Clustering by Mode Estimation

E. Ocelíková* and D. Klimešová**

* Technical University of Košice/ Department of Cybernetics and Artificial Intelligence, Košice, Slovak Republic
** Czech University of Agriculture, Prague, Faculty of Economics and Management
Department of Information Engineering, Prague 6 – Suchdol, Czech Republic
Eva.Ocelikova@tuke.sk, Klimesova@pef.czu.cz

***Abstract*—In this contribution we introduce a clustering scheme based on mode boundary detection procedures. Modes are characterized as compact regions of the data space with higher densities than their surrounding. A mode boundary as defined in this approach is an area of large local changes in the probability density functions. Examples of the performance of the clustering based on the so-obtained mode boundaries are given using artificially generated data sets.**

I. INTRODUCTION

Set of clustered objects can take as selection from basic statistic set of objects from mathematical statistics point of view which is represented multidimensional random variable. From this point of view probability density function of this set has significance. Many methods of cluster analysis have probability approaches to cluster creation. Each mode of this function corresponding to a one cluster [1][6].

Modes are characterized as compact regions of the data space with higher densities than their surrounding. When mode is characterized as local extreme of the underlying probability density function, it may be detected by hill-climbing procedures using some gradient search technique [2]. A mode boundary is defined as area of large local changes in the probability density function.

For mode boundary detection are used generalized gradient operators. The output of these operators are then used to determine a level of confidence that each sampling point in the data space lies on a mode boundary. A strong response from these operators is interpreted as an indication of the presence of a boundary element [3][10].

II. CLUSTERING BY MODE BOUNDARY DETECTION

The clustering problem is stated as finding the boundaries which separate the modes from their surrounding. The basic procedure of mode boundary detection has three major steps which are described hereafter [11].

A. Values density function estimation

Let $\mathbf{X} = \{X^1, X^2, ..., X^n\}$ be a set of *d*-dimensional vectors $X^i = (x_1^i, x_2^i, ..., x_d^i)$, $i=1,2,...,n$ describing clustered objects. d is number of attributes characterizing objects and n is number of these clustered objects.

Initially a diagonal transformation is execute which stretches or shrinks the axes of the data space in order to standardize the range of variation of each attribute:

$$\max_i x_j^i - \min_i x_j^i = h_j, \quad j = 1,2,...,d \qquad (1)$$

Each axis of data space is then partitioned into equal intervals. This discretization defines a set of R hypercubes of side length unity, so-called "unite lattice". The centers of these hypercubes are designed as P_r, $r = 1,2,...,R$. The centers have integer coordinates and frequently be called *sampling points.*

After unite lattice application on tested objects we have new space division in which density function will be tested in single hypercubes. Density function values will be estimated in non-empty hypercubes only. The centers of these non-empty hypercubes whose number never exceed the object number n. Density function value in simple hypercube will be determined as amount of objects which inhere in it.

Result of this first step is discrete set of estimated value basic density function at a subset of R sampling points, i.e.

$$f(P_r), \quad r = 1,2,...,R, \quad kde \quad R \le n. \qquad (2)$$

Figure 1 shows the unite lattice with hypercubes where every hypercube has illustrated its density function value.

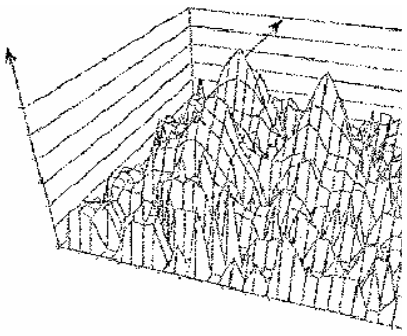

Figure 1. Density function values of hypercubes

B. Median filtering

For the mode boundary detection some preprocessing of hypercubes density function values is needed. It is necessary for enhance significant local variations of the density function and to reduce noise.

978-1-4244-3802-0/09/$25.00 ©2009 IEEE

The implementation of a generalized multidimensional median filter requires the definition of a neighborhood of hypercube P_r in the unit lattice.

Let $p_1^r, p_2^r, ..., p_j^r, ..., p_d^r$ are the integer coordinates of the hypercube P_r. The hypercubic neighborhood $V_m(P_r)$ of size $(2m+1)$, where m is radius of hypercubic surrounding (see Fig. 2), is defined by the set of sampling points $P_{r'}$ such as:

$$V_m(P_r) = \left\{ P_{r'} \mid p_j^r - m \le p_j^{r'} \le p_j^r + m,\ j = 1,2,...,d \right\} \quad (3)$$

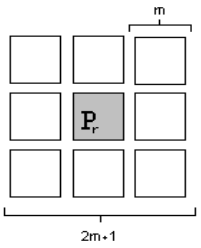

Figure 2. Surrounding of hypercube P_r, where m=1.

The filtered value of density function $f'(X)$ of $f(X)$ in hypercube P_r is then given as (4):

$$f'(P_r) = median\left[f(P_{r'}) \mid P_{r'} \in V_m(P_r) \right] \quad (4)$$

The surroundings of size $(2m+1)$, of each hypercube P_r is analyzed and hypercubic density function value of this surroundings will be include in to calculation of filtered density function value. This evaluate filtered density function value is the same for all hypercubes of surrounding $V_m(P_r)$.

Fig. 3 shows the change of density function after median filtering.

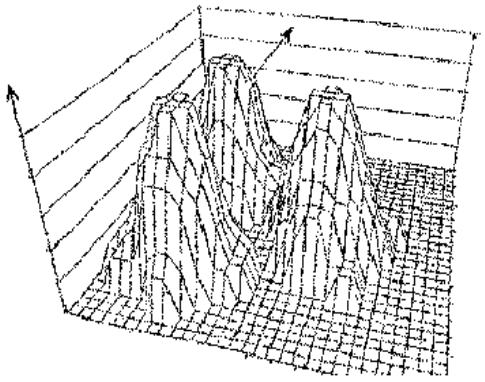

Figure 3. The change density function after median filtering

C. Boudary detection

Mode boundary detection is based on local operators that response to local amplitude changes and when applied to the filtered discrete estimation of the density function.

In this work we used Robert´s operator defined in d-dimensional space using (7):

$$\begin{aligned} G_{\alpha,\beta}(P_r) = &| f\left(p_1^r,...,p_\alpha^r,...,p_\beta^r,...,p_d^r\right) - \\ &- f\left(p_1^r,...,p_\alpha^r,...,p_\beta^r,...,p_d^r\right)| + \\ &+ | f\left(p_1^r,...,p_\alpha^r,...,p_\beta^r,...,p_d^r\right) - \\ &- f\left(p_1^r,...,p_\alpha^r,...,p_\beta^r,...,p_d^r\right)|, \end{aligned} \quad (5)$$

where $\alpha = 1,2,...,d$; $\beta = 1,2,...,d$.

The result of Robert ´s operator using is gradient $G(P_r)$. From now this gradient will be describe each hypercube instead of density function $f(P_r)$.The meaning of the gradient using instead of density function consists in fact, that gradient value of hypercube which is boundary element is high as compared to gradient value of the hypercubes which are situated inside or outside of cluster.

Furthermore, the effect of found boundary elements is higher when we use simple threshold of the computed gradients. Only hypercubes which gradient values is greater than e.g. 5% of the gradient maximum will be boundary elements.

III. RESULTS OF MODE BOUNDARY DETECTION METHOD

Boundary of clusters which have different profiles were marked sufficiently exactly by mode boundary detection method. The clusters contained large amount of objects, they have had different profiles. It was necessary for application of unite lattice all over data area. For the clusters which have size less than double size unit lattice hypercube, cluster boundary detection is not possible. The using of unite lattice with hypercubes of small sizes withheld satisfactory results too, because object density into the hypercubes on the cluster margin was little different as object density into hypercubes which were inside the cluster.

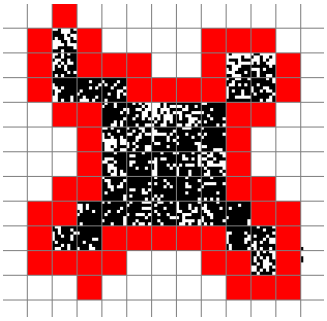

Figure 4. The boundary of the cluster with profile of intersect

The using of hypercubes with smaller sizes is suitable for area of large object density only. Figures 4 and 5 show results of mode boundary detection in 2D data space (by reason of graphic illustration). Figure 5 shows boundary of two clusters whereas one is nested into other.

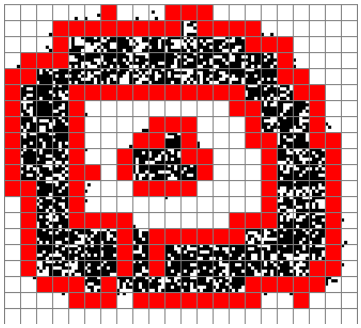

Figure 5. The detected boundaries of clusters nested one into the other

Different figures are selected purposely. Mode boundary detection methods enable to identify also such figure of clusters, which classical cluster methods determining of clusters on the basic of object distances, have problem.

ACKNOWLEDGMENT

This work is supported by the VEGA project No 1/0386/08 and the project information and knowledge support of strategic control - MSM 6046070904.

REFERENCES

[1] Devijver, P.A. - Kittler, J.: Pattern Recognition a Statistical Approach. Academic Press, London 1982

[2] Fukunaga, K.-Hosteller,L.D.: The Estimation of the Gradient of a Density Function with Applications in Pattern Recognition. IEEE Trans. Inform. Theory. IT-21, 32-40, 1975

[3] Herman, G.T.- Liu, H.K.: Dynamic Boundary Surface Detection by Hypersurface Fitting. IEEE Trans. Pattern Analysis Mach. Intell. PAMI-3, 482-486, 1981

[4] Kittler, J.: A Local Sensitive Method for Cluster Analysis. Pattern Recognition 8, 23-33, 1976

[5] Klimešová D- Ocelíková, E**.:** Study of Uncertainty and Contextual Modelling. *International Journal of Circuits, Systems and Signal Processing,* Issue 1, Volume 1, 2007, pp. 12-15

[6] Lukasová, A.-Šarmanová, J.: Methods of cluster analysis. SNTL, Praque, 1985

[7] Mizoguchi, R.-Shimura,M.: Non parametric learning without the teacher based on mode estimation, IEEE Trans. Comput. C25, 1976, pp.1109-1117

[8] Ocelíková,E :Multidimensional Data Recognition and Classification with Use Information Technology. In: *Proc. of the Faculty of Electrical Engineering and Informatics Research and Development Projects,* 2006, Košice, Slovakia. pp.29 – 30

[9] Ocelíková, E.-Zolotová, I.: Pattern Recognition on the Basis Intelligent and Information Technologies. In: Proc. of II. Int. Scientific Conference of the Faculty of Electrical Engineering and Informatics. May 15, 2001 Košice, pp.31-32

[10] Postaire, J.G. – Touzani, A.: Mode Boundary Detection by Relaxation for Cluster Analysis. Pattern Recognition, Vol.22, No.5, pp. 477-489, 1989

[11] Postaire, J.G.- Vasseuer,C.P.A.: A Fast Algorithm for non parametric Probability Density-estimation. IEEE Trans. Pattern Analysis Mach. Intell. PAMI-4, 663-666, 1982