} (i=1,…,k), where the column margins are $n_{.j} = \sum_{i=1}^{k} n_{ij}$. We will then refer to the j-th column profile as the corresponding conditional vector with components $n_{ij} / n_{.j}$. This profile vector may be a discrete conditional probability distribution law ; but often it estimates a profile or probability vector of the population, where for instance the set D of k data units represents a strata or a partition of some random sample of subjects in k classes (although we are not directly concerned in this paper with inferential aspects) . The column profiles have a major role in this work, as the similarity between pairs of variables will be measured using an appropriate function, the affinity coefficient, of their profiles.

The simple affinity coefficient between V_j and $V_{j'}$ (j, j´=1,.....,p) can be defined as the inner product between the square root column profiles associated to V_j and $V_{j'}$, that is:

$$a(j, j') = \sum_{i=1}^{k} \sqrt{\frac{n_{ij}}{n_{.j}} \frac{n_{ij'}}{n_{.j'}}}$$

Note that the formula above is straightforwardly adapted to the case where the data units have positive weights w_i, $\sum_{i=1}^{k} w_i = 1$.

It is easy to prove that the affinity coefficient has the following proprieties:

a) it is a symmetric similarity coefficient which takes values in [0,1], 1 for equal or proportional (initial) vectors and 0 for orthogonal vectors.
b) it measures the monotone tendency between the column profiles
c) is related to the Hellinger distance associated to the j –th and j'-th profiles

$$d^2_{a(j,j')} = \sum_i \left(\sqrt{\frac{n_{ij}}{n_{.j}}} - \sqrt{\frac{n_{ij'}}{n_{.j'}}} \right)^2 = \sum_i \left(\sqrt{n_{i/j}} - \sqrt{n_{i/j'}} \right)^2$$

by the usual relation $d^2_a = 2(1-a)$, applied to this particular Euclidean metric.

Moreover the affinity similarity coefficient (and the associated Hellinger distance) between two profiles:

d) is a measure of similarity (a distance, respectively) on the k-sphere with centre at the origin and radius equal to one: using the square root function allow us to obtain a transformed profile column vector with positive spherical coordinates $\sqrt{\frac{n_{ij}}{n_{.j}}}$, i=1,...,k.
e) is not changed if two or more equal or proportional data units are replaced by one single data unit with adjusted proportional contributions to the concerned pair of profiles – "principle of distributional equivalence", also verified by the qui-square distance (Benzecri, 1973) – or if more proportional data units are added to the data matrix
f) is not changed if more profile columns are added to the data matrix (larger V set size)
g) is independent of the D set size
h) is easily extended to practical situations where negative frequencies are present (they may arise in migration and economic studies for instance) ; a generalised affinity coefficient can then be defined by

$$a_g(j,j') = \sum_{i=1}^{k} \operatorname{sign}\left(\frac{n_{ij}}{n_{.j}}\right) \operatorname{sign}\left(\frac{n_{ij'}}{n_{.j'}}\right) \sqrt{\left|\frac{n_{ij}}{n_{.j}} \cdot \frac{n_{ij'}}{n_{.j'}}\right|}$$

where sign means the "signal of" and $|\ |$ means "absolute value". Here we refer to a profile column vector with spherical coordinates on the whole k-sphere. Normalization constraints are, for each column profile: $\sum_i \left|\frac{n_{ij}}{n_{.j}}\right| = 1$, $n_{.j} = \sum_i |n_{ij}|$. The generalised Hellinger distance can be found using the same relationship as above. This is a symmetric similarity coefficient which takes values in [-1,1].

i) for column profiles associated to binary variables the affinity coefficient turns out to be the Ochiai coefficient
j) can be adapted to comparing vectors of scores
k) may be used in hierarchical cluster analysis of non-homogenous data, where categorical variables of the three types mentioned above are present.
l) Can be generalised to real variables; in simulation studies with missing data the affinity coefficient shows a better behaviour than the Pearson correlation coefficient.

These properties (among others related to more complex data) turn out to be real advantages in using the affinity coefficient or the Hellinger distance as the basic coefficient to measure the similarity or the dissimilarity between column profiles. The Hellinger distance was already used as the basic distance of the Spherical Factor Analysis by M. Volle in 1979. In the next section we

introduce in a simple manner the concept of VAL similarity coefficient. Generalization to adaptive families then follows.

3. VAL probabilistic coefficient

Theorem: Under the reference hypothesis underlying the limit theorem of Wald-Wolfowitz (W-W), the affinity coefficient between column profiles a(j, j') is an observed value of a random variable A(j, j´) with asymptotic normal distribution.

A detailed proof of the theorem can be found for instance in Bacelar-Nicolau, 1988. The limit W-W theorem can be found in Fraser, 1975. The above reference hypothesis stand on a permutation test. The asymptotic parameters are computed as functions of the column profiles. Applications were presented for in many of the previously mentioned works.

From this theorem, we define a new random variable $A^*(j, j') = \frac{A(j, j') - \mu(A(j, j'))}{\sigma(A(j, j'))}$ which is asymptotic normal distributed N(0,1), as k grows to infinity. Simulation studies (Nicolau, 1992, Sousa, Silva, Bacelar-Nicolau, Nicolau, 2001) proved that convergence to normality goes very fast; if the asymmetry of the profiles is not too large, k greater than 10 gives already a good normal approximation.

The VAL probabilistic coefficient associated to this permutational model is defined by

$$\alpha(j, j') = \text{Prob}(A^*(j, j') \quad a^*(j, j') \quad \phi(a^*(j, j'))$$

where Φ represents the cumulative distribution function of the standard normal distribution.

The VAL coefficient between two profiles has the following properties

a) is a symmetric similarity coefficient which takes values in [0,1].
b) the greater is the probability to have similarity values lesser than the observed value, the greater is the Validity Link / probabilistic relationship between the pair (j, j') of column profiles.
c) verifies properties e)-g) of the affinity coefficient
d) is also applied to the generalised affinity coefficient by using the same limit W-W theorem
e) for column profiles associated to binary variables we have proved that there are two cluster of similarity coefficients which turn out to be respectively exactly or asymptotically equivalent to the Ochiai coefficient, since they give respectively exactly or asymptotically the same VAL probabilistic coefficient.
f) also verifies properties j)-l) of the affinity coefficient

Again discussion on the probabilistic concepts based on the affinity coefficient, the exploratory context of cluster analysis where they are used and the validity assessment of clustering results may be promoted at this point when we are teaching our courses. Several applications to real life problems are usually presented. Also a large bibliography can be added, analysed and discussed. Some papers were mentioned in the present extended abstract.

RESUMÉ

Nous référons le réseau des propriétés de l'approche probabiliste VAL à la classification hiérarchique, concernant des mesures de similarité et des distances associées basées sur la notion d'affinité entre vecteurs profil (ou de probabilité), aussi bien que la validité intrinsèque à la méthode.