A FUZZY CONTROL APPROACH FOR VOTE ESTIMATION

Jesús Cardeñosa, Pilar Rey

Abstract: This document presents the results of applying fuzzy control methods for the estimation of the lack of answer for the vote variable in the opinion polls carried out by the Sociological Research Centre (SRC).

Keywords: Fuzzy systems, voting systems

ACM Classification Keywords: J4 Social and Behavioral Sciences; I.2.3. Deduction and Theorem Proving

Introduction

This document presents the results of applying fuzzy control methods for the estimation of the lack of answer for the vote variable in the opinion polls carried out by the Sociological Research Centre (SRC). The first section describes the data used, emphasizing some problems which could be solved by means of classification procedures. The second section shows the designed fuzzy control system for classification, describing the procedures to generate the fuzzy rules and sets and their operation. Finally, we present the results and some conclusions and possible ways of improvement.

Type of Data

One of the missions assigned to the SRC by the law regulations is to carry out surveys and opinion polls to know the Spanish social reality; in particular, to learn about the Spaniards' vote intention for the general elections. In the opinion poll periodically carried out the first month of every trimester (known as "Opinion Barometer"), a direct question is posed about this vote intention. In addition, the survey includes other questions about the socio-demographic variables (sex, age, study level, labour situation, profession,…), and about other subjects related to vote behaviour that can contribute to improve knowledge about the reasons a person may have to vote for a specific political party.

In the surveys carried out by public institutions, the opinion polls are especially easy due to the simplification of some of the most technical phases of the survey processing. One of the reasons is the use of variables of qualitative type with a limited number of answering categories. In addition, only proportion estimates are almost exclusively used, and they originate smaller sampling errors than other kind of estimates. Another peculiarity of these surveys is the treatment given to the partial lack of answer (when the interviewed answers one or some of the questions, but not all of them): the category "*Don't know / Don't answer"* is added, and it is considered as any other one. By this procedure, the proportion estimated for the rest of the categories are skewed downwards, but the precision level demanded for the opinion surveys does not seem to be really great. In the case of the variable measuring the proportion of vote intention for every party ("*vote*", from now on), the high precision is demanded and it is precisely in this question where the partial lack of answer is usually bigger. To deal with this problem, the SRC usually make some *a posteriori* treatment (incorporation of expert's opinion, econometric modelling…). In any case, after elections are held, the media frequently points out critics and comments about the poor results obtained by these procedures regarding vote forecasting. Thus, when testing fuzzy systems, we will use "*vote"* as classification variable, with the purpose of using the classification results as an alternative procedure for the proportion estimates. The lack of answer would be replaced by classified values and the new proportions would

be calculated. Since this work is only a first approach to the use of fuzzy controllers for this type of data and we tried to build a simple system with few rules, the answers will be grouped in four categories: "*IU*", "*PP*", "*PSOE*" and "*OTHER*". (Being IU, PP, and PSOE acronyms of the main Spanish Political parties)

As input variables we will use the answer to four questions related to the vote: the assessment of the government management, the assessment of the first opposition party performance, the memory of the party voted for in the last national elections, and the ideological position (this latter resulting from asking the interviewed about his/her ideological position ranging from 1 to 10: 1 being the extreme left, and 10 the extreme right). The answer categories for the questions about the assessment of government management and the first opposition party performance are "*Very good*", "*Good*", "*Regular*", "*Bad"* and "*Very bad*". For memory of party voted for, they are also grouped in "*PP*", "*PSOE*", "*IU*" and "*OTHER*". This may lead to an excessive simplification, because besides grouping very different political parties (as in the case of vote intention variable), it also groups other categories that may show very different behaviours as "*Did not have age*", "*Does not remember*", "*Vote in white*" and "*Did not vote*". We have selected the data from the study number 2640 of the SRC Data Bank Catalogue (April 2006 Barometer), with a sample size of 2.500 interviews. The micro data corresponding to people residing in Autonomous Communities with other great parties (Catalonia, Galicia, and Basque Country) have been eliminated, and also those having partial lack of answer in some of the four input variables or in the output variable, resulting in 1.216 micro data to test the procedure.

Design of the Fuzzy Controller

An attractive point of the fuzzy control systems is the option of using simple rules that do not require special efforts for its design. As we have not experts in vote motivation, the classification rules are built by supposing that the interviewed answered the survey questions with some consistency. On the other hand, we are going to apply fuzzy rules in which each one will describe one of the possible classification categories. The antecedent part of the rules will be expressed by means of defined fuzzy sets in the answer categories sets of the four input variables, whereas the consequent part will be a crisp class label in the set of the classification categories. The general expression of these rules is:

R_i : If x_1 is A_{i1} and x_2 is A_{i2} and x_3 is A_{i3} and x_4 is A_{i4} then $y = y_i$, $i=1, 2, 3, 4$

where: x_1 = *assessment of government management*

x_2 = *assessment of first opposition party performance*

x_3 = *ideological position*

x_4 = *memory of the party voted for*

y_1 = *IU*

y_2 = *OTHER*

y_3 = *PP*

y_4 = *PSOE*

In a detailed form, the rules are:

– R_1: If the assessment of government management is negative, the assessment of first opposition party is negative, the ideological position is low and the memory of the party voted for is "*IU*", then the vote is "*IU*".

– R_2: If the assessment of government management is **negative**, the assessment of first opposition party is **negative**, the ideological position is **average** and the memory of the party voted for is "*OTHER"*, then the vote is "*OTHER*".

– R_3: If the assessment of government management is **negative**, the assessment of first opposition party is **positive**, the ideological position is **high** and the memory of the party voted for is "*PP*", then the vote is "*PP*".

– R_4: If the assessment of government management is **positive**, the assessment of first opposition party is **negative**, the ideological position is **low average** and the memory of the party voted for is "*PSOE*", then the vote is "*PSOE*".

It is important to point out that, although we have defined for "*memory of the party voted for"* the same answer categories than those for the "*vote"*, in the rules the meaning is very different, since for the "*memory of the party voted for"* variable we will define a fuzzy set for each party, whereas for the "*vote"* variable it regards to crisp class labels. After that, we move on to build the A_{jj} fuzzy sets of the antecedent part of the rules. In the first place, we built the "*Positive Assessment*" (*PA*) and "*Negative Assessment*" (*NA*) sets for the variables x_1 and x_2, which will be identical for both. As of the answer categories, their membership functions will respectively be in a natural way:

$$\mu_{PA}(x)=\begin{cases}0.00 & if\ x=very\quad bad\\ 0.25 & if\ x=bad\\ 0.50 & if\ x=regular\\ 0.75 & if\ x=good\\ 1.00 & if\ x=very\quad good\end{cases} \qquad and \quad \mu_{NA}(x)=1-\mu_{PA}(x)$$

where a set is the complementary of the other, taking for the complementary the strong standard negation (Figure 1).

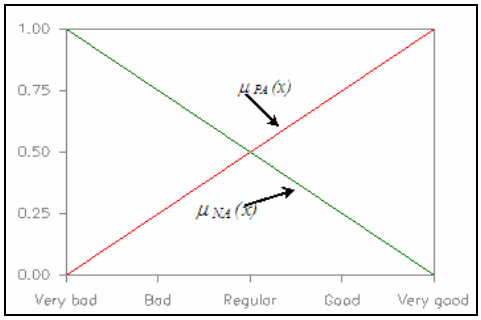

Figure 1

Now, in order to build the "*Low*" (*L*), "*Low Average*" (*LA*), "*Average*" (*A*) and "*High*" (*H*) fuzzy sets for the universal set of answers to the question about the ideological position, we will base on the average and the standard deviation for each classification category obtained from the sample. For later evaluation of the behaviour of the fuzzy classifier, we took into consideration the first 916 observations to estimate, being left the other 300 for the classification test. Table 1 shows the obtained values:

Party	Average	Standard deviation
IU	*3.00*	*1.06*
OTHER	*4.53*	*1.40*
PP	*6.67*	*1.31*
PSOE	*3.71*	*1.25*

Table 1

From this one, we built the Gaussian membership functions, in the form:

$$\mu_L(x) = e^{-\frac{(x-3,00)^2}{2.1,06^2}},\ x = 1,2,\ldots,10 \quad ,, \quad \mu_{LA}(x) = e^{-\frac{(x-3,71)^2}{2.1,25^2}},\ x = 1,2,\ldots,10$$

$$\mu_A(x) = e^{-\frac{(x-4,53)^2}{2.1,40^2}},\ x = 1,2,\ldots,10 \quad ,, \quad \mu_H(x) = e^{-\frac{(x-6,67)^2}{2.1,31^2}},\ x = 1,2,\ldots,10$$

that appears in Figure 2. They are taken as symmetrical functions around the average because that is how the studied frequencies in the micro data sample seem to behave.

Finally, we built the fuzzy sets for the results of the "*memory of the party voted for*" variable also from the information provided by the sample of the first 916 micro data, as the vote frequency (distribution by categories of the classification variable) for each group of "memory of the party voted for", as it appears in Table 2:

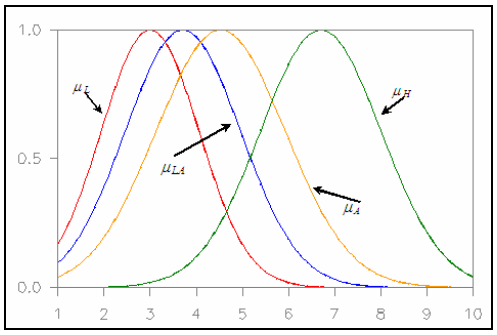

Figure 2

	Vote				
Memory of the party voted for	*IU*	*OTHER*	*PP*	*PSOE*	*Total*
IU	66.1	6.8	3.4	23.7	100.0
OTHER	2.9	37.5	23.0	36.6	100.0
PP	0.7	7.2	75.9	16.2	100.0
PSOE	1.0	7.0	6.1	85.9	100.0

Table 2

This gives rise to the membership functions in Figures 3 to 6:

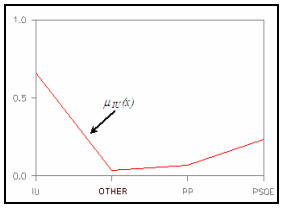

Figure 3

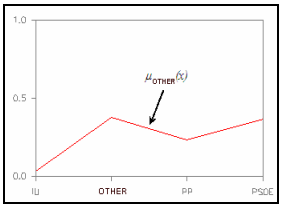

Figure 4

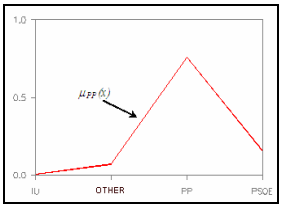

Figure 5

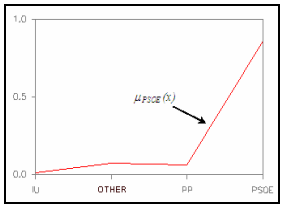

Figure 6

Once the fuzzy sets for the input variables were built, the next action was to follow the steps to design the system of fuzzy control.

Operation of the System

It will be necessary to choose the implication functions, t-norms, and so on, that allow the development of the system logout, once an input takes place. It has already been pointed out that the standard negation is used for the complementary. As implication function we chose the Mamdani implication, based on the t-norm of the minimum. For the t-norms (connective y) we used the product, and as t-conorm the standard fuzzy union (maximum operator). It has also been studied the use of the t-norm of the minimum instead of the product one: the first one showed the problem that frequently the maximum value to select the class was the same for different parties. The use of the t-norm of the product allows avoiding it, making possible the interaction among the antecedent proposals of each rule. On the other hand, we applied the Zadeh's compositional rule of inference locally to each fuzzy relation generated by the rules, coming later to combine the resulting sets in a disjunctive way.

If we suppose now that a numerical input to the system takes place, that is to say, the fact $x = x_0$ takes place or $x \in P^* = \{x_0\}$, where $x = [x_1, x_2, x_3, x_4]^T$ is the vector of input of the four variables, it will be $\mu_{P^*}(x) = \begin{cases} 1, & si\ x = x_0 \\ 0, & si\ x \neq x_0 \end{cases}$

We will then have, to make the inference, the four rules and the previous fact. In the first place, we will combine the fuzzy sets of the inputs of each rule applied to $P^* = \{x_0\}$ in a multiplicative way according to the selected t-norm, getting the activation rank of each rule as:

$$\beta_i(x_0) = \prod_{j=1}^{4} \mu_{ij}(x_{j0}),\ i = 1,2,3,4$$

where $x_0 = [x_{10}, x_{20}, x_{30}, x_{40}]^T$ and μ_{ij} are the corresponding membership functions. As the output of each rule is a crisp set (the class label of the corresponding party), it will have a membership function that will be, in fact, a characteristic function, properly:

$$\mu_{y_i}(y) = \begin{cases} 1, & si\ y = y_i \\ 0, & si\ y \neq y_i \end{cases} \quad , \quad i = 1,2,3,4$$

In order to apply Zadeh's compositional rule to each rule, it would be necessary to make:

$$\mu_{Q_i^*}(y) = \mathcal{J}\left(\beta_i(x_o), \mu_{y_i}(y)\right) = \min\left[\beta_i(x_o), \mu_{y_i}(y)\right] = \begin{cases} \beta_i(x_o), & si \quad y = y_i \\ 0, & si \quad y \neq y_i \end{cases}, \quad i = 1,2,3,4$$

where J is the Mamdani implication. Applying now the disjunctive combination of each rule outputs, the output of the classifier will be determined by the rule with the highest activation degree, that is to say,

$$y = y_{i^*}\ , i^* = \arg \max_{1 \le i \le 4} \beta_i(x_0)$$

Results

In order to test the classifier, we applied it to obtain the value of the *vote* for the 300 observations that have been left in the sample with this aim. If we compared the vote thus obtained with the real value provided by the interviewed person, we found that there was coincidence in 240 observations (80% of the data). In order to make the experiment more representative, we repeated all the steps four times, leaving 300 different observations every time for verification. Although the test number is not enough to extract significant consequences, it is observed that all the parameters taken into account stayed quite stable from a repetition to another one, which provides certain confidence in the reliability of the procedure. The results of the four iterations are shown in Table 3.

Iteration	*% of coincidence*
1	80.00
2	80.76
3	80.76
4	80.70
Average	80.56

Table 3

On the other hand, tests have also been made changing the membership functions obtained for other simpler ones, of trapezoidal and triangular type. It has been found that it is very simple to refine the parameters of those functions to significantly improve the results, but those refined parameters do not continue giving good results in other segments of sample. Although results may seem a little unexciting, if we compared them with 51%, at the most, that has been obtained with data of a similar survey, using classifiers based on Bayesian networks, it is rather positive. (These results are not totally comparable since the other survey used different input variables).

Conclusions

The experiment made here is only a first approach to obtain classifiers for the vote in opinion polls based on fuzzy controllers; therefore, instead of conclusions we are making some comments.

The final mission has been to improve the imputations of the “vote” lack of answer in the barometers of the SRC. It is necessary to point out that this experiment is rather more modest than to improve the estimations of the vote for the following elections. Also, it is necessary to remember that the interviewed people can be incongruous when they are asked about the vote and when they go to vote.

The shown results are quite encouraging, mainly considering that the design of the control system has not required of exhaustive previous analysis, but it has been obtained with a few rules based on common sense.

The system is simple and with very few rules to group the parties in only four groups, but it would necessarily become more complicated if it were tried out with all the political parties.

In the barometers there also are other questions included that can be used like input variables in the system: for example, questions about the assessment of the main leaders of the political parties, the confidence in the President of the government, etc. It is quite possible that its inclusion could allow improving results.

The tests made with other simpler functions of property indicate that it is also possible to improve the results using that route.

Bibliography

[Chen 2006] Chia-Chong Chen. Design of PSO-based Fuzzy Classification Systems. In: Tamkang Journal of Science and Engineering, Vol.9 nº1 pp 63-70. http://www2.tku.edu.tw/~tkjse/9-1/9-1-7.pdf

[Diez 2005] F.J. Diez. Introducción al razonamiento aproximado. Dpto. Inteligencia Artificial, UNED.

[Halkidi 2003] Vazirgiannis, Michalis, Halkidi, Maria, Gunopulos, Dimitrious. Uncertainty Handling and Quality Assessment in Data Mining. 2003, IX, 226 p., ISBN: 978-1-85233-655-4

[Yuan 1994] GJ Klir, B Yuan Fuzzy sets and fuzzy logic: theory and applications. Prentice-Hall, Inc. Upper Saddle River, NJ, USA1994

[Roubos 2002] Johannes A. Roubos, Magne Setnes, and Janos Abonyi. Learning fuzzy classification rules from labeled data. IN: Information Sciences. Vol 150. pp77-93. 2003

[Tanaka 1997] K. Tanaka An introduction to Fuzzy for Logic Practical Applications. Springer New York. 1997

Author's information

Jesús Cardeñosa – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: carde@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es

Pilar Rey – Banco de Datos. Centro de Investigaciones Sociológicas. C/ Montalban ,8; 28014 Madrid (Spain). e-mail: prey@cis.es; http://www.cis.es