EVALUATION OF HIERARCHICAL CLUSTERBENKAR METHOD

By Dr. Dimitrios Karapistolis

An ascending hierarchical classification of «objects» of a total i with cardinality card(I)=n, is a process which produces a sequence of partitions of the original set in subtotals non-empty and separate from each other, the so-called **classes**, one into the other, uniting every time only two classes that show under some metric at each step of grouping the smallest distance.

It is understood that the aim of ascending hierarchical classification is to group together all the statistical units in a population in a limited number of homogeneous classes, as per the behavior of certain variables, taking into account the total number of the variables so that each one may differ from the other as much as possible.

The classes are created based on an objective algorithm, which differs from the subjective methods which can be developed by any researcher. We say objective algorithm because the grouping of the statistical units is done without any a priori assumption in the original data table and based on a specific metric.

Of course, a table containing scores from various criteria may also include extreme values which must be taken into account in the procedure for the classification of «objects».

The result is that values of «objects» in the various criteria, which are included in a particular class, can vary considerably, something which negates to some extent the homogeneity of the replies in the class. The proposed method seeks to assess precisely the homogeneity of the classes of an ascending hierarchical classification.

Creation of the Ascending Hierarchical Classification

Let be table T(nxp) with n lines and p columns, In each classification regardless of the metric created, classes are formed each containing a population of «objects», which are depicted by the n lines of the data table. So for example in a market research questionnaire on each line of the table correspond the responses of each participant in total p questions which have been posed.

As is well known in a classification with the FACOR method, the Ward algorithm is used, so when we move from a division with l+1 classes in another division which has l classes, by merging two classes in one, with the criterion the reduction of inter-class inactivity in accordance with the theory of Huggens. In addition, we accept that all data are collected at the **center of gravity** of the class, which is **weighted** by the weight of the elements of the k class, which constitutes the barycenter of the class.

The grouping together of two observations or two classes in a class creates what we call a **node** of the hierarchy. Every node of the hierarchy symbolizes the center of gravity of the «objects» participating therein and the description of the classification is done with the dendrogram, whose nodes symbolize the subdivisions of the population.

If not interested in the overall hierarchy of n «objects» but only for a limited number of k classes, we only get one cut of the dendrogram at level e_1, namely to **«cut»** the dendrogram with a straight line at the point where the branches remaining satisfy the number of k classes which we want to maintain.

BENKAR method

As known, until today evaluation methods of a classification can be made with procedures which provide for either the use of neural networks or using classifiers of machine learning, who do not deliver probabilities but only an assessment of the performance of learning for the result obtained.

With the proposed method used in addition to the basic principles of Correspondence Analysis and the properties of Euclidean vector space R^N, given the probability distribution of objects to belong to certain classes of the hierarchy.

In particular, the k nodes (i.e. the k centers of classes) of a specific typology of the hierarchy, are created after first we pool for each column the values of the lines of table T(n,p) belonging to each class and then the k classes are considered as new lines, creating an augmented table T(n+k,p) which we analyze with Correspondence Analysis so to calculate the coordinates of the p-1 factor axes.

Using the coordinates of all the points of the cloud N(I) of the lines, of the cloud N(J) of the columns and the k nodes of the classification on the p-1 factor axes, resulting from the analysis of the table T(n+k,p) we create an orthonormal basis in the space $R^{(p-1)}$, where we place the p variables, the n statistical units, and the centers of k classes, to actual positions from where is drawn all the information supplied by the data table.

Following the use of Euclidean metrics, we can calculate the distances of each statistical unit from the k centers of classes, as suggested by the algorithm of Ward. The k distances of each statistical unit we transform in k probabilities, where the smallest of the k distances, corresponds to a greater likelihood that the statistical unit is close to the center of the class with the smallest distance, where a priori the smallest distance does not always correspond to the class identified by the process of classification using the algorithm of Ward.

This is because when there are statistical units with extreme values included in the class, the **homogeneity** of this class, in relation to the values of other statistical units which it includes is compromised, in so far as the center of each class, as stated above, is weighted by the weight of the elements of the k class in every step of creating a node of the hierarchy.

Therefore, it is appropriate to assess the correct positioning of the statistical units in a defined number of classes of the classification using the maximum probabilities arising from the conversion of the minimum distances of every statistical unit from the centers of the classes under Euclidean metrics.

Subsequently, the distribution of n maximum probabilities in m equal classes are formed, from which results the requested assessment of the original classification with the FACOR method.

By studying the distribution of the maximum probabilities, if the cumulative frequency (which is converted into a percentage) of the last two classes, which determines the number of statistical units which are classified by the two methods in the same classes is relatively small and even if the percentage which defines the scope of the last two classes is sufficiently satisfactory, the classification into k homogeneous classes shall be considered not to be satisfactory because the inconsistency in the classification of objects in k classes indicates that the k classes contain heterogeneous statistical units as to the values of p criteria.

But if the percentage of objects of the last two classes is satisfactory in relation to the percentage which defines the scope of the last two classes of the distribution of probabilities, they determine the evaluation of the ascending hierarchical classification with the FACOR method.

Application of the BENKAR method

For the application of the proposed method we will use a specific questionnaire containing qualitative variables (for the measurement of which was used a five-point Likert Scale, where 5 was the excellent impression), to which responded 1721 people, Part of the questionnaire concerned six questions regarding how foreign visitors grade (a) The sights of the city of Thessaloniki (b) the Greek cuisine (c) The nightlife of the city (d) The architectural style (e) The safety and (f) The friendliness of the locals.

The ascending hierarchical classification with the FACOR process was applied to the data obtained, For the evaluation of the classification and with criterion of partioning l_r we chose the intersection of the dendrogram into five classes. The data related to foreign visitors of Thessaloniki and the data contained in the survey was conducted in the context of the program ARCHIMEDES III titled «Data Analysis Technologies and Knowledge Management in designing tourist products»

The six variables present respectively as follows: Δ4, Δ5, Δ6, Δ7, Δ8, Δ9. Given the classification of 1721 people with the FACOR procedure, Table 1 shows their replies and the five classes to which the respondents belong.

Table 1: Values of the six variables and the five classes to which the respondents belong after the classification with the FACOR procedure

Tags	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9	Class FACOR
I1	4	4	0	4	5	5	3
I2	5	4	5	5	4	5	4
I3	3	4	3	1	2	3	2
.	.	.	.	.	.	.	.
1690	4	3	2	2	2	2	5
‘	‘	‘	‘	‘	‘	‘	‘
1719	5	5	5	2	2	5	2
1720	5	4	4	2	5	3	5
1721	5	5	4	3	4	5	5

Table 2 shows the augmented table created in step 3.

Table 2: Part of the augmented table T(n+k,P)

Tags	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
I1	4	4	0	4	5	5
I2	5	4	5	5	4	5
I3	3	4	3	1	2	3
.	.	.	.	.	.	.
I1690	4	3	2	2	2	2
.	.	.	.	.	.	.
I1721	5	5	4	3	4	5
K1	398	212	224	387	378	421
K2	727	750	810	592	391	620
K3	548	545	35	482	430	521
K4	3679	3612	3192	3600	3403	3437
K5	1580	1883	1907	1548	1788	1963

By following the steps 4 and 5, table T3(n,2k) is created which presents for each point-line the distance K_I (i=1,…,5) of points from the centers of the five classes and their respective probabilities P_I (i=1,...,5)

Table 3: Part of the distances table K_i of points-lines with the corresponding probabilities P_i

IND	K1	P1	K2	P2	K3	P3	K4	P4	K5	P5
I1	12.769	0.077	41.496	0.007	3.816	0.865	20.147	0.031	25.590	0.019
I2	5.770	0.032	3.719	0.078	21.824	0.002	1.264	0.672	2.229	0..216
I3	26.919	0.024	5.980	0.485	29.984	0.019	10.966	0.144	7.280	0.327
.	.	.	.	.	.	.	.	.	.	.
I1690	11.800	0.051	4.899	0.298	12.839	0.043	3.795	0.497	8.065	0.110
‘	.	.	.	.	.	.	.	.	.	.
I1719	25.445	0.019	4.003	0.755	35.039	0.010	12.399	0.079	9.378	0.138
I1720	9.281	0.003	5.912	0.006	21.472	0.001	1.164	0.162	0.514	0.829
I1721	9.253	0.015	4.365	0.069	15.999	0.005	2.087	0.302	1.470	0.609

Then, step 6 gives table T4(n,3) which presents for each point-line in the 1st column the classification with the FACOR procedure, in the 2nd column the new classification with the procedure of the minimum distance, and in the 3rd column the maximum probability to belong to the class corresponding to the minimum distance.

Table 4: Classification with the FACOR procedure and the Euclidean metrics in space R^5

Tags	FACOR	DIS	maxProb
I1	3	3	0.8651
I2	4	4	0.6717
I3	2	2	0.4851
.	.	.	.
I1690	5	4	0.497
‘	.	.	.
I1719	2	2	0.7552
I1720	5	5	0.829
I1721	5	5	0.6086

MAD software gives the following results:

Settled in the SAME classes: 1416 people
Settled in different classes: 305 people
Good adaptation rate: 82.28%

By continuing with step 7 we have the three distributions into five classes which emerged after the classification of the 1721 people

(a) with the ascending hierarchical classification (-CAH-)

(b) with the Euclidean metrics in space R^5 created by five factor axes after the application of Correspondence Analysis at the data table T(1726,6)

(c) a breakdown of the maximum probability of the 1721 people to belong in five different classes of ascending hierarchical classification

Table 5: The three distributions of the 7th step

CAH	n_i	f_i	DIS	n_i	f_i	Probability distribution	ni
K1	99	0.0575	**K1**	136	0.079	**T1: 0.2305 – 0.3855**	48
K2	181	0.1051	**K2**	225	0.1307	**T2: 0.3855 – 0.5408**	230
K3	129	0.0749	**K3**	142	0.0825	**T3: 0.5408 -0.6962**	315
K4	848	0.4927	**K4**	764	0.4439	**T4: 0.6962 – 0.8515**	335
K5	464	0.2696	**K5**	454	0.2638	**T5: 0.8515 – 1.0000**	488
	1721	**1**		**1721**	**1**		**1416**

Remark: The interpretation of class T5 is as follows: 488 people from the 1416 which were settled in the same classes, i.e. rate 34.5% has probability 0.8515 or more to belong to one of the five classes of classification with FACOR method.

Moving on to step 8 assessment of the precise classification of 1721 visitors with the BENKAR method, using the last two classes of probability distribution is as follows: 823 (=488+335) of the 1721 responders, which means 47.82% of the sample has a probability of 0.6962 to belong to ONE of the five classes of the classification.

Because the percentage 47.82% of all respondents is not satisfactory, despite the fact that the percentage 82.28% of the adjustment of the «objects» on five classes of two classifications is high enough, the data in Table T(1721,6) within the classes do not seem to have the necessary desired homogeneity, in relation to the values of the six variables.

By applying then the BENKAR method for how the 1721 visitors of Thessaloniki evaluate the first three questions of the study Δ1= «the cleanliness of the city», Δ2=«the natural beauty» and Δ3=«the values of products and services», we received the following results:

Table 6: Values of the six variables and the five classes to which the responders belong after the classification with the FACOR procedure

Tags	Δ1	Δ2	Δ3	Class FACOR
I1	3	5	4	5
I2	2	5	3	5
I3	1	3	4	5
.	.	.	.	.
1690	2	3	5	5
.	.	.	.	.
1719	1	2	3	5
1720	1	5	2	3
1721	2	5	3	5

MAD software gives the following results:

Settled in the SAME classes: 1638 people

Settled in different classes: 83 people

Good adaptation rate: 95.18%

Table 7: The three distributions of the 7th step

CAH	n_i	f_i	DIS	n_i	f_i	Probability distribution	ni
K1	119	0.0691	K1	135	0.0784	T1: 0.3700 -0. 4970	68
K2	294	0.1708	K2	348	0.2022	T2: 0.4970 – 0.6241	201
K3	221	0.1284	K3	219	0.1272	T3: 0.6241 – 0.7512	161
K4	691	0.4015	K4	681	0.3957	T4: 0.7512 – 0.8784	254
K5	396	0.23	K5	338	0.1963	T5: 0.8784 – 1.0000	954
	1721	1		1721	1		1638

Table 7 shows the following assessment: 1208 of the 1721 responders, which means 70.18% of the sample has a probability of 0.7512 to belong to ONE of the five classes of the classification.

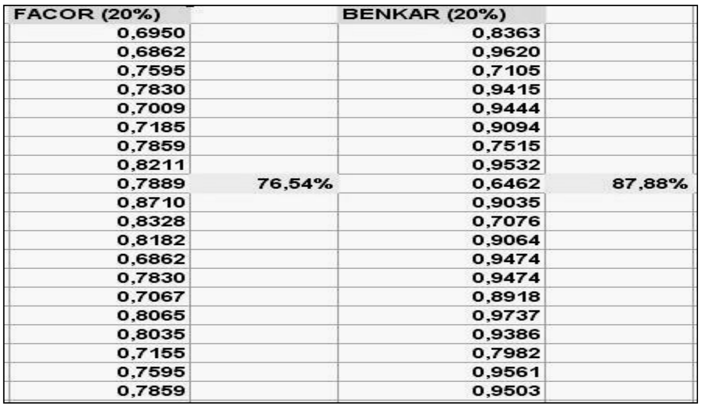
The percentage 70.18% of all respondents is satisfactory, as well as the fact that the percentage 95.18% of the adjustment of the «objects» on five classes of two classifications is high enough, the data in Table T(1721,3) within the classes it may be considered that the data have the necessary desired homogeneity, in relation to the values of the six variables.

Comparing the two assessments of the classifications firstly with the criteria Δ1-Δ3, and on the other hand with the criteria Δ4-Δ9, it emerges that the 1721 visitors of Thessaloniki have a uniform image of the city as to the first three criteria which were invited to assess, while for the other six criteria there were no answers of comparable homogeneity. This may be due to the fact that visitors came from 51 different countries of the world, so it can be deemed normal for them to hold similar views for criteria Δ1–Δ3, while for criteria Δ4-Δ9 visitors have quite different views due to the different traditions in place in their respective countries of origin.

Training of data with the Support Vector Machine -SVM-

Educating with 20 repetitions the data presented in Table 1, with the use of the Support Vector Machine, keeping in each repeat a random sample of 20% of the 1721 values, the one time with the rankings of «objects» with the FACOR method and the other with the BENKAR method gives the following results:

Table 8: Learning rates of 20 repetitions after the training of the classifications of data on the basis of the FACOR method and the BENKAR method.



FACOR (20%)		BENKAR (20%)	
0,6950		0,8363	
0,6862		0,9620	
0,7595		0,7105	
0,7830		0,9415	
0,7009		0,9444	
0,7185		0,9094	
0,7859		0,7515	
0,8211		0,9532	
0,7889	76,54%	0,6462	87,88%
0,8710		0,9035	
0,8328		0,7076	
0,8182		0,9064	
0,6862		0,9474	
0,7830		0,9474	
0,7067		0,8918	
0,8065		0,9737	
0,8035		0,9386	
0,7155		0,7982	
0,7595		0,9561	
0,7859		0,9503	

Table 8 shows that the BENKAR method predominates in the correct classification of objects. This occurs because the percentage assessment of programming the data in Table 1, which concerns the classification with the BENKAR method (87.88%) is higher than that obtained with the ascending hierarchical classification with the FACOR method (76.54 %). In addition to the BENKAR method on 20 repetitions of programming data assessment rates above the average value is much higher (13 to 20 repetitions over 90% with maximum 97.37%) from the respective percentages assessment with the FACOR method where the maximum value is only 87.10%.

Given that as stated above the SVM does not return probabilities, while the BENKAR method calculates the probability of each statistical unit belonging to a certain class, therefore the distribution of the maximum probability resulting from the proposed method may be regarded as objectively assessing the ascending hierarchical classification derived by the FACOR method.

Conclusion

The BENKAR method utilizing objective criteria such as the coordinates of points on the factor axes after the application on the data table of Correspondence Analysis, and the placement under these coordinates in the Euclidean vector space R^P, with the use of Euclidean metrics, provides the capability of the objective evaluation of the homogeneity of the «objects» who participate in the shaping of the classes of ascending hierarchical classification.

This capability of the BENKAR method can be applied to any table T(n,p) for which the data have been classified with any criterion of consolidation, given that in each case the BENKAR method sorts the «objects», after it places them in a orthonormal coordinate system R^p that creates the factor axes after the application of Correspondence Analysis on the data table.

The superiority of the BENKAR method in the classification of «objects» in k classes indicated by the ascending hierarchical classification shall be recorded and with the use of the Support Vector Machine, which assesses the programming of data at a rate higher than that which gives the same machine for the classification of the same data with the ascending hierarchical classification, using the FACORmethod.

In addition, the BENKAR method in contrast with the SVM, but also with any other method of grading, calculates the probability of the «objects» belonging to the classes formed, which is a main advantage of this method, with final conclusion the objective assessment of homogeneity of classes, which is one of the requirements in each classification.