A Complete Gradient Clustering Algorithm

Piotr Kulczycki[1,2] and Małgorzata Charytanowicz[1,3]

[1] Systems Research Institute, Polish Academy of Sciences,
Center for Stochastic Data Analysis Methods,
Newelska 6, PL-01-447 Warsaw, Poland
{kulczycki,mchmat}@ibspan.waw.pl
[2] Cracow University of Technology, Department of Automatic
Control and Information Technology, Cracow, Poland
[3] Catholic University of Lublin,
Institute of Mathematics and Computer Science,
Lublin, Poland

Abstract. A gradient clustering algorithm, based on the nonparametric methodology of statistical kernel estimators, expanded to its complete form, enabling implementation without particular knowledge of the theoretical aspects or laborious research, is presented here. The possibilities of calculating tentative optimal parameter values, and then – based on illustrative interpretation – their potential changes, result in the proposed Complete Gradient Clustering Algorithm possessing many original and valuable, from an applicational point of view, properties. Above all the number of clusters is not arbitrarily imposed but fitted to a real data structure. It is also possible to increase the scale of the number (still avoiding arbitrary assumptions), as well as the proportion of clusters in areas of dense and sparse situation of data elements. The method is universal in character and can be applied to a wide range of practical problems, in particular from the bioinformatics, management and engineering fields.

Keywords: Data analysis and mining, clustering, nonparametric statistical methods, kernel estimators, numerical algorithms.

1 Introduction

Consider the *m*-elements set of *n*-dimensional vectors:

$$x_1, x_2, \ldots, x_m \in \mathrm{R}^n \quad . \tag{1}$$

Generally, the task of clustering relies upon the division of the above set of data into subsets (clusters), each containing elements similar to one another, yet significantly differing from elements of other subsets. Such a general definition results in the mathematical apparatus not having a natural methodology, the existence becomes obvious of an excessive number of heuristic iterative procedures, each of them characterized by different advantages and disadvantages, as well as certain properties which may be of benefit in some problems and of no profit in others.

H. Deng et al. (Eds.): AICI 2011, Part III, LNAI 7004, pp. 497–504, 2011.

In the now classic paper [2] Fukunaga and Hostetler formulated a natural idea of clustering, making use of notable possibilities entering into widespread use of statistical kernel estimators at that time, today the main method of nonparametric estimation. The basis of the above concept is treating data set (1) as a random sample obtained from an *n*-dimensional random variable, calculating the kernel estimator of the density of its distribution, and making the clear assumption that particular clusters correspond to modes (local maxima) of the estimator. The presented method was formulated as a general idea only, leaving the details to the painstaking analysis of the user.

The aim of this publication is to present the Gradient Clustering Algorithm based on Fukunaga's and Hostetler's concept in its complete form, suitable for direct use without requiring users to have a deeper statistical knowledge or conduct laborious research. All parameters appearing here can be effectively calculated using convenient numerical procedures based on optimizing criteria. Moreover, making use of a near-intuitive interpretation of the concept of the gradient algorithm itself, as well as its theoretical base – kernel estimators, an analysis of the significance of particular parameters will be given, and the effects achieved through their possible change with respect to the above mentioned optimal values, depending on conditions of the problem in question and user preferences.

The main feature of the algorithm under research is that it does not demand strict assumptions regarding the desired number of clusters, which allows the number obtained to be better suited to a real data structure. In the paper, the parameter directly responsible for the number of clusters will be indicated. At a preliminary stage its value can be calculated effectively using optimizing criteria. It will also be shown how possible changes to this value influence the increase or decrease in the number of clusters, although without defining their exact number. Moreover, the next parameter is indicated, the value of which will influence the proportion between the number of clusters in dense and sparse areas of data set elements. Here also its value can be assumed based on optimizing reasons, or possibly submitted to modifications with the goal of increasing the number of clusters in dense areas of data set elements while simultaneously reducing or even eliminating them from sparse regions, or vice-versa. This possibility is particularly worth underlining as practically non-existent in other clustering procedures. Moreover, the appropriate relation between the two above mentioned parameters allows for a reduction, or even elimination of clusters in sparse areas, without influencing the number of clusters in dense areas of data set elements.

The broader description of the method presented here is available in the paper [4].

2 Statistical Kernel Estimators

Let the *n*-dimensional random variable X be given, with a distribution characterized by the density f. Its kernel estimator $\hat{f}: \mathrm{R}^n \to [0,\infty)$, calculated using experimentally obtained values for the *m*-element random sample (1), in its basic form is defined as

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h} \right) , \tag{2}$$

where $m \in \mathsf{N} \setminus \{0\}$, the coefficient $h > 0$ is called a smoothing parameter, while the measurable function $K : \mathsf{R}^n \to [0, \infty)$ of unit integral $\int_{\mathsf{R}^n} K(x)\,\mathrm{d}x = 1$, symmetrical with respect to zero and having a weak global maximum in this place, takes the name of a kernel. The choice of form of the kernel K and the calculation of the smoothing parameter h is made most often with the criterion of the mean integrated square error.

Thus, the choice of the kernel form has – from a statistical point of view – no practical meaning and thanks to this, it becomes possible to take into account primarily properties of the estimator obtained and/or aspects of calculations, advantageous from the point of view of the applicational problem under investigation (for broader discussion see [3 – Section 3.1.3], [8 – Sections 2.7 and 4.5]).

The fixing of the smoothing parameter h has significant meaning for quality of estimation. Too small a value causes a large number of local extremes of the estimator $\hat{f}$ to appear, which is contrary to the actual properties of real populations. On the other hand, too big values of the parameter h result in overflattening of this estimator, hiding specific properties of the distribution under investigation. In practice, the value of the smoothing parameter h can be calculated with confirmed algorithms available in literature. In the multidimensional case the most common and universal cross-validation method is proposed [3 – Section 3.1.5], [7 – Section 3.4.3], however in the one-dimensional case the convenient plug-in method [3 – Section 3.1.5], [8 – Section 3.6.1] can be recommended.

For the basic definition of kernel estimator (2), the influence of the smoothing parameter on particular kernels is the same. Advantageous results are obtained thanks to the individualization of this effect, achieved through so-called modification of the smoothing parameter [3 – Section 3.1.6], [7 – Section 5.3.1]. It relies on mapping the positive modifying parameters s_1, $s_2, \ldots, s_m$ on particular kernels, described as

$$s_i = \left(\frac{\hat{f}_*(x_i)}{\bar{s}} \right)^{-c}, \tag{3}$$

where $c \in [0, \infty)$, $\hat{f}_*$ denotes the kernel estimator without modification, $\bar{s}$ is the geometrical mean of the numbers $\hat{f}_*(x_1)$, $\hat{f}_*(x_2)$, ... , $\hat{f}_*(x_m)$, and finally, defining the kernel estimator with modification of the smoothing parameter in the following form:

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} \frac{1}{s_i^n} K\left(\frac{x - x_i}{h s_i} \right). \tag{4}$$

Thanks to the above procedure, the areas in which the kernel estimator assumes small values (e.g. in the range of "tails"), are additionally flattened, and the areas connected with large values – peaked, which allows to better reveal individual properties of a distribution. The parameter c stands for the intensity of the modification procedure. Based on indications for the criterion of the integrated mean square error, the value

$$c = 0.5 \tag{5}$$

can be suggested.

Practical applications may also use some other additional procedures, generally improving the quality of the estimator, and others – optional – possibly fitting the model to an existing reality. For the first group one should recommend a linear transformation [3 – Section 3.1.4], [7 – Section 4.2.1], while for the second, the boundaries of a support [3 – Section 3.1.8], [7 – Section 2.10].

Detailed information regarding kernel estimators can be found in the monographs [3], [7], [8].

3 Complete Gradient Clustering Algorithm

Consider – as in the Introduction – the m-elements set of n-dimensional vectors (1). This will be treated as a random sample obtained from the n-dimensional random variable X, with distribution having the density f. Using the methodology described in Section 2, the kernel estimator $\hat{f}$ can be created. Take the natural assumption that particular clusters are related to its modes, or local maxima of the function $\hat{f}$, and mapping onto them elements of set (1) is realized by transposing those elements in the gradient $\nabla\hat{f}$ direction, with the appropriate fixed step.

The above is carried out iteratively with the Gradient Clustering Algorithm, based on the classic Newtonian procedure, defined as

$$x_j^0 = x_j \quad \text{for} \quad j = 1, 2, \ldots, m \tag{6}$$

$$x_j^{k+1} = x_j^k + b\frac{\nabla\hat{f}(x_j^k)}{\hat{f}(x_j^k)} \quad \text{for} \quad j = 1, 2, \ldots, m \quad \text{and} \quad k = 0, 1, \ldots, k^* \,, \tag{7}$$

where $b > 0$ and $k^* \in \mathbb{N} \setminus \{0\}$. In practice it is recommended that $b = h^2/(n+2)$.

In order to refine the above concept to the state of a complete algorithm, the following aspects need to be formulated and analyzed in detail:

1. formula of the kernel estimator $\hat{f}$;
2. setting a stop condition (and consequently the number of steps k^*);
3. definition of a procedure for creating clusters and assigning to them particular elements of set (1), after the last, k^*-th step;
4. analysis of influence of the values of parameters on results obtained.

The above tasks are the subjects of the following sections.

3.1 Formula of the Kernel Estimator

For the needs of further parts of the concept presented here, the kernel estimator $\hat{f}$ is assumed in a form with modification of smoothing parameter of standard intensity (5).

The kernel K is recommended in the most universal normal form [3 – Section 3.1.3], [8 – Sections 2.7 and 4.5] due to its differentiability in the whole domain, convenience for analytical considerations connected with gradient, and assuming positive values, which in every case guards against division by zero in formula (7).

3.2 Setting a Stop Condition

It is assumed that algorithm (6)-(7) should be finished, if after the consecutive k-th step the following condition is fulfilled

$$\left|D_k - D_{k-1}\right| \le aD_0 \ , \tag{8}$$

where $a > 0$ and

$$D_0 = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d(x_i, x_j) \tag{9}$$

$$D_{k-1} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d(x_i^{k-1}, x_j^{k-1}) \ , \quad D_k = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d(x_i^{k}, x_j^{k}) \ , \tag{10}$$

while d means Euclidean metric in R^n. Therefore, D_0 and D_{k-1}, D_k denote sums of distances between particular elements of set (1) before starting the algorithm and after the (k–1)-th and k-th step, respectively. Primarily it is recommended that $a = 0.001$; the potential decrease of this value does not significantly influence the obtained results, although increases require individual verification of their correctness.

Finally, if after the k-th step condition (8) is fulfilled, then

$$k^* = k \tag{11}$$

and consequently this step is treated as the last one.

3.3 Procedure for Creating Clusters and Assigning Particular Elements to Them

At this stage the following set is investigated

$$x_1^{k^*}, \ x_2^{k^*}, \ldots, x_m^{k^*} \ , \tag{12}$$

consisting of the elements of set (1) after the k^*-th step of algorithm (6)-(7). Following this, the set of mutual distances of the above elements

$$\left\{ d(x_i^{k^*}, x_j^{k^*}) \right\}_{\substack{i=1,2,\ldots,m-1 \\ j=i+1,i+2,\ldots,m}} \tag{13}$$

should be defined. Taking the above set as a sample of a one-dimensional random variable, the auxiliary kernel estimator $\hat{f}_d$ of mutual distances of the elements of set

(12) ought to be calculated. Regarding the methodology of kernel estimators presented in Section 2, normal kernel is once again proposed, as is the use of the procedure of smoothing parameter modification with standard value of parameter (5), and additionally left-sided boundary of a support to the interval $[0, \infty)$.

The next task is to find – with suitable precision – the "first" (i.e. for the smallest value of an argument) a local minimum of the function $\hat{f}_d$ belonging to the interval $(0, D)$, where $D = \max_{\substack{i=1,2,\ldots,m-1 \\ j=i+1,i+2,\ldots,m}} d(x_i, x_j)$. For this purpose one should treat set (13) as a random sample, calculate its standard deviation σ_d, and next take in sequence the values x from the set

$$\{ 0 , \ 0.01 \cdot \sigma_d , \ 0.02 \cdot \sigma_d , \ \ldots \ , \ [\mathrm{int}(100 \cdot D) - 1] \cdot 0.01 \cdot \sigma_d \} , \tag{14}$$

where $\mathrm{int}(100 \cdot D)$ denotes an integral part of the number $100 \cdot D$, until the finding of the first (the smallest) of them which fulfils the condition

$$\hat{f}_d (x - 0.01 \sigma_d) > \hat{f}_d (x) \quad \text{and} \quad \hat{f}_d (x) \leq \hat{f}_d (x + 0.01 \sigma_d) . \tag{15}$$

Such calculated value [1] will be denoted hereinafter as x_d, and it can be interpreted as half the distance between „centers" of potential clusters lying closest together.

Finally, the clusters will be created. To this aim one should:

1. take the element of set (12) and initially create a one-element cluster containing it;
2. find an element of set (12) different from the one in the cluster, closer than x_d; if there is such an element, then it should be added to the cluster; in the other case – proceed to point 4;
3. find an element of set (12) different from elements in the cluster, closer than x_d to at least one of them; if there is such an element, then it should be added to the cluster and point 3 repeated;
4. add the obtained cluster to a "list of clusters" and remove from set (12) elements of this cluster; if this so-reduced set (12) is not empty, return to point 1; in the other case – finish the algorithm.

3.4 Analysis of Influence of the Values of Parameters on Results Obtained

It is worth repeating that the presented clustering algorithm did not require a preliminary, often arbitrary in practice, assumption concerning number of clusters – their size depending solely on the internal structure of data, given as set (1). In the application of the Complete Gradient Clustering Algorithm in the presented above basic form, the values of the parameters used are effectively calculated taking optimizing reasons into account. However, optionally – if the researcher makes the decision – by an appropriate change in values of kernel estimator parameters it is

[1] If such a value does not exist, then one should recognize the existence of one cluster and finish the procedure.

possible to influence the size of number of clusters, and also the proportion of their appearance in dense areas in relation to sparse regions of elements in this set.

As mentioned in Section 2, too small a value of the smoothing parameter h results in the appearance of too many local extremes of the kernel estimator, while too great a value causes its excessive smoothing. In this situation lowering the value of the parameter h in respect to that obtained by procedures based on the criterion of the mean integrated square error creates as a consequence an increase in the number of clusters. At the same time, an increase in the smoothing parameter value results in fewer clusters. It should be underlined that in both cases, despite having an influence on size of cluster number, their exact number will still depend solely on the internal structure of data. Based on research carried out one can recommend a change in the value of the smoothing parameter of between –25% and +50%. Outside of this range, results obtained require individual verification.

Next, as mentioned in Section 2, the intensity of modification of the smoothing parameter is implied by the value of the parameter c, given as standard by formula (5). Its increase smoothes the kernel estimator in areas where elements of set (1) are sparse, and also sharpens it in dense areas – in consequence, if the value of the parameter c is raised, then the number of clusters in sparse areas of data decreases, while at the same time increasing in dense regions. Inverse effects can be seen in the case of lowering this parameter value. Based on research carried out one can recommend the value of the parameter c to be between 0 (meaning no modification) and 1.5. An increase greater than 1.5 requires individual verification of the validity of results obtained. Particularly it is recommended that $c = 1$.

Practice, however, often prevents changes to the clusters in dense areas of data – the most important from an applicational point of view – while at the same time requiring a reduction or even elimination of clusters in sparse regions, as they frequently pertain to atypical elements (outliers). Putting the above considerations together, one can propose an increase of both the standard scale of the smoothing parameter modification (5) and the value of the smoothing parameter h calculated on the criterion of the mean integrated square error, to the value h^* defined by the formula

$$h^* = \left(\frac{3}{2}\right)^{c-0.5} h \ . \tag{16}$$

The joint action of both these factors results in a twofold smoothing of the function $\hat{f}$ in the regions where the elements of set (1) are sparse. Meanwhile these factors more or less compensate for each other in dense areas, thereby having practically no influence on the detection of these clusters. Based on research carried out one can recommend a change in the value of the parameter c from 0.5 to 1.0. Increasing it to above 1.0 demands individual verification of the validity of results obtained. Particularly it is recommended that $c = 0.75$.

More details of the method presented above, with illustrative examples, can be found in the paper [4].

4 Final Comments

The algorithm described in this paper was comprehensively tested both for random statistical data as well as generally available benchmarks. It was also compared with other well-known clustering methods, k-means and hierarchical procedures. It is difficult to confirm here the absolute supremacy of any one of them – to a large degree the advantage stemmed from the conditions and requirements formulated with regard to the problem under consideration, although the Complete Gradient Clustering Algorithm allowed for greater possibilities of adjustment to the real structure of data, and consequently the obtained results were more justifiable to a natural human point of view. A very important feature for practitioners was the possibility of firstly functioning using standard parameters values, and the option of changing them afterwards – according to individual needs – by the modification of two of them with easy and illustrative interpretations. These properties were actively used in three projects from the domains of bioinformatics (categorization of grains for seed production [1]), management (marketing support strategy for mobile phone operator [5]) and engineering (synthesis of fuzzy PID controller [6]).

References

1. Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Lukasik, S., Zak, S.: Gradient Clustering Algorithm for Features Analysis of X-Ray Images. In: Pietka, E., Kawa, J. (eds.) Information Technologies in Biomedicine, vol. 2, pp. 15–24. Springer, Berlin (2010)
2. Fukunaga, K., Hostetler, L.D.: The estimation of the gradient of a density function, with applications in Pattern Recognition. IEEE Transactions on Information Theory 21, 32–40 (1975)
3. Kulczycki, P.: Estymatory jadrowe w analizie systemowej. WNT, Warsaw (2005)
4. Kulczycki, P., Charytanowicz, M.: A Complete Gradient Clustering Algorithm Formed with Kernel Estimators. International Journal of Applied Mathematics and Computer Science 20, 123–134 (2010)
5. Kulczycki, P., Daniel, K.: Metoda wspomagania strategii marketingowej operatora telefonii komorkowej. Przeglad Statystyczny 56, 116–134 (2009)
6. Lukasik, S., Kowalski, P.A., Charytanowicz, M., Kulczycki, P.: Fuzzy Models Synthesis with Kernel-Density Based Clustering Algorithm. In: Ma, J., Yin, Y., Yu, J., Zhou, S. (eds.) Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 3, pp. 449–453. IEEE Computer Society, Los Alamitos (2008)
7. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
8. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1994)