Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images

Małgorzata Charytanowicz, Jerzy Niewczas, Piotr Kulczycki, Piotr A. Kowalski,
Szymon Łukasik, and Sławomir Żak

Abstract Methods based on kernel density estimation have been successfully applied for various data mining tasks. Their natural interpretation together with suitable properties make them an attractive tool among others in clustering problems. In this paper, the Complete Gradient Clustering Algorithm has been used to investigate a real data set of grains. The wheat varieties, Kama, Rosa and Canadian, characterized by measurements of main grain geometric features obtained by X-ray technique, have been analyzed. The proposed algorithm is expected to be an effective tool for recognizing wheat varieties. A comparison between the clustering results obtained from this method and the classical *k*-means clustering algorithm shows positive practical features of the Complete Gradient Clustering Algorithm.

1 Introduction

Clustering is a major technique for data mining, used mostly as an unsupervised learning method. The main aim of cluster analysis is to partition a given population into groups or clusters with common characteristics, since similar objects are grouped together, while dissimilar objects belong to different clusters [4, 11]. As a result, a new set of categories of interest, characterizing the population, is discovered. The clustering methods are generally divided into six groups: hierarchical,

M. Charytanowicz and J. Niewczas
Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin, Konstantynów 1 H, PL 20-708 Lublin, Poland
e-mail: {mchmat,jniewczas}@kul.lublin.pl

P. Kulczycki, P.A. Kowalski, S. Łukasik, and S. Żak
System Research Institute, Polish Academy of Sciences, Newelska 6, PL 01-447 Warsaw, Poland
Department of Automatic Control and Information Technology, Cracow University of Technology, Warszawska 24, PL 31-155 Cracow, Poland
e-mail: {kulczycki,pakowal,slukasik,slzak}@ibspan.waw.pl

partitioning, density-based, grid-based, and soft-computing methods. These numerous concepts of clustering are implied by different techniques of determination of the similarity and dissimilarity between objects. A classical partitioning *k*-means algorithm is concentrated on measuring and comparing the distances among them. It is computationally attractive and easy to interpret and implement in comparison to other methods. On the other hand, the number of clusters is assumed here by user in advance and therefore the nature of the obtained groups may be unreliable for the nature of the data, usually unknown before processing.

The rigidity of arbitrary assumptions concerning the number or shape of clusters among data can be overcome by density-based methods that let the data detect inherent data structures. In the paper [9], the Complete Gradient Clustering Algorithm was introduced. The main idea of this algorithm assumes that each cluster is identified by local maxima of the kernel density estimator of the data distribution. The procedure does not need any assumptions concerning the data and may be applied to a wide range of topics and areas of cluster analysis [3, 9, 10].

The main purpose of this work is to propose an effective technique for forming proper categories of wheat. In the earliest attempts to classify wheat grains a geometry and set of parameters were defined. The size, shape and colour of grain because of their heritable characters, can be used for wheat variety recognition. Accomplished studies showed that digital image processing techniques commonly used in multivariate analysis give reliable results in classification process [13, 15, 17]. In this paper, the algorithm proposed in [9] will be used to identify wheat varieties, using their main geometric features.

2 Complete Gradient Clustering Algorithm (CGCA)

In this section, the Complete Gradient Clustering Algorithm, for short the CGCA, is shortly described. The principle of the proposed algorithm is based on the distribution of the data; the implementation of the CGCA needs to estimate its density. Each cluster is characterized by a local maximum of the kernel density estimator. As a result, regions of high densities of objects are recognized as clusters, while areas with sparse distributions of objects divide one group from another. Data points are assigned to clusters by using an ascending gradient method, i.e. points moving to the same local maximum are put into the same cluster. The algorithm works in an iterative manner until a termination criterion has been satisfied.

2.1 Kernel Density Estimation

Suppose that $x_1, x_2, \ldots, x_m$ is a random sample of m points in n-dimensional space from an unknown distribution with density f. Its kernel estimator can be defined as

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

where the positive coefficient h is called the smoothing parameter or bandwidth, while the measurable function $K : R^n \to [0, \infty)$ of unit integral $\int_{R^n} K(x)\mathrm{d}x = 1$, unimodal and symmetrical with respect to zero, takes the name of a kernel [5, 14].

It is generally accepted, that the choice of the kernel K is not as important as the choice of the coefficient h and thank to this, it is possible to take into account the primarily properties of the estimator obtained. Most often the standard normal kernel given by

$$K(x) = \frac{1}{2\pi^{n/2}} e^{-\frac{x^{\mathrm{T}}x}{2}} \tag{2}$$

is used. It is differentiable up to any order and assumes positive values in the whole domain.

The practical implementation of the kernel density estimators requires a proper choice of the bandwidth h. In practice the best value of h is mostly taken as the value that minimizes the mean integrated square error. A frequently used bandwidth selection method is based on the approach of least-squares cross validation [5, 14]. The value of h is chosen to minimize the function $M : (0, \infty) \to R$ given by the rule:

$$M(h) = \frac{1}{m^2 h^n} \sum_{i=1}^{m} \sum_{j=1}^{m} \widetilde{K}\left(\frac{x_j - x_i}{h}\right) + \frac{2}{mh^n} K(0), \tag{3}$$

where $\widetilde{K}(x) = K^{*2}(x) - 2K(x)$ and K^{*2} is the convolution square of the function K; for the standard normal kernel (2):

$$K^{*2}(x) = \frac{1}{(4\pi)^{n/2}} e^{-\frac{x^{\mathrm{T}}x}{4}}. \tag{4}$$

In this case the influence of the smoothing parameter on particular kernels is the same. The individualization of this effect may be achieved through the modification of the smoothing parameter. This relies on introducing the positive modifying parameters $s_1, s_2, \ldots, s_m$ mapped on particular kernels, described by the formula

$$s_i = \left(\frac{\hat{f}_*(x_i)}{\widetilde{s}}\right)^{-c}, \tag{5}$$

where $c \in [0, \infty)$, $\hat{f}_*$ is the kernel estimator in its basic form (1) and $\widetilde{s}$ denotes the geometrical mean of the numbers $\hat{f}_*(x_1)$, $\hat{f}_*(x_2)$, …, $\hat{f}_*(x_m)$. The value of the parameter c implies the intensity of modification of the smoothing parameter. Based on indications for the criterion of the mean integrated square error the value 0.5 as c is proposed. Finally, the kernel estimator with modification of the smoothing parameter is defined as

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} \frac{1}{s_i^n} K\left(\frac{x - x_i}{hs_i}\right). \tag{6}$$

Additional procedures improving the quality of the estimator obtained, such as a linear transformation and support boundary, as well as the general aspects of the theory of statistical kernel estimators are found in [5, 6, 14]. Exemplary practical applications are presented in the publications [1, 3, 7, 8, 10].

2.2 Procedures of the CGCA

Consider the data set containing m elements $x_1, x_2, \ldots, x_m$ in n-dimensional space. Using the methodology introduced in Subsect. 2.1, the kernel density estimator $\hat{f}$ may be constructed. The idea of the CGCA is based on the approach proposed by Fukunaga and Hostetler [2]. Thus given the start points:

$$x_j^0 = x_j \text{ for } j = 1, 2, \ldots, m, \tag{7}$$

each point is moved in an uphill gradient direction using the following iterative formula:

$$x_j^{k+1} = x_j^k + b \frac{\nabla \hat{f}(x_j^k)}{\hat{f}(x_j^k)} \text{ for } j = 1, 2, \ldots, m \text{ and } k = 0, 1, \ldots\ , \tag{8}$$

where $\nabla \hat{f}$ denotes the gradient of kernel estimator $\hat{f}$ and the value of the parameter b is proposed as $h^2/(n+2)$ while the coefficient h is the bandwidth of $\hat{f}$.

The algorithm will be stopped when the following condition is fulfilled:

$$|D_k - D_{k-1}| \leq \alpha D_0, \tag{9}$$

where D_0 and D_{k-1}, D_k denote sums of Euclidean distances between particular elements of the set x_1, x_2, …, x_m before starting the algorithm as well as after the $(k-1)$-th and k-th step, respectively. The positive parameter α is taken arbitrary and the value 0.001 is primarily recommended. This k-th step is the last one and will be denoted hereinafter by k^*.

Finally, after the k^*-th step of the algorithm (7)-(8) the set

$$x_1^{k^*}, x_2^{k^*}, \ldots, x_m^{k^*}, \tag{10}$$

considered as the new representation of all points x_1, x_2, …, x_m, is obtained. Following this, the set of mutual Euclidean distances of the above elements:

$$\left\{ d(x_i^{k^*}, x_j^{k^*}) \right\}_{\substack{i=1,2,\ldots,m-1 \\ j=i+1,i+2,\ldots,m}} \tag{11}$$

is defined. Using the methodology presented in Subsect. 2.1, the auxiliary kernel estimator $\hat{f}_d$ of the elements of the set (11), treated as a sample of a one-dimensional random variable, is created under the assumption of nonnegative support. Next, the first (i.e. obtained for the smallest value of an argument) local minimum of the function $\hat{f}_d$ belonging to the interval $(0, D]$, where D means the maximum value of the set (11), is found. This local minimum will be denoted as x_d, and it can be interpreted as the half-distance between potential closest clusters. Finally, the clusters are created. First, the element of the set (11) is taken; it initially create a one-element cluster containing it. An element of the set (11) is added to the cluster if the distance between it and any element belonging to the cluster is less than x_d. Every added element is removed from the set (11). If there are no more elements belonging to the cluster, the new cluster is created. The procedure of assigning elements to clusters is repeated as long as the set (11) is not empty.

Procedures described above constitute the Complete Gradient Algorithm in its basic form. The values of the parameters used are calculated automatically, using optimization criteria. However, by an appropriate change in values of these parameters it is possible to influence the size of number of clusters, and also the proportion of their appearnce in dense areas in relation to sparse regions of elements in this set. Namely, lowering (raising) the value of smoothing parameter h results in raising (lowering) the number of local maxima. A change in the value of that parameter of between -25% and +50% is recommended. Next, raising the intensity c of the smoothing parameter modification results in decreasing the number of clusters in sparse areas of data and increasing their number in dense regions. Inverse effects can be seen in the case of lowering this parameter value. The value of the parameter c to be between 0 and 1.5 is recommended. Finally, an increase of both parameters c and h can be proposed. Then the additional formula

$$h^* = \left(\frac{3}{2}\right)^{c-0.5} h \tag{12}$$

is used for calculating the smoothing parameter h^*, where the value of the parameter h is calculated on the criterion of the mean integrated square error. The joint action of both these factors results in a twofold smoothing of the function $\hat{f}$ in the regions where the elements of the set $x_1, x_2, \ldots, x_m$ are sparse. Meanwhile these factors more or less compensate for each other in dense areas, thereby having small influence on the detection of clusters located there. Detailed information on the CGCA procedures and their influences on the clustering results is described in [9].

3 Materials and methods

The proposed algorithm has been applied for wheat variety recognition. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13 × 18 cm X-ray KODAK plates. Figure 1 presents the X-ray images of these kernels.

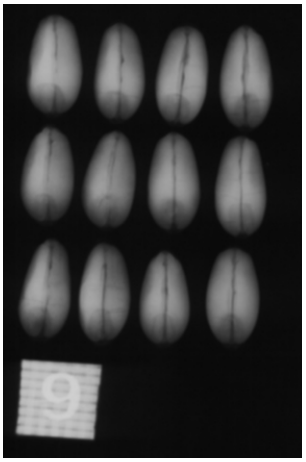


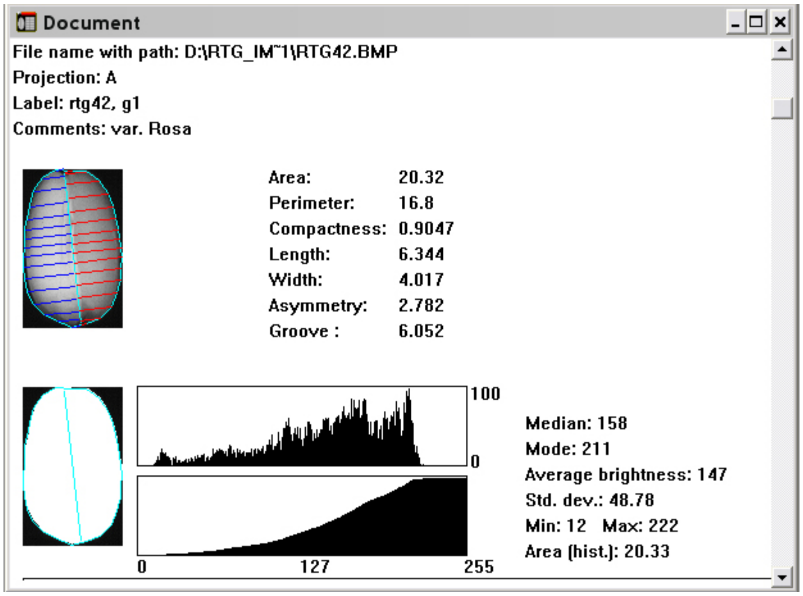
Fig. 1 X-ray photogram (13 × 18 cm) of kernels

The X-ray photograms were scanned using the Epson Perfection V700 table photo-scanner with a built-in transparency adapter, 600 dpi resolution and 8 bit gray scale levels. Analysis procedures of obtained bitmap graphics files were based on the computer software package GRAINS, specially developed for X-ray diagnostic of wheat kernels [12, 16]. To construct the data, seven geometric parameters of wheat kernels: area A, perimeter P, compactness $C = 4\pi A/P^2$, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove, were measured from a total of 210 samples (see Fig. 2). All of these parameters were real-valued continuous.

In our investigations, the data was reduced to be two-dimensional after applying the Principal Component Analysis [4] to validate the results visually.

4 Results and discussion

The data's projection on the axes of the two greatest principal components, with wheat varieties being distinguished symbolically, is presented in Fig. 3. Samples were labeled by numbers: 1-70 for the Kama wheat variety, 71-140 for the Rosa wheat variety, and 141-210 for the Canadian wheat variety. To discuss the clustering



File name with path: D:\RTG_IM~1\RTG42.BMP
Projection: A
Label: rtg42, g1
Comments: var. Rosa

Area: 20.32
Perimeter: 16.8
Compactness: 0.9047
Length: 6.344
Width: 4.017
Asymmetry: 2.782
Groove : 6.052

Median: 158
Mode: 211
Average brightness: 147
Std. dev.: 48.78
Min: 12 Max: 222
Area (hist.): 20.33

Fig. 2 Document window with geometric parameters of a kernel and statistical parameters of its image (as a unit of measure millimeters were used)

results obtained by the CGCA, mistakenly classified samples are displayed with their labels (see Fig. 3).

Using procedures described in Subsect. 2.2 allowing elimination of clusters in sparse areas, the CGCA created three clusters corresponding to Rosa, Kama, and Canadian varieties, containing 69, 65, and 76 elements respectively. Thus the samples 9 and 38, which belong to the Kama wheat variety are incorrectly grouped into the cluster associated with the Rosa wheat variety. What is more, the samples 125, 136, 139, which belong to the Rosa wheat variety, and the samples 166, 200, 202, which belong to the Canadian wheat variety are mistakenly classified into the cluster associated with the Kama wheat variety. In addition, the samples 20, 27, 28, 30, 40, 60, 61, 64, 70, which belong to the Kama wheat variety are mistakenly classified into the cluster associated with the Canadian wheat variety. It is worth noticing however, that in the case of samples 9 and 38, misclassification can be justifiable – both samples lie close to the area of a high density of the Rosa wheat variety samples. The same problem is discerned with samples 125, 136, 139 and 166, 200, 202, which are placed close to samples of the Kama wheat variety. Similarly, mistakenly classified samples 20, 27, 28, 30, 40, 60, 61, 64, 70 lie very close to samples of the Canadian wheat variety. Thus, taking into consideration characteristics of wheat varieties, the CGCA seems to be an effective technique for wheat variety recognition.

Clustering results, containing numbers of samples classified properly and mistakenly into clusters associated with Rosa, Kama, and Canadian varieties, are shown in Table 1.

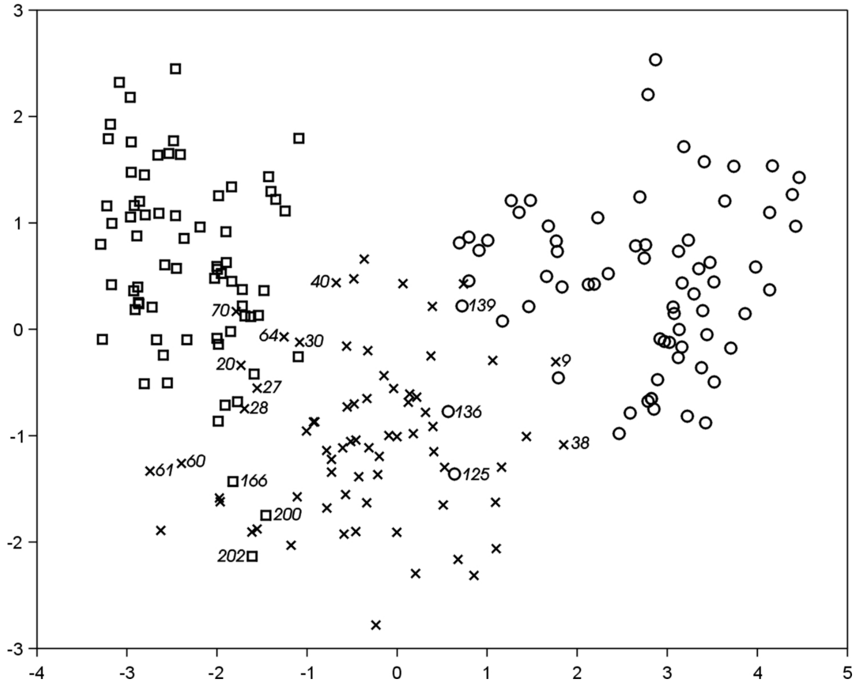


Fig. 3 Wheat varieties data set on the axes of the two greatest principal components: (○) the Rosa wheat variety, (×) the Kama wheat variety, (□) the Canadian wheat variety

Table 1 Clustering results for the wheat varieties data set

	Number of elements in clusters		
Clusters	Correctly classified	Incorrectly classified	Total
Rosa	67	2	69
Kama	59	6	65
Canadian	67	9	76

According to the results of the CGCA, out of 70 kernels of the Rosa wheat variety, 67 were classified properly. Only 2 of the Kama variety were classified mistakenly as the Rosa variety. For the other two varieties, the CGCA created clusters containing 65 elements (the Kama variety) and 76 elements (the Canadian variety). In regard to the Kama variety, 59 kernels were classified correctly, while 6 of the other varieties were incorrectly identified as the Kama variety. For the Canadian variety, 67 kernels were correctly identified and 9 kernels of the Kama variety were mistakenly identified as the Canadian variety. The results of Kama and Canadian varieties are not so satisfactory as for Rosa and this implies that these two varieties could not be so clearly distinguished as the Rosa variety, when using main geometric parameters.

Table 2 Correctness percentages for the wheat varieties data set

Wheat Varieties	Correctness %
Rosa	96
Kama	84
Canadian	96

The percentages of correctness of the CGCA are presented in Table 2. The proposed algorithm achieved an accuracy of about 96% for the Rosa wheat variety, 84% for the Kama wheat variety, and 96% for the Canadian wheat variety.

The comparable percentages of correctness of classification has been obtained when the *k*-means algorithm with arbitrary taken cluster number of 3 was used. It is worth stressing however, that this algorithm availed of the a priori assumed correct number of clusters, which in many applications may not be known, or even such a "correct" – from a theoretical point of view – number might not exist at all. The CGCA instead does not require strict assumptions regarding the desired number of cluster, which allows the number obtained to be better suited to a real data structure. Moreover, in its basic form values of parameters may be calculated automatically, however there exists the possibility of their optional change. A feature specific to it is the possibility to influence the proportion between the number of clusters in areas where data elements are dense as opposed to their sparse regions. In addition, by the detection of one-element clusters the algorithm allows the identification of outliers, which enables their elimination or designation to more numerous clusters, thus increasing the homogeneity of the data set.

5 Conclusions

The proposed clustering algorithm, based on kernel estimator methodology, is expected to be an effective technique for wheat variety recognition. It performs comparable with respect to the classical *k*-means algorithm, however requires no a priori information about the data. The data reduced after applying the Principal Component Analysis, contained apparent clustering structures according to their classes. The amount of 193 kernels, giving almost 92% of the total, was classified properly. The wheat varieties used in the study showed differences in their main geometric parameters. The Rosa variety is better recognized, whilst Kama variety and Canadian variety are less successfully differentiated. Further research is needed on grain geometric parameters and their ability to identify wheat kernels.

References

1. Charytanowicz M, Kulczycki P (2008) Nonparametric Regression for Analyzing Correlation between Medical Parameters. In: Pietka E, Kawa J (eds) Advances in Soft Computing - Information Technologies in Biomedicine. Springer-Verlag Berlin Heidelberg
2. Fukunaga K, Hostetler LD (1975) The estimation of the gradient of a density function, with applications in Pattern Recognition. IEEE Transactions on Information Theory 21:32–40
3. Kowalski P, Łukasik S, Charytanowicz M, Kulczycki P (2008) Data-Driven Fuzzy Modeling and Control with Kernel Density Based Clustering Technique. Polish Journal of Environmental Studies 17:83–87
4. Krzyśko M, Wołyński W, Górecki T, Skorzybut M (2008) Systemy uczace sie. WNT, Warszawa
5. Kulczycki P (2005) Estymatory jadrowe w analizie systemowej. WNT, Warszawa
6. Kulczycki P (2007) Estymatory jadrowe w badaniach systemowych. In: Kulczycki P, Hryniewicz O, Kacprzyk J (eds) Techniki informacyjne w badaniach systemowych. WNT, Warszawa
7. Kulczycki P (2008) Kernel estimators in industrial applications. In: Prasad B (ed) Soft Computing Applications in Industry. Springer-Verlag, Berlin
8. Kulczycki P, Charytanowicz M (2005) Bayes Sharpening of Imprecise Information. International Journal of Applied Mathematics and Computer Science 15:393–404
9. Kulczycki P, Charytanowicz M (2010) A Complete Gradient Clustering Algorithm Formed with Kernel Estimators. International Journal of Applied Mathematics and Computer Science, in press
10. Kulczycki P, Daniel K (2009) Metoda wspomagania strategii marketingowej operatora telefonii komórkowej. Przeglad Statystyczny 56:116-134
11. Mirkin B (2005) Clustering for Data Mining: A Data Recovery Approach. Chapman and Hall/CRC, London
12. Niewczas J, Woźniak W (1999) Application of "GRAINS" program for characterisation of X-ray images of wheat grains at different moisture content. Xth Seminar "Properties of Water in Foods". Warsaw Agricultural University, Department of Food Engineering
13. Niewczas J, Woźniak W, Guc A (1995) Attempt to application of image processing to evaluation of changes in internal structure of wheat grain. International Agrophysics 9:343–347.
14. Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London
15. Shouche SP, Rastogi R, Bhagwat SG, Sainis JK (2001) Shape analysis of grain of Indian wheat varieties. Computers and Electronics in Agriculture 33:55–76
16. Strumiłło A, Niewczas J, Szczypiński P, Makowski P, Woźniak W (1999) Computer system for analysis of X-ray imges of wheat grains. International Agrophysics 13:133-140
17. Utku H, Koksel H, Kayhan S (1998) Classification of wheat grains by digital image analysis using statistical filters. Euphytica 100:171–178