Exemplary Applications of the Complete Gradient Clustering Algorithm in Bioinformatics, Management and Engineering

P. Kulczycki[1,2], M. Charytanowicz[1,3], P.A. Kowalski[1,2], S. Lukasik[1,2]

**[1] Polish Academy of Sciences, Systems Research Institute,
Centre of Information Technology for Data Analysis Methods, Warsaw, Poland.**

**[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.

e-mail: {Piotr.Kulczycki, Malgorzata.Charytanowicz, Piotr.A.Kowalski, Szymon.Lukasik}@ibspan.waw.pl

Abstract: This publication deals with the applicational aspects and possibilities of the Complete Gradient Clustering Algorithm – the classic procedure of Fukunaga and Hostetler, prepared to a ready-to-use state, by providing a full set of procedures for defining all functions and the values of parameters. Moreover, it describes how a possible change in those values influences the number of clusters and the proportion between their numbers in dense and sparse areas of data elements. The possible uses of these properties were illustrated in practical tasks from bioinformatics (the categorization of grains for seed production), management (the design of a marketing support strategy for a mobile phone operator) and engineering (the synthesis of a fuzzy controller).

Keywords: exploratory data analysis, clustering, nonparametric methods, kernel estimators, seed production, mobile phone operator, fuzzy controller.

1. Introduction

Clustering is becoming a fundamental procedure in exploratory data analysis [2]. However it lacks natural mathematical apparatus, such as – for example – differential calculus for investigating the extremes of a function. In this situation the ambiguity of an interpretation (important mainly in practical applications) as well as particular factors of the definition itself (e.g. the meaning of “similarity” and consequently “dissimilarity” of elements) imply a huge variety of concepts and thus of clustering procedures. On one hand this significantly hinders the research, but on the other it allows to better suit the applied method to the specifics and requirements of a investigated task. This work deals with the applicational properties of the so-called Complete Gradient Clustering Algorithm, illustrated in examples of practical problems of bioinformatics, management and engineering.

2. Main contents

Consider the *m* elements data set comprised of *n*-dimensional vectors

$$x_1,\ x_2, \dots, x_m \in \mathbb{R}^n \quad , \tag{1}$$

treated here as a sample obtained from an *n*-dimensional real random variable. In their now seminal paper [3], Fukunaga and Hostetler formulated a natural and effective concept of clustering, making use of the significant possibilities of statistical kernel estimators [5, 11, 12], which were becoming more widely applied at that time. The basis for this concept is accepting data set (1) as a random sample obtained from a certain *n*-dimensional random variable, calculating a kernel estimator for its distribution density and making a natural assumption that particular clusters are related to modes (local maxima) of the resulting estimator (in consequence “valleys” of the density function become borders for such-formed clusters). The method presented then was formulated as a general idea only, leaving detailed analysis to the user. In the paper [7], the Fukunaga and Hostetler algorithm was supplemented and finally given in its complete form, useful for application without the necessity of deeper statistical knowledge or laborious calculations and investigations. It is characterized by the following features:

1. all parameters can be effectively calculated using numerical procedures based on optimizing criteria;
2. the algorithm 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;
3. the parameter directly responsible for the number of clusters is indicated; it will also be shown how possible changes – e.g. with regard to values calculated using optimizing criteria (see point 1 stated above) – to this value, influence the increase or decrease in the number of clusters without, however, defining their exact number;
4. moreover, the next parameter can be easily indicated, the value of which will influence the proportion between the number of clusters in dense and sparse areas of elements of data set (1); here also the value of this parameter can be assumed based on optimizing criteria (see again point 1); it will also be shown here that potential lowering of the value of this parameter results in a decrease in the number of clusters in dense regions of data as the number of clusters in sparse areas increases, while a potential raise in its value has the opposite effect – increasing the number of clusters in dense areas while simultaneously reducing or even eliminating them from sparse regions of data set (1);
5. the appropriate relation between the two above-mentioned parameters allows for a reduction, or even elimination of clusters in sparse areas, practically without influencing the number of clusters in dense areas of data set elements;
6. the algorithm also creates small, even single-element clusters, which can be treated as atypical elements (outliers) in a given configuration of clusters, which makes possible their elimination or assignation to the closest cluster by a change – described in points 3 and particularly 4 or 5 – in the values of the appropriate parameters.

The features in point 4, and in consequence 5, are particularly worth underlining as practically non-existent in other clustering procedures. In practical applications it is also worth highlighting the implications of points 1 and 2, and potentially 3. Unusual possibilities are offered by the property expressed in point 6.

The Complete Gradient Clustering Algorithm presented here 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 aforementioned projects from the domains of bioinformatics, management and engineering concerning the categorization of grains for seed production [1], the design of a marketing support strategy for a mobile phone operator [9] and the synthesis of a fuzzy controller for the reduction of a rule set [4, 10], respectively.

More details of the material presented above is available in the paper [8], which will appear soon.

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. (Editors) Information Technologies in Biomedicine. Springer-Verlag, Berlin (2010), Vol. 2, pp. 15-24.

[2] Everitt B.S., Landau S., Leese M.: *Cluster Analysis*. Arnold, London (2001).

[3] Fukunaga K., Hostetler L.D.: *The estimation of the gradient of a density function, with applications in Pattern Recognition*. IEEE Transactions on Information Theory, Vol. 21 (1975), pp. 32-40.

[4] Kowalski P.A., Lukasik S., Charytanowicz M., Kulczycki P.: *Data-Driven Fuzzy Modeling and Control with Kernel Density Based Clustering Technique*. Polish Journal of Environmental Studies, Vol. 17 (2008), pp. 83-87.

[5] Kulczycki P.: *Estymatory jadrowe w analizie systemowej*. WNT, Warsaw (2005).

[6] Kulczycki P., Hryniewicz O., Kacprzyk J. (Editors): *Techniki informacyjne w badaniach systemowych*. WNT, Warsaw (2007).

[7] Kulczycki P., Charytanowicz M.: *A Complete Gradient Clustering Algorithm Formed with Kernel Estimators*. International Journal of Applied Mathematics and Computer Science, Vol. 20 (2010), pp. 123-134.

[8] Kulczycki P., Charytanowicz M., Kowalski P.A., Lukasik S.: *The Complete Gradient Clustering Algorithm: Properties in Practical Applications*. Journal of Applied Statistics. In press (2012).

[9] Kulczycki P., Daniel K.: *Metoda wspomagania strategii marketingowej operatora telefonii komorkowej*. Przeglad Statystyczny, Vol. 56 (2009), pp. 116-134.

[10] Lukasik S., Kowalski P.A., Charytanowicz M., Kulczycki P.: *Fuzzy Models Synthesis with Kernel-Density Based Clustering Algorithm*. In: Proc. Fifth International

Conference on Fuzzy Systems and Knowledge Discovery, Ma J., Yin Y., Yu J., Zhou S. (Editors), Jinan (China), 18-20 October 2008, Vol. 3 (2008), pp. 449-453.

[11] Silverman B.W.: *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London (1986).

[12] Wand M.P., Jones M.C.: *Kernel Smoothing*. Chapman and Hall, London, (1994).