ON MEASURES OF CLUSTER VALIDITY

Jiří Šustal

Technical University of Brno
Faculty of Electrical Eng., Dept. of Mathematics,
Hilleho 6, 60200 Brno, Czechoslovakia

In this short contribution I shall give an overview of some recent results concerning measures of cluster validity.

A fuzzy partition algorithm is aimed to obtain a fuzzy partition (or several fuzzy partitions), given by numbers u_{ij} where u_{ij} evaluates how much an element x_j from the set $X = \{x_1, \dots, x_n\}$ belongs to the cluster i where $1 \leq i \leq c$, c being the number of clusters. We can ask, how much the fuzzy clusters are hard. If they are hard enough, then the uncertainty connected with the classification of an element x_j into the most likely class (determined by the highest u_{ij}) will be relatively small. In an opposite case it will be high. This uncertainty about an element x_j can be called local. If all elements x_j from X are taken into account, then we get the so called global uncertainty. As pointed out in Bezdek's book /1/ p. 98, even if striving for a genuinely fuzzy partition, we cannot completely cancel hard partitions, because in the end by this concept the suitability of a fuzzy partition is judged. Hence if we could somehow formalize the concept of uncertainty, it would help us to compare fuzzy partitions and we would get a measure of cluster validity. This is the line we are pursuing here.

Let $\bar{u}_j$ denote the set of elements $u_{1j}, \dots, u_{cj}$ and let U denote the set of all elements u_{ij}. Suppose that an appropriate measure $H(\bar{u}_j, c)$ of the local uncertainty about the classification of an element x_j is already at our dispésal. Then how can the measure $G(U,c)$ of the global uncertainty be constructed? Several possibilities lend themselves, e.g. (for a nonnegative $H(\bar{u}_j, c)$)

$$G(U,c) = \sum_j H(\bar{u}_j, c) \qquad (1)$$

$$G(U,c) = \sum_j w_j H(\bar{u}_j,c), \quad w_j \geq 0 \tag{2}$$

$$G(U,c) = \sup_j H(\bar{u}_j,c). \tag{3}$$

Formula (1) posesses the additivity property, i.e., for $U=U_1 \cup U_2$, $U_1 \cap U_2=\emptyset$, we have $G(U,c)=G(U_1,c)+G(U_2,c)$. Hence $U_1 \subset U \Rightarrow G(U_1,c) \leq G(U,c)$.

Now let us consider two well-known measures of cluster validity. We can show that they bear features of the above outlined procedure of constructing global measures from the local ones. These measures are

the partition coefficient:

$$F(U,c)= \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^2,$$

and the partition entropy:

$$E_1(U,c)=- \frac{1}{n} \sum_j \sum_i u_{ij}\log_a(u_{ij}), \quad a>1.$$

For our purpose it will be better to introduce slightly modified measures

$$F_1(U,c)=1-F(U,c)=1- \frac{1}{n} \sum_j \sum_i (u_{ij})^2,$$

$$F_2(U,c)=nF_1(U,c)=n- \sum_j \sum_i (u_{ij})^2,$$

$$E_2(U,c)=nE_1(U,c)=- \sum_j \sum_i u_{ij}\log_a(u_{ij}).$$

The modofications F_1,F_2 can be better understood if we remember that $F(U,c) \leq 1$ where $F(U,c)=1$ only in the case of zero uncertainty, i.e., if all $\bar{u}_j$'s are of the form $(0,\ldots,0,1,0,\ldots,0)$.

The underlying local measures for the above formalas can be easily guessed.

$$F(\bar{u}_j,c)=1- \sum_i (u_{ij})^2 = \sum_i u_{ij}(1-u_{ij}) \tag{4}$$

which is in fact Vajda's entropy /2/.

$$E(\bar{u}_j,c) =- \sum_i u_{ij}\log_a(u_{ij}) \tag{5}$$

is of course the Shannon entropy.

We can see that E_1,E_2,F_1,F_2 result from the local measures (4),(5) using formulas (2),(1),(2),(1) respectively,

where the weights for the formula (2) are chosen as $w_j = \frac{1}{n}$.

So far the underlying local measure has been guessed rather casually departing from the known formulas for the global measure. A more systematic procedure would proceed reversely, first to state properties for the local measure, then to construct appropriate local formulas, and only afterwards to give global measures. Proceeding along these lines I have considered about 10 properties, which can be found in /3/. One example of a local measure constructed by this procedure is

$$H(\bar{u}_j,c)=1-\sum_{2}^{c}\lambda_i(u_{1j}-u_{ij})-u_{1j}\sum_{c+1}^{\infty}\lambda_i, \qquad (6)$$

where $\lambda_i = \frac{1}{(i-1)i}$ (or more generally $\sum_2 \lambda_i=1$, $\lambda_i > 0$) and where $\{u_{ij}\}$, $i=1,2,\ldots,c$, form a nonincreasing sequence, i.e., $u_{1j} \geq u_{2j} \geq \ldots \geq u_{cj}$ otherwise a rearrangement would be necessary.

Having a global measure $G(U,c)$ we can not only compare 2 fuzzy partitions but we can also decide about the suitability of a single fuzzy partition. To this end we need some benchmark b_G to be able to say that the partition is good enough if $G(U,c) < b_G$, or that it is too uncertain if $G(U,c) \geq b_G$.

Again as a basis for our considerations it is prferable to start with the local uncertainty. For the above formulas (4),(5),(6), one can take as the decision level the value $b=H((\frac{2}{3}, \frac{1}{3}),2) = H((\frac{2}{3},\frac{1}{3},0,\ldots,0)c)$. This choice of $\bar{u}_j$ is rather accidental and, depending on the situation, also other $\bar{u}_j$ are feasible. For the global measure we then get following decision levels

for the formula (1): $b_G=n.b$,

for the formula (2): $b_G=\sum_j w_j.b$ ($=b$ if $\sum w_j=1$),

for the formula (3): $b_G=b$.

REFERENCES

/1/ Bezdek, J.C., Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York, 1981.

/2/ Vajda I., A contribution to the informational analysis of pattern, in: Methodologies of pattern recognition, ed. by S. Watanabe, Academic Press, New York, 1969.

/3/ Šustal J., On the uncertainty of fuzzy classifications, in: Approximate reasoning in decision analysis, ed. by Gupta M., Sanchez E., North Holland, Amsterdam, 1982.