ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 196, 2006

*Jerzy Korzeniewski**

PROPOSAL OF NEW CLUSTER ANALYSIS ALGORITHM

Abstract. One of well-known groups of cluster analysis methods is the group of methods based on density estimation. In the paper we propose a new method of defining clusters which consists of two steps. In the first step we find local maxima of the joint distribution thus establishing clusters centres. In the second step we assign observations to one of existing clusters centres. The number of clusters is assumed to be known. In both steps we use similar technique based on the kernel density estimator with the Epanechnikov kernel. The performance of the method is analyzed in an example of application to the Gordon (1999) data. In the analysis the Rousseeuw indices are used to assess clusters cohesion as well as and some comparisons with other methods of defining clusters are presented. The results look promising.

Key words: cluster analysis, density estimation, kernel estimation, Epanechnikov kernel.

1. INTRODUCTION

Let us consider arbitrary set of n points from d-dimensional Euclidean space. Multidimensional kernel estimate based on kernel K and window size h calculated at point x is given by the formula

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{1}$$

The optimal kernel in the sense of minimal mean square error is the Epanechnikov kernel given by the formula

$$K_E(x) = \begin{cases} 0.5\, c^{d}_{-1}(d+2)(1 - x^T x) & \text{if} \quad x^T x < 1 \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

* Ph.D., Chair of Statistical Methods, University of Łódź.

The gradient of the density estimator i.e.

$$\nabla\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n}\nabla K\left(\frac{x-x_i}{h}\right) \tag{3}$$

will be equal to

$$\nabla\hat{f}(x) = \frac{1}{nh^d c_d}\,\frac{d+2}{h^2}\sum_{x_i \in S_h^{(x)}}[x-x_i] = \frac{n_x}{nh^d c_d}\,\frac{d+2}{h^2}(1/n_x \sum_{x_i \in S_h^{(x)}}[x-x_i]) \tag{4}$$

The quantity in the brackets i.e.

$$M_h(x) = \{1/n_x \sum_{x_i \in S_h^{(x)}}[x-x_i]) = 1/n_x \sum_{x_i \in S_h^{(x)}} x_i - x \tag{5}$$

is called the sample mean shift. One may prove (Comaniciu, Meer 2000) that the sequence of consecutive centres of sample/window is convergent to the local maximum of density function. The convergence is quite fast therefore we will use this kernel estimator based on the Epanechnikov kernel in both steps on the proposed method.

In the first step we will use the sample mean shift method to find the centres of the predetermined number of clusters. In the second step we may also use the sample mean shift method to determine the direction in which the window "moves" and, in this way we will find cluster to which each point should be assigned. The remaining part of the paper is devoted to the description of the proposed algorithm and some clustering assessment methods are described in a more detailed way. In the third part of the paper the performance of the method is assessed in an example of application to the Gordon (1999) data.

2. ALGORITHM DESCRIPTION

The first step is an iterative one. In the first iteration we draw randomly k points, where k is the number of clusters that has to be assumed. We find the points of convergence for each of the k points in the sequences of consecutive shifts of windows of size h (the same size for every point for all sequence items). If the number of different points of convergence is equal to k and any two different limit points meet the condition

$$d(x_i, x_j) > h \quad \text{for} \quad i \neq j, \tag{6}$$

where x_i, x_j are limit points for i, $j = 1, 2, \dots k$, we accept these points as clusters centres (to be modified in next steps). If there are less than k points of convergence (i.e. some sequences converge to the same points) or if condition (6) is not met, we forget about the drawn k points and we draw next k points. Once that we have established some k cluster centres we modify them iteratively i.e. at each iterative step we draw randomly k points and if we arrive at k different points of convergence satisfying condition (6) we take weighted sums of these points and existing clusters centres i.e. centre c_i at the j-th iteration is determined by the formula

$$c_i = ((j-1)c_{i-1} + x_{i,j})/j, \tag{7}$$

where $x_{i,j}$ is the one of the limit points of k points at the j-th iteration that is closest to the centre c_i. Centres modification is performed in natural succession i.e. we start with $i = 1$ then $i = 2$ and so on. While "adding" new limit points to existing cluster centres we do not trouble to insure any kind of optimization i.e. to add limit points to closest centres. Such optimization would require defining the succession or importance of centres and, thus, another parameter. As it turns out such optimization is not necessary because very seldom it takes place that limit points are assigned to "wrong" centres.

In the algorithm's first step described above, the choice of parameter h is crucial to the proper performance of the whole algorithm. Some researchers call parameter h "cosmic" as there is no indication of its value that would be suitable for clustering. We applied the following procedure of determining the value of parameter h. All coordinates of h are determined in the same way on the basis of the projections of all observations on a given coordinate. Let $y_1, \dots, y_n$ be the values of all observations projected on a fixed coordinate. Let r be the smallest positive Euclidean distance between two values out of y_i i.e.

$$r = \min_{i \neq j} |y_i - y_j| \quad \text{and} \quad r > 0. \tag{8}$$

We will use the well known statistical formulae for the number m of classes in order properly present statistical population consisting of n observations

$$m \leqslant 30 \quad m \leqslant 5 \ln n \quad m = \sqrt{n} \quad m = 1 + 3.322 \log n. \tag{9}$$

We will accept m to be the greatest of these three recommendations. We calculate the width of each class (equal for all classes) by dividing the greatest distance between two values out of $y_1, ..., y_n$ by m. The value of parameter h will be equal to half of the median of the medians of distances between each two consecutive local maxima of the classes numbers of objects. The set of medians is constructed in the following way. First median of the set is calculated for the case in which the beginning of the first class is equal to the smallest value of y_i, the second median of the set is calculated for the case in which all classes are shifted to the right by r, the third median of this set is calculated for the case in which all classes are shifted to the right by $2r$, and so on, until the beginnings of the classes exceed the ends of the classes from the first case. The idea behind defining parameter h as equal to half of the average distance between local maxima of the projection distribution density function is that this value is the perfect value for the observations lying in the neighbourhood of local minima of the density function, to decide in which direction (to which local maximum) they should be clustered by the density kernel estimate based on the window of size h. By the distance between two consecutive local maxima we understand the distance between the centres of two consecutive classes strictly more numerous than each of their two neighbouring classes. The value of parameter h determined in this way may fail to give proper clustering only if in some data regions there are many local maxima located closely to one another and in some other data regions there is a smaller number of local maxima located further from one another. In such cases the value of the parameter should be determined locally.

The second step of the algorithm is focused on assigning every observation to one of the cluster centres determined in the first stage. The simplest way is to assign every observation to the cluster represented by the closest cluster centre. This way does not work properly which one can check on almost any data set to be found in literature. The reasons for this behavior are obvious, observation should be assigned to the clusters the distance from which, or the distance from the "meaningful" part of which is smallest. The distance from the clusters centres is not crucial. Another simple way is to assign observations sequentially i.e. in each step we assign the observation which has the smallest distance from one of the clusters (i.e. the smallest distance to the closest member of each of the clusters created up to the current step) to this cluster. This way does not work properly as one can check easily in a number of examples. The reason this time is the fact that sequential assigning of observations may cause "approaching" of clusters to observation not assigned yet independently of the distance between an observation and its closest neighbors i.e. an observation may be assigned to an erroneous cluster because the observations closest neighbors have not been assigned yet to any of the clusters created up to the current step.

In the second stages of the algorithm we propose the following procedure which seems most natural and gives good results. Every observation is assigned to the cluster represented by the centre which is closest to the limit of the mean shift procedure for this observation. Window size (different at each step) of this procedure is equal (in each dimension) to the Euclidean distance between the point generated in the current step of the procedure and the closest of all the clusters centres.

3. ALGORITHM APPLICATION AND ASSESSMENT

Let us apply our algorithm to the clustering of the Gordon (1999) data. These data set consists of 300 observations generated from three different two dimensional normal distributions (100 observations from each). The centres of these distributions are located at the midpoints of the sides of equilateral triangle whose sides are of length 10. For each of the three distributions, the major axis of its variance-covariance matrix lies along the side of the triangle and has length 4, with the minor axis having length 1. There is a fair amount of overlap between each pair of the three distributions (cf. Figure 1).

As the number of observations n is equal to 300 then the implementation of the first step of our algorithm was performed for $m = 20$ classes (according to (9) m should range from 18 to 28, but all these values result in very similar values of h). The first coordinate (corresponding to the horizontal axis) of the window size h was equal to about 1/13 of the sample width (greatest observation minus smallest) on the horizontal axis. The second coordinate of h was equal to about 1/12 of the sample width on the vertical axis. Using this window size we arrived at the clusters centres depicted as three big black dots after a small number of iterations – clusters centres had stabilized after not more than 12 iterations. The results of applying the second step of the algorithm are shown in Figure 1. In the same figure the clusters find by Gordon are also presented. Gordon used the following method. In the first step a subset of 75 data items was selected from dense regions of the plane by sequentially identifying objects with minimum average distance to their fifth nearest neighbor (amongst objects that had not yet been selected). This step gave three, as the author puts it, "visually-evident" partition classes. In the second step the sample variance-covariance matrices of these three classes were evaluated and all 300 objects were assigned to the class whose Mahalanobis distance to them was smallest.

We assessed the quality of the two clustering methods by means of cluster cohesion indices proposed by Rousseeuw (see Gordon 1999). For each object i for $i = 1, \ldots, 300$ we calculated index $s(i)$ following the formula

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}, \text{ where } a(i) = \sum_{j \in \{C_r^{-i}\}} \frac{d_i}{n_r - 1} \text{ and } b(i) = \min_{s \neq r}\left\{\sum_{j \in \{C_s\}} \frac{d_{ij}}{n_s}\right\}, \tag{10}$$

where for d_{ij} we used the Euclidean distance. Positive value of index $s(i)$ suggest that object i belongs to the proper cluster while negative value suggests something contrary.

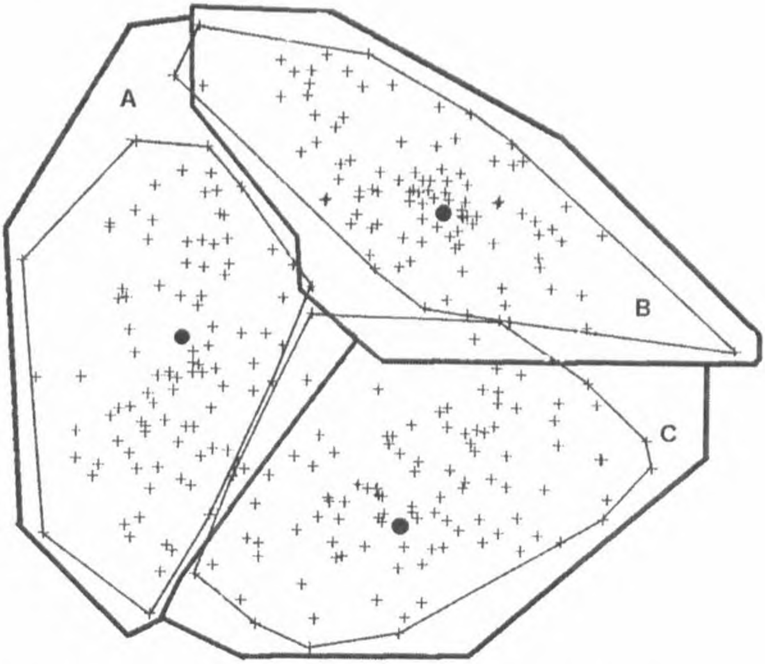

Fig. 1. Example of clustering two dimensional Gordon (1999) data. Crosses represent data, thin lines represent cluster boundaries found with the Gordon method, thick lines are boundaries of clusters established with the new method

Gordon clustering gave two negative indices, about –0.3 in value, and one fractionally negative –0.008, while our algorithm gave one slightly negative value –.01.

The overall comparative assessment of both clustering methods points to the fact that the Gordon method is more parametrized because the number of 75 observations was chosen arbitrarily, the phrase "visually-evident" may in some cases be also very questionable and the fifth closest neighbor is also an arbitrary choice. In our opinion it is safer and more robust to put some more attention to the proper choice of parameters (or exactly one parameter as it is in the case of the Epanechnikov kernel) for methods based on density function estimation to derive methods giving the same or better results.

REFERENCES

Gordon A. D. (1999), *Classification*, Chapman and Hall, Boca Raton–London.

Comaniciu D., Meer P. (2000), "Mean Shift Analysis and Applications", *IEEE Transactions Pattern Analysis Machine Intelligence*, **24**(5), 603–619.

Jerzy Korzeniewski

PROPOZYCJA NOWEGO ALGORYTMU DO ANALIZY SKUPIEŃ

(Streszczenie)

Jedną z dobrze znanych grup metod analizy skupień są metody oparte na szacowaniu gęstości. W artykule zaproponowana jest nowa metoda wyszukiwania skupień, która składa się z dwóch kroków. W pierwszym kroku znajdujemy maksima lokalne rozkładu łącznego, które przyjmujemy jako centra skupień. W drugim kroku każda obserwacja przyłączana jest do jednego z centrów. Zakładamy z góry liczbę skupień. W obydwu krokach używamy tej samej techniki opartej na estymatorze jądrowym funkcji gęstości z jądrem Epanecznikowa. Działanie metody jest przeanalizowane na przykładzie danych Gordona (1999). W analizie wykorzystano indeksy Rousseeuwa spoistości skupień, jak również przedstawiono porównanie z innymi metodami analizowania skupień. Wyniki wyglądają obiecująco.