ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Arkadiusz Maciuk**

MULTIDIMENSIONAL DATA CLASSIFICATION – COMPARISON OF *ISODATA* AND APPROXIMATION BY POINTS METHODS

Abstract. The effect of division is dependent not only on the criteria of division but also on the chosen method. The standard algorithm of multidimensional data classification ISODATA divides the given set into an assumed number of separable subsets in such a way that the division fulfills best the accepted criteria. An alternative method is approximation by the chosen number of points which in result indicates the areas of a set with large congestion of the elements. The paper compares effects of using both methods listing their advantages and drawbacks. Apart from presenting the results of division of various sets some characteristics of classification are discussed which are an effect of the choice of one of the above mentioned methods.

Key words: classification, division method, multivalent data.

I. THE *ISODATA* METHOD

The set of parameterized data of many features can be identified with subset of multidimensional Banach space. Norm of this space defines distance which can be the basis of classifications of these data. The *ISODATA* method and the method of approximation by points are two nonheurestic methods of multidimensional data classification which use the notion of distance. Let (X, w) be set X contained in Banach space E with measure w. Lets assume that measure w for set X is finite and enumerable. If symbol $\|\cdot\|$ stands for norm, then $\|x-p\|$ is distance between x and p.

What is now a classic method – *ISODATA* divides primal set of elements into k separable subsets in the way that the sum of measures of dispersion of these subsets is minimal. The aim of the procedure, having given k number of subsets and parameter $q \geq 1$, is finding the minimum of function $\min_{\{X_1,\ldots,X_k\}} \sum_{j=1}^{k} \sum_{x \in X_j} \left\| \mathbf{x} - \mathbf{p}_j \right\|^q w(\mathbf{x})$, or function:

* Ph.D., Department of Mathematics and Cybernetics, University of Economy, Wroclaw.

$$\min_{\{X_1,\ldots,X_k\}} \sum_{j=1}^{k} \left(\frac{1}{w(X_j)} \sum_{x \in X_j} \left\| x - p_j \right\|^q w(x) \right)^{1/q}, \qquad (1)$$

where the symbol $w(x)$ stands for the measure of element x belonging to X, $w(X_j)$ stands for the measure and p_j – the representative of subset X_j. The aim of using the *ISODATA* method is therefore finding among all the possible divisions of set X into k parts such division where the sum of subsets dispersion is the smallest. In the simplest case, when the norm of space E is the Euclides norm, and the parameter $q = 2$, representative of subset X_j is an arithmetic mean and its measure of dispersion is the square root of the (sampling) variance. This procedure was first created by Ball and Hall (1967). Up till now many changes and versions of this method were formulated. For example Jajuga (1987) broadened *ISODATA* method to Mahalanobis norm $\|x\| = x^T M x$, where M is positive definite matrix, and called such classification the spherical classification.

The *ISODATA* method verifies an assumption that X is k-connected set that is a set consisting of k separable subsets. To be more precise set X with measure w is a k–connected set if there is a covering of set X by k separable spheres which includes almost all the elements of set X in relation to measure w, the total field of which is smaller then any covering of this set by $k-1$ spheres. The calculation procedure of this method consists of the initial step, and two steps repeated iterately, until the moment, in which next iteration does not alter calculated values. The initial step is choosing the k different points $P_0 = \{p_{1,0}, \ldots, p_{k,0}\}$ from space E, called further on "the initial procedure points". They can be, for example, the points of set X chosen at random. First step is the division of set X $\mathcal{X}_0 = \{Xp_{1,0}, \ldots, Xp_{k,0}\}$ in this way, that every element $x \in X$ is assigned to set with index j when and only when $\left\| \mathbf{x} - \mathbf{p}_{j,0} \right\| = \min_s \left\| \mathbf{x} - \mathbf{p}_{s,0} \right\|$, where $s \in \{1,\ldots,k\}$. In the second step the representative for each of so formed subsets is chosen, that is the point p_j is found which realizes the minimum of function $f_{(X_j, w)}(p) = \sum_{\mathbf{x} \in \mathbf{X}_j} \left\| \mathbf{x} - \mathbf{p} \right\|^q w(\mathbf{x})$. A set of such representatives is the set $P_1 = \{p_{1,1}, \ldots, p_{k,1}\}$. The choice of set P_0 determines the division of set X into $\mathcal{X}_0$ and division $\mathcal{X}_0$ determines the representative set P_1. Next P_1 determines $\mathcal{X}_1$, and so on. The procedure is continued, iterately marking $(\mathcal{X}_{k-1}, P_k)$, until the moment when two consecutive divisions are the same.

II. APPROXIMATION BY POINTS METHOD

The method of approximation by points is based on an assumption that set X is generated by k different points or in other words that set X is k - modal. The result of set approximation by a point can be interpreted as indication of a modal of the monomodal set, and the effect of approximation set by k points as an indication of modal values of k- modal set.

The verification of a thesis, that $X = \{p\}$ can be conducted as follows. The equation of point p is the equation realized only by one point $x = p$. $\|x - p\| = 0$ is such an equation. The measure of non fulfillment of the condition that the set X is equal to point p, is function $\Delta_{(X,w)}(p) = c\sum_{x\in X}\|\mathbf{x}-\mathbf{p}\|^q w(\mathbf{x})$, where $w(x)$ is measure of x, $c > 0$ and $q \geq 1$ – given parameters. The value of function is non-negative and equal to zero when and only when $X = \{p\}$ in relation to measure w. The more X is "different" from set $\{p\}$, the larger the value of this function. Also, the point realizing minimum of this function in relation to p is "the best representative" (generator) (X, w), or in other words, the point approximating (X, w).

The measure of a postulate, that set $X = \{p_1, \ldots, p_k\}$ is the non-negative function of zero value when and only when $X \subseteq \{p_1, \ldots, p_k\}$. It can be any function:

$$\Delta_{(X,w)}(p_1, \ldots, p_k) = c\sum_{x\in X}\|x - p_1\|^{q_1} \ldots \|x - p_k\|^{q_k} w(x), \qquad (2)$$

where $q_1, \ldots, q_k \geq 1$ given parameters, $c > 0$ normalizing constant. The approximation of (X, w) by k points means finding the set of k points $\{p_1, \ldots, p_k\}$ which realizes minimum of this function. The minimum of function (2) can be found by using method of iterate modification of weights, which indicates sequences of point sets convergent to this minimum. The expression "modification of weights" is derived from the fact that the representative of the same X is chosen iterately modifying each time the measure (the weight of points). Let $P_0 = \{p_{1,0}, p_{2,0}, \ldots, p_{k,0}\}$, where k is given, be any chosen subset of X called further on "the set of initial procedure points". Point $p_{1,1}$ is the point which realizes the minimum of function (2) in relation to variable p_1, with given $p_2 = p_{2,0}, \ldots, p_k = p_{k,0}$. Point $p_{2,1}$ is the point realizing the minimum of function (2) in relation to variable p_2 with given $p_1 = p_{1,1}, p_3 = p_{3,0}, \ldots, p_k = p_{k,0}$. Point $p_{3,1}$ is the point realizing the minimum of function (2) in relation to variable p_3 with given $p_1 = p_{1,1}$, $p_2 = p_{2,1}, p_4 = p_{4,0}$, and so on. All points of set P_0 are exchanged one by one receiving this way set P_1. P_3, P_4 and all consecutive sets are constructed in the

same way, until the location of points in two consecutive sets is practically the same.

As in the case of *ISODATA* method, norm of function given by formula (2) can be any, however the simplest case is Euclides norm and parameters $q_1 = \ldots = q_k = 2$. Then next approximation points will simply be arithmetic means of set $\boldsymbol{X}$, each time with different weights.

III. THE MATTER OF ESTABLISHING THE NUMBER OF CLASSES

In such classification methods the issue is a proper selection of parameter k that is establishing correctly number of subsets - the classes. The solution of this problem can be applying the procedures for $k = 1$, then for $k = 2$, for $k = 3$ and so on – and consequently the comparison of effects of procedure application. In the case of *ISODATA* method verification of correctness of selection of the number of classes is equivalent to establishing the sum of dispersion measures of individual division subsets. That is establishing value of function (1) for the division received as a result of using the procedure. If for k subsets this value is smaller than for $k - 1$, then the thesis about existence of k different classes is more credible than the thesis about existence of only $k - 1$ classes.

In the case of the approximation by points method, this verification is more complicated: dispersion measure is calculated for each approximating point separately. Let set $\boldsymbol{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\}$ be the result of procedure application. The correctness of approximation by point $\boldsymbol{p}_1$, with given points $\boldsymbol{p}_2, \ldots, \boldsymbol{p}_k$ can be analyzed identifying $\|\boldsymbol{x} - \boldsymbol{p}_2\|^{q_2} \ldots \|\boldsymbol{x} - \boldsymbol{p}_k\|^{q_k} w(\boldsymbol{x})$ with $w_1(\boldsymbol{x})$. Value $s^{q_1}(\boldsymbol{X}, w) = \frac{1}{c_1} \sum_{x \in X} \| \boldsymbol{x} - \boldsymbol{p}_1 \|^{q_1} w_1(\boldsymbol{x})$, where $c_1 = \sum_{x \in X} w_1(\boldsymbol{x})$ is partial dispersion measure for point $\boldsymbol{p}_1$ taking into account the position of points $\boldsymbol{p}_2, \ldots, \boldsymbol{p}_k$. In the similar way partial measures of dispersion can be calculated for any point $\boldsymbol{p}_j$, where $j \in \{1, \ldots, k\}$. What we receive this way is k of indicators, and each of them is a measure of $(\boldsymbol{X}, w_j)$ dispersion, where w_j is a modified measure w. Figure 1 and 2 contain comparison of results of both methods and give dispersion measures in the simplest case: when norms is Euclides norm and parameters $q_1 = \ldots = q_k = 2$. Total (sum) variance of all subsets in the case of *ISODATA* method (left column) and partial variance of approximating points (right column).

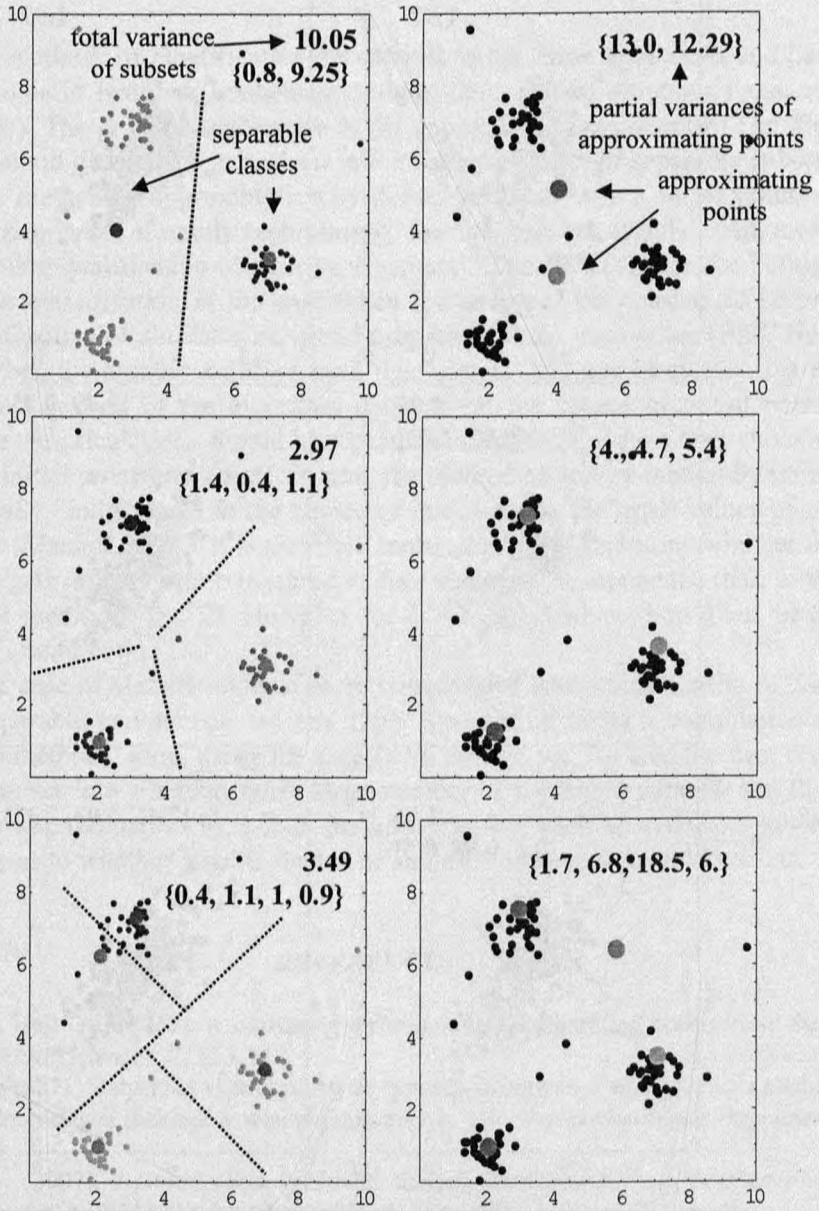

Figure 1. Comparison of effects of classification by *ISODATA* method (left column) and method of approximation by points (right column) for $k = 2$, 3 and 4

Source: own elaboration.

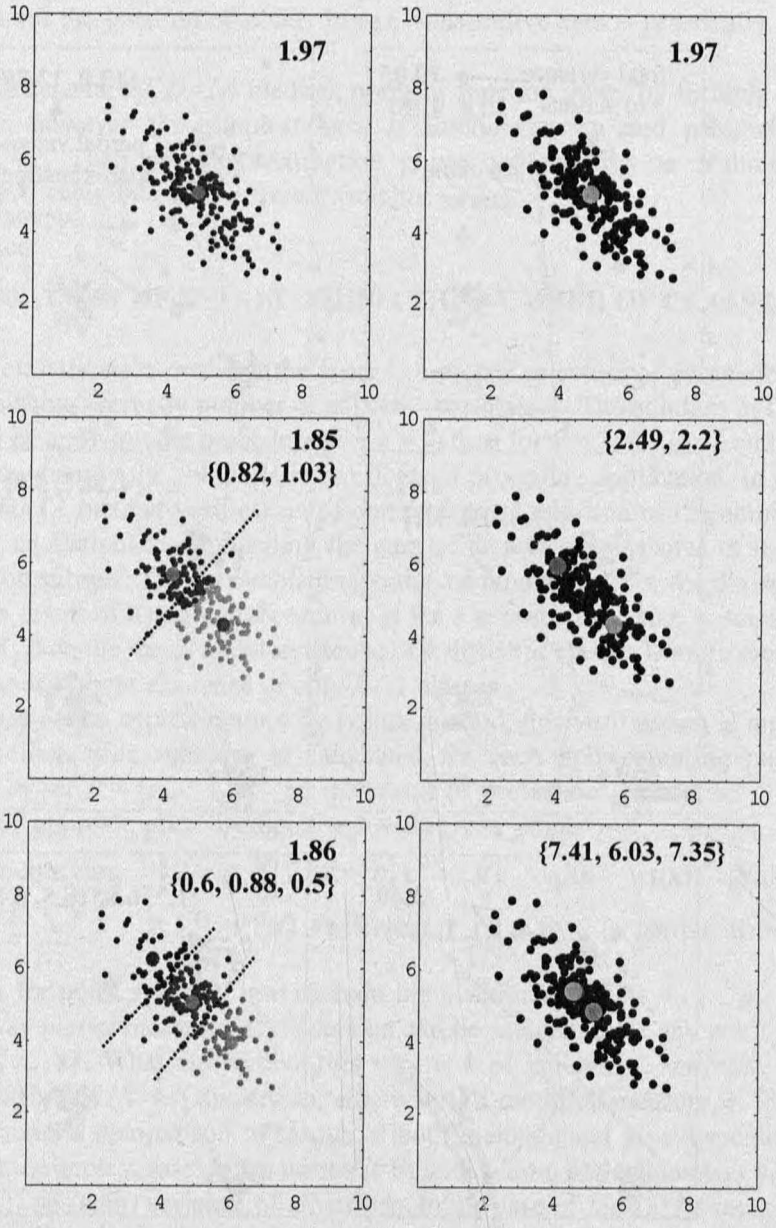

Figure 2. Comparison of effects of classification by *ISODATA* method (left column) and method of approximation by points (right column) for $k = 1$, 2 and 3 – questionable example of classification

Source: own elaboration.

IV. FEATURES OF THE METHODS

Both methods of classification are subject to the same limitations and have similar numeric features (nonheurestic algorithms, similar amount of numeric operations). The principal difference is the approach to classification. The *ISODATA* method divides the primal set into a settled number of separable subsets, while the method of approximation by points indicates only a set of points of a certain size "most correctly representing" the analyzed set, together with measures enabling qualification of their "correctness". The *ISODATA* method allows for proper classification in the case when the analyzed set consists of several subsets of points of similar size, clearly separated from each other (Ball, Hall, 1967). When the number k differs from the "correct" number of classes, the effects of application of the algorithm depends on the choice of initial points. Therefore the calculations should be repeated several times – each time choosing different initial procedure points. In turn, the method of approximation by points is practically "indifferent" to the choice of initial points for small values of parameter k (Maciuk, 2007). It is therefore more suitable to determine whether the given set gets divided into two, three or four separable components, than in the *ISODATA* method (Figs. 2). However for $k > 4$ the latter method gives better results of classification.

In the case of classification of more complicated sets which consist of dozens of separable components we can apply the method being a compilation of these two methods. First, using the *ISODATA* method we can find the best division of the set into a settled fairly large number of separable subsets, and than using the approximation by points method examine each so received subsets separately as to whether it splits further on into two or three separable subsets.

REFERENCES

Ball G.H., Hall D.J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science*, 12, 153-165.

Jajuga K. (1987). Statystyka ekonomicznych zjawisk złożonych – wykrywanie i analiza niejednorodnych rozkładów wielowymiarowych. *Prace naukowe AE we Wrocławiu,* nr 371.

Maciuk A. (2007). Aproksymacja punktami metodą modyfikacji wag. *Mathematical Economics*, nr 4(11). Wydawnictwo AE we Wrocławiu.

Arkadiusz Maciuk

KLASYFIKACJA WIELOWYMIAROWYCH DANYCH – PORÓWNANIE METODY *ISODATA* I METODY APROKSYMACJI PUNKTAMI

Efekt podziału zależy nie tylko od ustalenia kryteriów podziału, ale także od wyboru metody dzielenia. Standardowy algorytm klasyfikacji wielowymiarowych danych *ISODATA* dzieli wyjściowy zbiór na ustaloną liczbę rozłącznych podzbiorów tak, aby podział ten jak najkorzystniej spełniał przyjęte kryteria. Alternatywą wobec niego jest algorytm oparty na metodzie aproksymacji ustaloną liczbą punktów, którego efektem jest wskazanie obszarów zbioru o dużym stopniu zagęszczenia elementów. Artykuł zawiera porównanie efektów użycia tych metod ze wskazaniem zalet i wad. Omawia też pewne własności klasyfikacji wynikające z konsekwencji wyboru jednej z dwóch omawianych metod.