ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 225, 2009

*Jerzy Korzeniewski**

A PROPOSAL OF NEW CLASSIFICATION ALGORITHM

Abstract

In the paper a new method of classifying points to a predetermined number of classes is presented. The method is based on the use of the sample/window mean shift technique to obtain a synthetic insight into the data set structure. The method's performance is tested on Euclidean space data sets generated by the Milligan's CLUSTGEN programme through comparison with the grouping obtained by the *k*-means method. The criterion applied are the Rousseeuw's silhouette indices are used as a criterion for camparison.

Key words: classification algorithm, mean shift method, silhouette indices.

1. Introduction

The new algorithm is based on the sample mean shift method used to estimate the local maxima of the density function of a random vector. A detailed description of this technique may be found in Comaniciu and Meer (2000). Here, we only indicate the idea behind this technique. Let $\{x_i\}_{i=1\dots n}$, be a set of *n* points from *d*-dimensional Euclidean space. The quantity

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

where $S_h(x)$ denotes the sphere of radius *h* and centre *x*, is called the window/sample mean shift. If we apply mean shift to a given window once, we get a new window centre to which (or rather to points contained in the new

[*] Ph.D., Department of Statistical Methods, University of Łódź.

centre window) we may apply mean shift once again and so on. The sequence of mean shifts constructed in such a way always moves the window in the direction of the greatest increase in density. Therefore, if we keep on moving the sample by the vector given by formula (1) we will get convergence towards the centre of the local density maximum (see. Comaniciu, Meer, 2000). By the limiting point of a given starting point we will understand the centre of the last window in the sequence of the mean shift procedures. The location of the maximum found by the mean shift sequence depends on the size h of the window. The smaller the value of h the more local is the character of the maximum; the greater the value of h the more global is the maximum. In particular, if the window size h is greater than the greatest distance between any two data points, every data point will be shifted towards the same limiting point.

We will apply the mean shift technique to every point of a data set, and in this way will get a smaller number of the limiting points which represent the whole data set as some data points have the same limiting points. The criterion deciding about assigning limiting points (and at the same time the original set points) to classes will basically be the weight of the limiting points i. e. the number of points which a given limiting point represents. First, however, we have to know the window size h for which the mean shift technique will be performed. This window size will be found by means of the horizontal phase method similar to the one used in earlier publications to determine the number of clusters (see Korzeniewski 2005). To explain the idea of the search for a sensible window size h let us imagine a two dimensional data set consisting of three identical, equally spaced, unimodal clusters. The cluster centres are 100 units away from one another. Every point will be shifted in the mean shift procedure (for $h < 100$) to the very centre of its cluster because the cluster density increases with getting closer to cluster's centre (clusters are unimodal). Therefore, if we draw 3 points and consider the condition of at least one of the distances between two limiting points being smaller than the window size of the mean shift procedure, we may observe that the probability of meeting this condition should remain constant no matter if the window size h is equal to 20, 30 or 70 units. The reason for this is that the probability of meeting the condition is equal to the probability of drawing 3 points which belong to a smaller number of classes than 3 i.e. to one cluster or to two different clusters. For such window sizes every of the 3 drawn points will be shifted to the centre of its cluster, hence, one of the distances between the limiting points will be eqaul to 0 (because at least 2 points are from the same cluster) thus meeting the condition. If we drew 3 points from 3 different clusters the condition, obviously, would not be met. Now, that the horizontal phase (constant probability of meeting the condition) is established, it seems natural to adopt the window size that lies in the middle of the horizontal phase as a good transition from data points to limiting points parameter. The next problem to be solved is to find a way of clustering the limiting points to form well defined classes. We propose to form classes (the number of which is given) on the basis of the limiting points with greatest weight or on the basis of the pooled two limiting points with greatest weight. In the latter case the two pooled points have to be mutual closest neighbours – this trial turned out to be successful for small and moderate numbers of classes (≤ 10).

2. Algorithm formulation

We will divide the algorithm into two stages. The first stage will result in finding the horizontal phase and, subsequently, in representing the whole data set by, usually much smaller number of the limiting points. The role of the second stage will be to cluster the limiting points to arrive at the final division into classes. Let us assume, therefore, that we have a data set of *n* points from *d*-dimensional Euclidean space and that the points form *k* well defined classes.

Stage One

Step 1. We find the median of the pairwise distances from 500 pairs of points.

Step 2. We draw without replacement *k* data points and for each point we find the corresponding limiting point in the mean shift procedure for a fixed window size *h*.

Step 3. We check if among all pairs of limiting points there exists at least one pair of points with the distance smaller than *h*.

Step 4. We repeat step 2 and step 3 2000 times in order to find the probability of meeting the condition from step 3.

Step 5. We repeat steps 2, 3 and 4 for all window sizes *h* from interval (0, max. distance) with *h* increasing discreetly by small increments e.g. 1/100 of the median. We get the dependance of the probability of meeting the condition from step 3 on the window size *h*.

Step 6. We find the horizontal phase of the curve representing the dependance i.e. the segment of the curve of the length equal to 1/10 of the median (i.e. 10 consecutive increments of window size *h*) for which the chi-suqare statistic (to measure the uniformity of fractions and in this way the horizontality) has the smallest value.

Once we have found the horizontal phase we forget about he original data set, and from now on it is represented only by the set of the limiting points of the mean shift procedure for window size *h* lying in the middle of the horizontal phase. Every limiting point has weight i.e. the number of original data points having this point as their limiting point.

Stage Two

Step 1. The limiting points are clustered sequentially to form k classes in the following way: each sequential class is defined either by the heaviest limiting point or two jointly heaviest limiting points (whichever variant gives heavier class). In the latter case i.e. clustering two jointly heaviest limiting points, we impose a side condition that both points have to be mutual closest neighbours.

Step 2. When k classes have been formed (by repeating stage one k times) and some limiting points are still left we incorporate these points into one of the k classes according to the following sequential rule: starting from the heaviest limiting point we incorporate it into the class which contains the closest limiting point.

3. Performance analysis

In order to assess the performance of the algorithm proposed we generated 240 data sets using the Milligan's CLUSTGEN programme, each set made up of 100 points. The sets were generated in 12 cases i.e. 20 sets with 2, 3, 4 and 5 classes in the Euclidean spaces of dimensions 4, 6 and 8. In each case the number of classes k was equal to the true number of classes. We compared the results with the classification obtained by the k-means method with k randomly chosen starting points. The comparison was based on the Rousseuw's silhouette index which for the i-th element is given by the formula

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{2}$$

where $a(i)$ is the average distance between the i-th element and all other points in its class $b(i)$ is the average distance to points in the nearest class. The negative value of $s(i)$ suggests that the i-th point should rather belong to some other class. The actual comparison criterion was the percentage of data set points with negative value of $s(i)$.

The results showed that there was no significance difference neither with respect to the number of classes nor to the dimension of the space. The arithmetic mean of the percentage of points with negative indices was equal to 5.8% while for the k-means it was equal to 16.2%. The proposed algorithm may be assessed as interesting, though, one has to remember that there are probably better methods than k-means clustering. An interesting characteristic of the algorithm is that it is nonparametric and it can be applied to real life data sets – the sets generated by CLUSTGEN are mixtures of normal distributions. This is the path for further author's investigations.

References

Comaniciu D., Meer P. (2000), *Mean shift analysis and applications. Pattern analysis and application*

Gordon A. D. (1999), *Classification*, Chapman & Hall.

Korzeniewski J. (2005), *Comparative assessment of some chosen methods of determining the number of clusters in a data set*, „Acta Universitatis Lodziensis", Folia Oeconomica, (to appear).

Jerzy Korzeniewski

Propozycja nowego algorytmu klasyfikacyjnego

W artykule przedstawiona jest nowa metoda klasyfikowania punktów zbioru danych do klas, których liczba jest zadana. Metoda oparta jest na wykorzystaniu techniki średniego przesunięcia okna/próby do uzyskania syntetycznego wglądu w strukturę zbioru danych. Działanie metody jest sprawdzone na zbiorach danych z przestrzeni euklidesowych wygenerowanych przy pomocy programu CLUSTGEN poprzez porównanie wyników z grupowaniem uzyskanym metodą *k*-średnich. Kryterium porównawczym są indeksy sylwetkowe Rousseeuwa.