Serdica J. Computing **2** (2008), 19–30

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

APPLICATION OF CLASSIFICATION METHODS TO A PROBLEM RELATED TO SPECIFIC GROUPS OF E-GOVERNMENT USERS*

Vesela Angelova, Avram Eskenazi

Abstract. One of the important tasks of the EU ELOST project on E-government and Low Socio-Economic Status Groups (LSG) was to compare experts' opinions on fundamental problems of the subject. This papers shows how the application of specific classification methods to experts' formalized answers could lead to some non-trivial and objective conclusions about interdependencies and the interrelation between e-government policies/tools and experts' background and country affiliation.

1. Introduction.

1.1. The Problem. This work has been inspired by and became part of the EU SSA No 27287 ELOST project "E-Government for Low Socio-Economic Status Groups (LSG)". The aim of the project derives from the well known fact that the pace at which countries deploy e-Government services, including measures taken to increase their use by LSGs, vary considerably across Europe. In

ACM Computing Classification System (1998): I.5.4.
Key words: classification methods, application, e-government.
*This paper was partly supported by ELOST – a SSA EU project – No 27287.

order to arrive at policy proposals, the project is carrying out several activities, one of them being a cross-national comparative assessment of e-Government services. Various input data were provided to this assessment. One stream of this data came from personal interviews with relevant key actors and decision makers in governments and local authorities in each of the six participating country.

A total number of 41 interviews were carried out. Due to different reasons the answers of only part of them followed a preliminary prepared formalized questionnaire. The aim was to evaluate the E-government policy tools for the six countries by the selected experts from each country. According to the aim, the questionnaire was constructed on the following topics of questions: personal data of the respondent, identification of the representatives of the LSG, policies, tools, future development of e-government tools.

The LSG's chosen after a preliminary analysis and included in the questionnaire are:

a. Unemployed persons
b. People with low/very low income
c. Homeless
d. People with a low education level
e. Immigrants
f. Ethnic minorities
g. Refugees
h. People in isolated or underdeveloped regions
i. Prisoners

The type of the experts-respondents was determined on the basis of the first group of questions. The next group of 5 questions (No 5 to 9) emphasizes the clarification of the LSGs. It looks for answers such as: which groups are the less profited by the digital services; which groups are particularly important with regard to e-government policymaking; what reasons enforce the digital division among people and compel the governments to take measures against the digital inequality.

The third topic relates to the support of the e-government policies for LSGs in the respective country and includes one question (No 10).

The fourth topic includes 4 questions (No 11 to 14) directed to the application of e-government tools by LSG as reality or future plans. The questions pick up information about the sufficiency of the public facilities; to what degree LSG use e-government tools and for what purpose; what forms of participation are available to LSGs, etc.

The last topic consists of 2 questions (No 15 and 16) and investigates the experts' views and expectations for e-government developments in the future as: services; the reflection of the new technologies on e-government or on LSGs; the reduction of the digital division among the people.

Most questions required an answer from a scale ranking from 0 to 3. Only the experts from Bulgaria and Germany, as well as one of Finland complied, the others preferred to answer in a free form. Hence, the answers of the first subset of experts were further used in our investigation. The aim was to identify, if possible, clusters (groups) of tools having led to identical or similar results, or clusters of countries applying similar policies/practices/tools, or more generally speaking, to establish in a more objective way important interdependencies between countries, tools and results.

1.2. The methods. As is well known, pattern recognition aims at classifying data (patterns) based on either a priori knowledge or statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait—often proximity according to some defined distance measure. Elements within a cluster should ideally be as homogeneous as possible. But there should be heterogeneity between clusters. Various methods are known to find out clusters for a given set of data.

The approach we proposed and used here (as brief described in 2. below) is particularly suitable to our case – a relatively small set of patterns – experts with their opinions/answers. Obviously, such a set of less than 30 objects could hardly be investigated with statistical methods. This is why we applied our classification method, based on the so-called test approach. (Juravlev [1] and Kudriavtzev are considered to be the pioneers of this approach, further refined in specific directions [2], [3]). The main advantage of the method in our case is that it is possible to obtain objective and reliable results by using relatively small subsets of patterns, in any case even below 10.

Among the well known other possible methods are the following

I. K-Means

K-Means(X,k) partitions the points in the m-by-n data matrix X into k clusters. (The number k needs to be determined at the onset.) The goal is to divide the objects into k clusters such that some metric relative to the centroids of

the clusters is minimized. This iterative partitioning minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. Rows of X correspond to points, columns correspond to variables, or in other words the data are a list of n-measurement vectors. K-Means computes clusters differently for the different supported distance measures [4]:

- 'sqEuclidean'—Squared Euclidean distance (default). Each centroid is the mean of the points in that cluster.
- 'city-block'—Sum of absolute differences, i.e., the L1-distance. Each centroid is the component-wise median of the points in that cluster
- 'cosine'—One minus the cosine of the included angle between points (treated as vectors). Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length.
- 'correlation'—One minus the sample correlation between points (treated as sequences of values). Each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation.
- 'Hamming'—Percentage of bits that differ (only suitable for binary data). Each centroid is the component-wise median of points in that cluster.

In other applications K-Means returns a clustering of data, given the initial estimates of the cluster centres (seeds) and weights for each element of measurement vectors in data.

II. The Jaccard index

The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the sample sets.

Given two objects, A and B, each with n binary attributes, the Jaccard coefficient is a useful measure of the overlap that A and B share with their attributes.

III. The Dice coefficient and similarities

The association measures are maximum-likelihood estimates for various coefficients of association strength. As such, all the measures are subject to large sampling errors, especially for low-frequency data. One measure is interesting because of its similarity to mutual information. The best-known coefficient from this group is the Dice coefficient, the correlation between two discrete events, e.g., the extraction of collocations from text corpora.

Both the Dice coefficient and the similar Jaccard coefficient have found widespread use in the field of information retrieval. The Dice and Jaccard measures are fully equivalent, i.e., there is a monotonic transformation between their scores [5].

2. Our Approach.

We have proposed another approach: 8 algorithms to determine a measure to a cluster (without a centroid). They are based on the terms *Irreducible test* and *Irreducible representative set*. This approach has been successfully applied to various areas, particularly to evaluate the quality of different software products. These two sets of objects are extracted from a teaching table for clustering (m–by–n data, divided in k clusters). Integer measurements, not only binary, are possible:

An Irreducible test is a minimal subset of variables, for which the measurements in vectors for different clusters are different.

An Irreducible representative set for the cluster Ki is a minimal subset of variables, for which the values (measurements) in the vectors of Ki and all others vectors are different.

Hence, the tests are dissimilarity units for all clusters, the representative sets are dissimilarity units for one cluster and all another clusters.

Example: Let suppose that e1=(1,1,0,1) and e2=(0,0,0,1) belong to the cluster K1, e3=(0,1,1,0) belongs to K2.

Cluster	Sample	c1	c2	c3	c4
K1	E1	1	1	0	1
	E2	0	0	0	1
K2	E3	0	1	1	0

Teaching table

Then all the tests are {1,2},{3},{4} for which the measurements are different in K1 and K2.

With respect of the membership of a given variable in such units, this variable receives a certain weight to reflect its contribution to the diversity of the vectors in the teaching table.

We propose 3 types of weights:

- p_i (part of all tests whose member is the variable i),
- q_i (the same with respect of the length of the test),

- r_i (the same applied to the representative sets).

By definition the vectors in the cluster K of the teaching table have 100% coincidence of values (measurements) for all tests and representative sets in this cluster, and 0% for measurements in other clusters. If a new vector is given, according to the number of coincidence of measurements for all tests and representative sets, we can establish the similarity of this new vector (object) to each cluster, based on the dissimilarity units.

Our algorithms for pattern recognition are of three types:

1. Using the minimum distance of the objects' weighted sum – A1, A2, A3 (respectively with p_i, q_i, r_i). When vectors' components are discrete values, their sum forms an information weight. This gives us a way for ranging the vectors, which form clusters. The "new" object belongs to the cluster with the nearest object.

2. Using the distance measure 'city-block' with weights – A4, A5, A6 (respectively with p_i, q_i, r_i). This metric is the total distance between all objects in a cluster and the new object.

3. By voting, taking into consideration the number of coincidences between the "new" object and the tests or the representative sets of a class – A7, A8.

Our clustering is a hierarchical clustering – each vector is considered to be one cluster at onset. Iteratively, when two vectors are recognized as closest to each other by the algorithms A1–A8, they will be associated. A1–A6 are convenient for variables with quantitative attributes, A7 and A8 for not fully defined and for qualitative attributes. (The value 0 is meaningful, not as in the Jaccard method.)

3. Development and Experiments. In order to carry out the experiments, we updated and adapted an existing program, implementing the approach described.

From the gathered information about e-Government policies and tools for Low Socio-economic Groups we took 12 of the interviews with experts as 12 vectors (AP, JK, HG, CH, SK, DI, SZ, HT, VM, PM, IT and/or AA) of data measurements of 16 questions – the answers supplied.

We tried various experiments in order to establish to what extent we could draw non-trivial conclusions by applying the methods described above to the data of the interviews.

I. Let us consider question 5 only – with 9 subgroups (characteristics) and rank of importance from 0 to 3. We obtain (DE means that the expert comes

from Germany, FI Finland, and BG Bulgaria):

	a	b	c	d	e	f	g	h	i		
E1	3	2	1	1	1	1	1	2	0	AP	DE
E2	3	3	2	2	3	3	3	3	2	JK	DE
E3	1	1	0	0	0	–	–	0	1	HG	DE
E4	2	2	1	1	1	1	1	1	0	CH	DE
E5	3	1	1	0	2	2	1	0	0	SK	FI
E6	3	3	2	2	1	1	1	3	0	DI	BG
E7	2	2	2	2	0	3	0	2	0	DZ	BG
E8	2	3	1	3	1	3	1	2	2	HT	BG
E9	3	3	1	2	2	3	2	3	3	VM	BG
E10	3	2	2	1	3	2	0	3	0	PM	BG
E11	2	1	0	1	–	1	–	2	–	IT	BG

In so far as E12 is identical to E6, E12 was excluded, so we keep working with 11 samples. After the initial evaluation of the table, we establish that the heaviest sample is E2 and the lightest one E3. Hence, the biggest difference is observed betweeng two of the German experts. As far as the characteristics are concerned, the calculations show that "h" is the heaviest, whilst "e", "f" and "g" are the lightest. The meaning of this is that "h" is the most powerful in differentiating the clusters, whilst "e", "f" and "g" are the weakest.

By consecutively excluding each object (interviewed expert), we try to attach it to some of the remaining by using the eight algorithms. We get as a value the class (cluster) which this object is closest to according to the respective algorithm.

	A1	A2	A3	A4	A5	A6	A7	A8
E1	7	7	7	4	4	4	4	4
E2	9	9	9	9	9	9	6	9
E3	11	11	11	11	11	11	–	5
E4	5	5	5	5	1	1	1	1
E5	4	4	4	4	4	3	1	3
E6	8	10	10	10	10	1	2	2
E7	1	1	1	11	11	11	1	4
E8	6	6	6	11	11	9	1,4	4
E9	2	2	2	2	2	2	2	2
E10	8	8	6	6	6	6	1,2,6	2,7
E11	5	5	4	1	1	4	–	4

According to the data so obtained, we can join E1 and E4, as well as E2

and E9 – they attach them to each other. Unilaterally, to the first group E11 and E7 are attached, and then E3 and E5 to them.

Another group, consisting of E6, E8, and E10, might be considered as a third cluster. We could compare these results with those to be obtained by using some of the other algorithms, as described at the beginning – number of coincidences or various types of distances (Euclidean, cosine, etc). For example the simple coincidences (the size of the intersection) between the samples are as follows:

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11
E1	–	1	0	7	4	5	3	4	2	4	3
E2		–	0	0	1	5	3	3	5	4	0
E3			–	0	3	0	1	0	0	0	2
E4				–	3	4	3	4	1	3	3
E5					–	3	1	2	3	3	1
E6						–	3	3	4	4	1
E7							–	3	1	4	2
E8								–	3	0	2
E9									–	2	0
E10										–	1
E11											–

One can see that samples E1 and E4 have the highest number of coincidences (7) and this corresponds to the results obtained trough the basic approach. On the other hand, we have an equal number of coincidences for E2, E6 and E9 – this does not correspond to the results of our approach. This is just to show that the later is more sophisticated and reflects more complex cases.

II. Considering question 6 only with the same 9 subgroups and ranks of importance from 0 to 3, we obtain

	a	b	c	d	e	f	g	h	i		
E1	3	1	0	0	2	1	2	1	–	AP	DE
E2	2	1	1	0	2	2	2	1	1	JK	DE
E3	2	2	2	2	2	2	2	2	2	HG	DE
E4	2	1	0	1	1	1	1	1	1	CH	DE
E5	3	1	0	1	2	2	2	2	1	SK	FI
E6	3	3	0	0	2	2	2	2	1	DI	BG
E7	1	1	1	2	0	2	–	2	0	DZ	BG
E8	1	0	0	0	0	1	1	2	2	HT	BG
E9	2	1	1	2	1	1	–	1	–	VM	BG
E10	3	2	2	1	3	2	0	3	0	PM	BG
E11	2	1	0	1	–	1	–	2	–	IT	BG
E12	3	3	1	1	2	2	2	3	–	AA	BG

Firstly, we note that E10 and E11 have given the same answers as to question 5. The heaviest sample is E12, the lightest E8. That means that for this question the biggest difference is between two Bulgarian experts.

As with question 5, the most substantial contribution to the differentiation of samples again gives "h". The lowest contribution comes from "i". As a whole, the number of irreducible tests and of irreducible representative sets here is half of those of question 5. Following the same procedure as for question 5, we obtain the number of the closest class for each excluded object.

We can unify E4 and E11 into one class (cluster), as well as into another one – E10 and E12.

	A1	A2	A3	A4	A5	A6	A7	A8
E1	4	4	4	2	2	2	4,8	6
E2	9	9	7	1	1	1	9	1
E3	10	10	12	7	7	7	–	11
E4	11	11	11	11	11	11	11	11
E5	6	6	6	11	11	11	6	11
E6	5	5	5	5	5	5	5	5
E7	2	2	2	11	11	9	–	8
E8	4	4	11	4	4	11	1	1
E9	11	11	7	4	4	4	4	4
E10	3	3	12	12	12	12	12	3,12
E11	4	4	4	4	4	4	4	4
E12	10	10	10	6	10	6	10	6

At first glance E5 and E6 seem to be very close, but the resemblance of E5 with E11 on one hand and of E6 and E12 on the other makes this union impossible. If we apply K-Means (see above) the following clusters will be obtained: {E10, E6, E12, E3}, {E4, E9, E11, E1, E2, E5} and {E7, E8}. Obviously the basic approach is more "severe". At least because K-Means can't treat missing information and in order to overcome this drawback arbitrarily replaces the missing information with an average value (in our case 1.5 was used as the middle of the interval [0, 3).

III. Considering question 7 only with 9 subgroups with ranks of importance from 0 to 3, we have

	a	b	c	d	e	f	g	h	i		
E1	2	2	1	0	1	2	1	2	–	AP	DE
E2	2	1	1	0	1	1	1	1	0	JK	DE
E3	2	2	2	1	1	1	1	2	2	HG	DE
E4	2	1	0	1	1	1	1	1	1	CH	DE
E5	3	2	1	2	1	1	1	3	2	SK	FI
E6	1	1	0	0	–	0	–	–	–	DI	BG
E7	1	1	0	0	–	0	–	–	–	DZ	BG
E8	1	0	0	0	0	0	0	1	0	HT	BG
E9	1	1	0	1	1	1	–	1	–	VM	BG
E10	0	0	0	1	2	2	0	2	1	PM	BG
E11	1	0	0	0	–	1	–	1	–	IT	BG
E12	2	1	0	0	0	1	0	1	–	AA	BG

Firstly, we note that E4 has given the same answers to question 6. The heaviest sample is E5, the lightest E8. The most substantial contribution to the differentiation of samples this time is given by "a", "b" and "f". The lowest contribution is for "h" and "i". Following the same procedure as for questions 5 and 6, we obtain the number of the closest class for each excluded object.

When applying this procedure, this time we notice that sometimes the representative sets and tests give different estimations. However the end results show closeness between E1 and E3, and E5 comes near to them – hence they form one cluster (class). Another group of proximity is formed by E2 and E4; E9 and E12 come near to them and all these four constitute a second cluster. A last cluster is formed by E6 and E8, as well as by E11, which comes near to them.

	A1	A2	A3	A4	A5	A6	A7	A8
E1	3	3	3	3	3	2	–	2
E2	4	4	4	4	4	12	4	12
E3	1	1	1	1	1	1	–	1
E4	2	2	2	9	9	9	9	9
E5	3	3	3	3	3	3	–	1,3
E6	8	8	8	8	8	8	8	8
E7	10	10	11	8	8	6	–	6,9,11
E8	6	6	6	6	6	6	6,11	6,11
E9	12	12	12	4	4	4	4	11
E10	9	9	12	11	11	11	–	1
E11	6	6	7	8	8	8	8	9
E12	2	2	9	–	–	2	2,4	6

IV. Considering question 8 only – with 9 subgroups with ranks of importance from 0 to 3 (without answers to E10), we obtain as clusters {E2, E3, E4, E9}, {E1, E5, E6, E11} and {E7, E8, E12}. The heaviest sample is E5, the lightest is E8, as with question 7. The most substantial contribution to the differentiation of samples now is given by "a", "d" and "h". The lowest contribution comes from "c".

4. Conclusions. On the basis of the results obtained, we can quite easily draw conclusions about the degree to what the various characteristics (in our case – the various LSGs) differentiate the classes of expert opinions. As already seen, the experts included in our experiments have the highest degree of disagreement (as questions 5 and 6 are concerned) on the role of group "h" – people in isolated or underdeveloped regions. This is independent on the individual experts' peculiarities. On the contrary – on the question No 5 about "the rank of importance as a target population from a public point of view" there is almost no difference in the views of the experts considered about the following LSGs:

e. Immigrants

f. Ethnic minorities

g. Refugees

Another possible direction for drawing meaningful conclusions is to try identifying personal/professional similarities between experts belonging to the same cluster. That means to what extent or which personal/professional particularities of the experts cause them to have the same or almost the same opinion on one or even to several questions. As already seen, experts E1 and E5 belong to the same cluster for questions 5, 6, 7, and 8. If we compare their personal characteristics, we establish that they both work on a national level in a governmental organization (a ministry), have experience between 3 and 7 years, estimate their own competence respectively as very good and excellent. However, they have different types of education.

We must admit that more general and extensive conclusions are not recommendable with this volume and level of data available. The problem was that a large part of the experts interviewed (through carefully selected) did not felt competent in all fields of interest or not on all LSGs (take as an example the Bulgarian professor of medicine of Roma origin who is unquestionably one of the best experts in Bulgaria on Roma problems/solutions, but chose not to answer to many the questions, because stated he was not competent enough on them). On the other hand, some of the experts, whilst competent on a particular matter,

did not accept to rank their answer even with the relatively "easy" ranking scale of 0 to 3 and preferred not to answer at all to such questions.

Consequently, the experiments demonstrated that in principle the mathematical methods adapted and applied could lead to meaningful and instructive conclusions, provided that more relevant data is accumulated and processed.

REFERENCES

[1] Juravlev J. I. et al. Recognition and classification problems with standard testing information, *J. of Computing and mathematical Physics* (*Zhurnal vychislitel'noy matematiki i matematicheskoj fiziki*), **20** No 5 (1980), 1294–1309 (in Russian).

[2] Eskenasi A. Evaluation of Software Quality by Means of Classification Methods. *J. of Systems and Software*, **10**, No 3 (1989), 213–216.

[3] Eskenazi A., V. Angelova. A New Method for Software Quality Evaluation. *J. of New Generation Computing Systems*, **3**, No 1 (1990), 47–53.

[4] The MathWorks Inc., MatLab Software – Statistics Toolbox, 1984–2007.

[5] Evert S. `www.collocations.de` – Association Measures, 2004.

Vesela Angelova
Communication
and Information Security
Directorate, Ministry of Interior
29, Shesti septemvri str.
1000 Sofia
e-mail: `vesela.a@gmail.com`

Avram Eskenazi
Software Engineering Department
Institute of Mathematics and Informatics
Acad. Bonchev Str., Bl. 8
1113 Sofia
e-mail: `Eskenazi@math.bas.bg`

Received September 18, 2007
Final Accepted January 17, 2008