SUPPLEMENT TO KYBERNETIKA VOLUME *28* (1992), PAGES 58 – 61

FUZZY CLUSTERING AND THE WAY OF DEFINITION OPERATION

EMIR VELEDAR

In this article, by grouping and comparing cotton producing countries according to their level of development of cotton production, we will show that the use of specific operations that are normally used in fuzzy set theory should be based on particular information specific to each individual task, as opposed to the usual a priori use of such operations as is usually done.

The relationship between the problems of fuzzy set theory and automatic classification (cluster analysis) has been stressed by many researchers, who have emphasized the application of fuzzy set theory techniques to cluster analysis. Their main assumption has been that the clusters are fuzzy sets (see Backer [1]). The appeal of such an approach is that it avoids the strict classification of points into sets, but instead considers the degree of membership necessary for certain criteria. In doing this a measure of fuzziness is applied not only to determine the membership of certain points, but also to classify classes as a whole (cf. [1]).

At the core of all these operations are well-known fuzzy set operations – minimum and product, but the justification for the use of such operations for cluster analysis remains unclear. The aim of this article is to discuss how appropriate this assumption is. While we do not plan to restrict our discussion unnecessarily, the primary focus of this article is two-dimensional analysis. When solving automatic classification problems, distance between each pair of points along a coordinate is always known. Given these distances, we then calculate multidimensional measures of distance. In fuzzy set theory measures of distance are defined by the formula:

$$\mu(S_1, S_2) = \min(\mu(S_1), \mu(S_2)) \tag{1}$$

$$\mu(S_1, S_2) = \mu(S_1) \star \mu(S_2) \tag{2}$$

where S_1 and S_2 are objects or their characteristics. These formulas are used to calculate one-dimensional object estimates based on several fundamental characteristics which are essential at that capacity. However the selection of such an estimate is not motivated by anything in particular. The basis of this analysis is discussed in Kovalerchuk [3]. The justifications for this operation, when referring to cluster analysis, consist of treating the probability measures as if they are defined by Kolmogorov axioms as a membership function. These membership functions can be defined through relative frequencies, but they can also be considered subjective probabilities, or can be derived on the basis of

subjective estimates. Also it is assumed that the probability measure is not completely known. In other words, there are elements of the probability space in which the significance of probabilities is not known. It is also assumed that there is certain information that can be used additionally to define the essential significance of probabilities. This additional information can be obtained through a dialogue with decision makers (cf. [4]).

We will take as an example the grouping and comparing of cotton producing countries according to the level of development of their cotton production.

Initially, we will use cotton production and consumption statistics from the FAO (cf. [5]) for 1985 and population statistics from OUN (cf. [6]) for each country C. Therefore we will consider two characteristics:

$$\begin{aligned} A(C) &- \text{cotton production} \\ B(C) &- \text{cotton consumption.} \end{aligned}$$

In order to simplify our analysis, we will define them as A and B. Then in order to apply our method, it is essential to obtain general characteristics and to give them probability status. If it is possible to consider these characteristics as probabilities, the solution to a particular problem would involve calculating general characteristics (common probabilities) by multiplying them and comparing countries based on the size of the results. Countries whose products do not significantly differ would be classified into one cluster, a relatively simple task of one-dimensional cluster analysis. In this process, if a problem appears, one can apply certain mathematical methods (see Veledar [7]). One appropriate method is that suggested by Zhuravel and Ionin [9], where one maximizes intergroup distance criteria and minimizes intragroup dispersion, a Hungarian solution used in dynamic programming. We will now discuss how the values we are interested in can be transformed into probabilities.

In order to do this we will consider a situation where "country C does not produce less cotton than A", A being the actual production of cotton in the year under consideration. Let M be the maximum cotton production per head in all cotton producing countries. We introduce A/M. It will be considered the measure of production per head for each country. The value of A/M is treated as the probability $P(x \geq A)$, or in other words the cotton production probability is not lower than the actual value of A. The probability space is formed by the events:

$$\begin{aligned} S_1(C) &= \text{"production per head in country } C \text{ is not less than } A\text{"} \\ S_2(C) &= \text{"production per head in country } C \text{ is less than } A\text{"}. \end{aligned}$$

More complex probability spaces require more extensive calculations including the solution of linear equations. This is why we are only considering spaces with two elements. We define the probabilities:

$$P(S_1(C)) = A(C)/M \tag{3}$$

$$P(\bar{S}_2(C)) = 1 - A(C)/M. \tag{4}$$

We have introduced a probability space with events:

$S_2(C)$ – "cotton production per head in country C is not lower than B"
$\bar{S}_2(C)$ – "cotton production per head in country C is lower than B",

with probabilities:

$$P(S_2(C)) = B(C)/H \tag{5}$$
$$P(\bar{S}_2(C)) = 1 - B(C)/H \tag{6}$$

H is the maximum consumption per head of cotton for all countries under consideration and $B(C)$ is cotton production in country C. The above probability measures, as was stressed earlier, represent conditional probabilities, the condition being that cotton production is not lower than A. This can be expressed by statements such as:

$P(S_3(C))$ "the measure of cotton consumption per head in country C, with the condition that cotton production be not lower than A, and is equal to A/M."

In mathematical terms:

$$P(S_3(C)) = P(x \geq B \,|\, y \geq A) = A/M \tag{7}$$

where x is the consumption and y is the production.

Thus we have all necessary probabilities

$$P(x \geq B \,|\, y \geq A) \star P(y \geq A) = P(x \geq B) \wedge (y \geq A).$$

The countries are grouped in the following way:

The first group consists of the most developed countries: Turkey, USSR, Israel, Egypt, USA. In each of these countries, the common probability is not less than 0,3.
Pakistan forms a group in and of itself (0,25).
The next group is: Australia, Brazil, China, Peru (0,07–0,1).
The last group is: Sudan, Columbia, Mexico and India (0,02–0,04).

Conclusion. The method we have chosen, if we consider only its formal characteristics, resembles methods based on "fuzzy" conjunctions – the product, an operation which is well known in fuzzy set theory. However, there is a principal methodological difference. In fuzzy set theory, minimums and products are given as postulates. This is justified by providing examples of systems that operate functionally, based on such operations. Mamdani's and Assilian's article [10] is an example of this and it is frequently used as a reference. Zadeh [11], in particular, uses Mamdani's argument in answer to Traybus's criticisms.

Our example makes it obvious that one can apply this operation from a different perspective, by treating the membership function as a probability measure that is not completely defined. If you have a concrete situation you can opt for additional measures

given additional information. In our example an estimate imposed externally has been shown to be an accurate estimate, when the probability measure is introduced in a particular way. What is really known is not $P(x \leq B)$ but $P(x \leq B \,|\, y \leq A)$ and that makes the definition of $P((x \leq B) \wedge (y \leq A))$ accurate although, in general cases:

$$P((x \leq B) \wedge (y \leq A)) \leq \min(P(x \leq B),\ P(y \leq A)).$$

This problem can also be formulated for three or more parameters. It is possible to take these measures derived from such methods as rough estimates, because the original problem, to compare countries based on their consumption and production, has been replaced by a problem in which it is assumed that production and consumption are not lower than current levels. In order to obtain accurate measures iterative procedures can be used to precisely define conditional probabilities as it was done in Bellman and Zadeh [2].

Our main conclusion is that in using one or another operation over fuzzy sets in cluster analysis, in order to group objects on the basis of multidimensional data, decisions should be based on specific information from a concrete example, and not a priori.

REFERENCES

[1] E. Backer: On fuzzy optimization method of decomposition in cluster analysis. Reliability and Maintainability Symposium, 1980.

[2] R. Bellman and L. A. Zadeh: Decision making in fuzzy environment. Manag. Sci. *17* (1970), 141–164.

[3] B. Kovalertchuk: Veroyatnostnaya interpretaciya peresetcheniya i obyedineniye razmytykh mnozhestv. Tashkent 1986.

[4] I. B. Kurdjumov, M. V. Mosolova and V. B. Nazajkinskii: Zadatcha mnohocelevoy optimizatsii s nechotkimi usloviyami. Izvestiya AN SSSR, Tekhn. kibernetika 1979, 6, 3–8.

[5] Cotton outlook, No. 3 – 18.

[6] World statistics in brief, U. N., N. Y. 1985, 108 pages.

[7] E. Veledar E: Metodi i algoritmi fuzzy klaster analize. Faculty of Economics, Mostar 1990.

[8] E. Veledar and B. J. Kovalerchuk: Vjerovatnosna interpretacija presjeka i unije Fuzzy skupova kao osnova za klasifikaciju. II majski skup sekcije za klasifikacije i nomenklature SSDJ, Mostar 1988.

[9] N. M. Zhuravel, B. G. Ionin and N. N. Ionina: Ispolzovanie procedur optimizacii pri klassifikacii objektov i faktorov. Nauka, Novosibirsk 1978, pp. 147–161.

[10] E. H. Mamdani and S. Assilian: An experiment in linguistic synthesis with a fuzzy logic controller. Internat. J. Man–Machine Stud. 7 (1975), 1–13.

[11] L. A. Zadeh: Fuzzy sets or probability theory. Proc. IEEE *68* (1980), 421.

Dr. Emir Veledar, Faculty of Economics, 88000 Mostar. Yugoslavia.