Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp406-410)

The Result Analysis of the Cluster Methods by the Classification of Municipalities

PAVEL PETR, KAŠPAROVÁ MILOSLAVA
System Engineering and Informatics Institute
Faculty of Economics and Administration
University of Pardubice
Studentská 95, 532 10 Pardubice
Czech Republic
miloslava.kasparova@upce.cz
http://www.upce.cz

Abstract: The goal of this paper is the result analysis of chosen cluster methods by the classification of municipalities. A nonhierarchical cluster method with a variable number of clusters is used for the determination of a convenient number of clusters. A Kohonen map is chosen as an alternative approach to standard cluster methods. Nonhierarchical cluster methods with the fixed number of clusters are used for the classification of municipalities, too. The results of methods are analyzed. The classification of municipalities is the starting basis for the creation of a prediction model (PM). The PM is determined for the employees of the Regional Authority to support their decision making by financing of the subsidized bus connection lines.

Key – words: cluster analysis, cluster method, neural nets, clustering, municipalities

1 Introduction

It was possible to do the municipality classification for a PM creation. The PM is meant for the employees of the Regional Authority to support their decision making by the financing of the subsidized bus connection lines. The municipalities' classification in to clusters will result in a creation of municipality input chains which are in bus line connections. Every municipality that is in the bus line connection will be represented by the number of clusters. The classification is performed by chosen methods of cluster analysis [2, 3, 4].

Cluster analysis groups data objects into clusters in such a way that the objects belonging to the same cluster are similar, while those belonging to different ones are dissimilar.

An existence of n objects is an initial condition for the usage of the cluster analysis. The n municipalities where the object j is the municipality j, are the objects of the clustering. Every object is described by p characteristics. A vector of measurement $\underline{O_j}$ that contains p characteristics in formula (1):

$$\underline{O_j} = \{z_{1j}, z_{2j}, \ldots, z_{pj}\} \tag{1}$$

for j-th object O_j, $(j = 1, 2, \ldots, n)$. The input set of the objects which are determined for the clustering, it is possible to write in formula of objects matrix $\boldsymbol{O}$. The general formula of objects matrix $\boldsymbol{O}$ is following (2):

$$\boldsymbol{O} = \begin{bmatrix} z_{11} & z_{12} & \ldots & z_{1p} \\ z_{21} & z_{22} & \ldots & z_{2p} \\ z_{j1} & z_{j2} & \ldots & z_{jp} \\ z_{n1} & z_{n2} & \ldots & z_{np} \end{bmatrix}, \tag{2}$$

where : n – object characteristics number;

p – number of characteristics.

The task of clustering is then to divide the set of objects into the disjunctive clusters.

The decision making about the object clustering in cluster is realized on the basis of the similarity by application of metric [3, 4, 5]. A sum of the square errors to centre of clusters E [4] is chosen as a criterion of the quality of clustering. It is defined in this way:

Let $\Omega = \{M_1, M_2, \ldots, M_k\}$ is the clustering of objects set in k clusters $M_1 = \{O_{11}, O_{12}, \ldots, O_{1n_1}\}$, $M_2 = \{O_{21}, O_{22}, \ldots, O_{2n_2}\}, \ldots, M_k = \{O_{k1}, O_{k2}, \ldots, O_{kn_k}\}$,

where O_{hj} is object j of h-th cluster M_h. Then E is determined in formula (3):

$$E = \sum_{h=1}^{k} \sum_{j=1}^{n_h} d^2\left(O_{hj}, T_h\right), \qquad (3)$$

where: $d^2\left(O_{hj}, T_h\right)$ - the square of the Euclidian metric of object O_{hj} to the centre T_h of cluster M_h.

T_h - the centre of cluster M_h; it is determined by the vector of mean values of characteristics i of objects in cluster M_h in formula $T_h = (t_{h1}, t_{h2}, \ldots, t_{hp})$, for its characteristics i, where $i = 1, 2, \ldots, p$, is (4):

$$t_{hi} = \frac{1}{n_h} \sum_{j=1}^{n_h} z_{hji}, \qquad (4)$$

where: n_h - number of objects in cluster M_h;
z_{hji} - characteristic i of object j in cluster M_h.

A set of typical point is defined for the realization of cluster methods. The typical point characterizes the typical object – municipality. It comes out from p-dimensional Euclidian space E_p. It is more approaches for the determination of initial typical points, for example [4]:

- the selection of first s points from randomly organized set of points;
- marking of points by serial numbers *1, 2, ... ,n* and a generation randomly s serial numbers from set *{1, 2, ... , n}*, points which belong to them then are initial typical points;
- the selection s points which are chosen by an analyst.

2 Problem Definition

The goal is to classify municipalities on the basis of the same or similar values of their characteristics. The classification of municipalities is the starting basis for PM [1]. It is realized by values of the given characteristics using the cluster analysis methods.

Characteristics of municipalities are following:

Information – land registers	Population	Occurrence of facilities	Technical facilities of the municipality	Other
KO, PCO, PK, KV	*POC, PMO, PZO, POV15-64, PVOC, PVZ, PVM*	*SKO, POS, POL, ZDZ*	*PLY, VOD, KAN*	*M-O*

The metric of input data was created firstly. It contents 452 municipalities of the chosen region. The municipalities were identified by the municipality code *(KO)* and characterized by the number of municipality parts *(PCO),* the number of land registers *(PK),* the local area in hectares *(KV),* the number of population in municipalities *(POC),* the number of men *(PMO)* and women *(PZO)* that live in the municipality, the number of population at the age 15 to 64 (*POV15-64)*, the average age of population in municipalities (*PVOC*), the average age of women *(PVZ)* and men *(PVM)*, the occurrence of: schools *(SKO),* post offices *(POS),* the police *(POL*), medical facilities *(ZDZ);* technical facilities of municipalities – the gas *(PLY),* the duct *(VOD),* the sewerage *(KAN)* and the determination of municipalities or a village *(M-O)* according to [6].

The reason of municipalities' classification is:

- a generalization of municipalities' characteristics;
- a creation of the simplified input chain configuration.

3 Cluster Analysis Methods

The municipalities' classification was realized in two steps. Firstly it was necessary to determine a convenient number of clusters. There were used a nonhierarchical cluster method [7] with a variable number of clusters and a Kohonen map [8, 9, 10].

Secondly nonhierarchical cluster methods with a fixed number of clusters were used: the Forgy method [4, 11], the K-Means method [2, 3] and the Jancey method [4, 12]. The classification of municipalities in the fixed number of clusters k was realized by these methods. The parameter E in formula (3) was a criterion of the quality of the clustering.

3.1 Nonhierarchical Cluster Method

For the solution of an algorithm are determined the following input parameters:

- the set of the input objects, which is the matrix of municipalities $\boldsymbol{O}$ in formula (2);
- the matrix of the typical points *T(s,p),* where s is the number of inputs typical points *(s = 15)* and p is the number of the municipality (the approach the selection of the first s points was chosen for the creation of the typical point matrix);
- the value of a mergence threshold and dividing of clusters *H,* where *H = 0.6*;
- maximal number of iterations *I* where *I = 2n;*
- minimal number of objects in cluster *MP* where *MP* $\in$ *{1, 2,..., 15}*.

The iterations of algorithm consist of steps. The algorithm is described in [7].

3.1.1 Results of the Nonhierarchical Cluster Method

The 69 clusters were created on the basis of the parameter *MP*, for *MP* = 1. The resulting parameter *E* was *E* = 58.3. The objects were classified in to 60 clusters by the parameter *MP* = 2 and with *E* = 60.12. The lowest *k* = 16 was extracted by *MP* = 15 and *E* = 79.88. By the increase of the parameter *MP* the number of the created clusters *k* is decreased. A graphical representation of the dependence of *E* and *k* on *MP* is in Fig. 1.

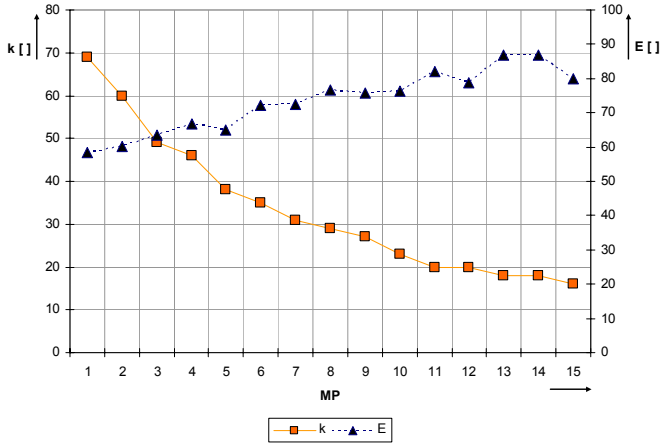

Fig. 1 The graphical representation of the dependence of *E* and *k* on *MP*

The increase of the parameter *PM* in the cluster results in the increase of the criterion *E*. The low values of criterion E were achieved by the creation of the big number of clusters $38 \leq k \leq 69$. The stabilization of cluster number was done by the value of the parameter $11 \leq MP \leq 15$. The values *E* = 82.14 and *k* = 20 were for *MP* = 11, the values *E* = 79.88 and *k* = 16 was for *MP* = 15, more in the Table 1.

Table 1: Number of the clusters k and the parameter E by the dependence on the parameter MP

MP	11	12	13	14	15
E	82.14	78.88	86.84	86.84	79.88
k	20	20	18	18	16

3.2. Kohonen Map

The Kohonen map (self-organizing map) was used as an alternative cluster method. The Kohonen map [8, 9, 10] is a special kind of neural network that performs unsupervised learning. It takes the input vectors in formula (1) for *j = 1, 2, ... n* and performs a type of spatially organized clustering, or feature mapping, to group similar records together and collapse the input space to a two-dimensional space that approximates the multidimensional proximity relationships between the clusters. The Kohonen map consists of two layers of neurons or units: an input layer and an output layer. The input layer is fully connected to the output layer, and each connection has an associated weight. Another way to think of the network structure is to think of each output layer unit having an associated center, represented as a vector of inputs to which it most strongly responds (where each element of the center vector is a weight from the output unit to the corresponding input unit) [8].

The parameters of the Kohonen map are represented as weights between input units and output units, or alternately, as a cluster center associated with each output unit, more for example in [3, 8].

3. 2. 1 Results of the Kohonen Map

The clustering differed in dimension and time of neural net training in interval from 1 to 10 minutes.

The clustering of objects in a big number of clusters by different dimensions of topological grid of the Kohonen map results from the achieved results. The lowest number of clusters k_{min} = 18 was achieved by the map dimension 4 x 5 and the time of neural net training 2 minutes. The biggest number of clusters k_{max} = 56 was obtained by the map dimension 8 x 8.

3. 3 Result Comparison

The goal of the methods was to determine the appropriate number of clusters of municipalities. It was achieved the appropriate number of clusters k_{opt} by the big value of the criterion *E* by the nonhierarchical cluster method.

The number of clusters *k* rises by the dimensions increase of the Kohonen map. According to a visualization of resulting clusters of objects by different dimensions it is possible to find out hypothetical clusters of objects. An example of representation of clusters is in Fig. 2.

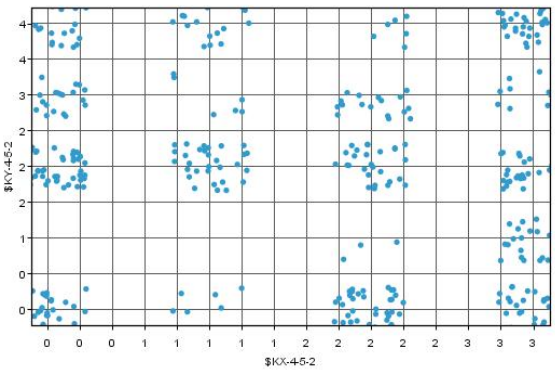

Fig. 2 The graphical representation of the supposed number of clusters from the Kohonen map

3. 4 Nonhierarchical Cluster Methods with Fixed Number of Clusters

The nonhierarchical cluster methods by the fixed number of clusters are used in the next step for municipalities' classification. There are: the K-Means method, the Forgy method and the Jancey method.

The input matrix of objects $\boldsymbol{O}$ in formula (2) is loaded at the start of the clustering by these methods. The clustering is realized for k = 15, 16, 17, 18, 19, 20.

3. 4. 1 K-Means Method

The *k*-means method [3, 4, 8] is a clustering method, used to group records based on similarity of values for a set of input fields. The basic idea is to try to discover k clusters in such a way that the records within each cluster are similar to each other and distinct from records in other clusters. *K*-means is an iterative algorithm; an initial set of clusters is defined, and the clusters are repeatedly updated until no more improvement is possible (or the number of iterations exceeds a specified limit) [8]. A cluster algorithm is described for example in [8].The criterion of the quality of clustering is the parameter E in formula (3).

Results of the K-Means Method

The values of parameter E for every k are in the Table 2.

Table 2: Results by the K-Means method

The number of clusters k	15	16	17	18	19	20
The general sum of the squares errors E	134.30	133.41	131.64	122.18	121.14	112.80

3. 4. 2 Forgy Method

The method [11] is based on the alternation of two steps. The first step is a calculation of the typical points of clusters of objects. The second step is the creation of clusters of objects by the clustering of every object to the typical point to which the object is nearest. The iterations are realized till a stable dividing in the final clusters is created, more in [4, 11]. The criterion of the quality of clustering is the parameter E in formula (3).

Results of the Forgy Method

The values of parameter E for every k are in the Table3.

Table 3: Results by the Forgy method

The number of clusters k	15	16	17	18	19	20
The general sum of the squares errors E	128.84	130.68	117.61	116.00	116.63	113.03

3. 4. 3 Jancey Method

The Jancey method [4, 12] is almost the same as the Forgy method. It differs in the way of the calculation of new typical points of the created clusters [4]. The criterion of the quality of clustering is the parameter E in formula (3).

Results of the Jancey Method

The values of parameter E for every k are in the Table 4.

Table 4: Results by the Jancey method

The number of clusters k	15	16	17	18	19	20
The general sum of the squares errors E	122.57	126.22	118.51	112.30	112.22	111.23

3. 5 Result Comparison

It was achieved the lowest parameter E = 112.80 by number of clusters k = 20 by the K-Means method. The considerable decrease of parameter E = 122.18 was by k = 18.

It was achieved the lowest parameter E = 113.03 by k = 20 by the Forgy method. The considerable decrease of parameter E = 117.61 was by k = 17.

By the Jancey method almost the same results of parameter E by k = 18 and k = 19 were achieved, very similar to k = 20, too.

The graphical representation of parameter E for every k and the methods: Jancey, Forgy and K-Means is in Fig. 3.

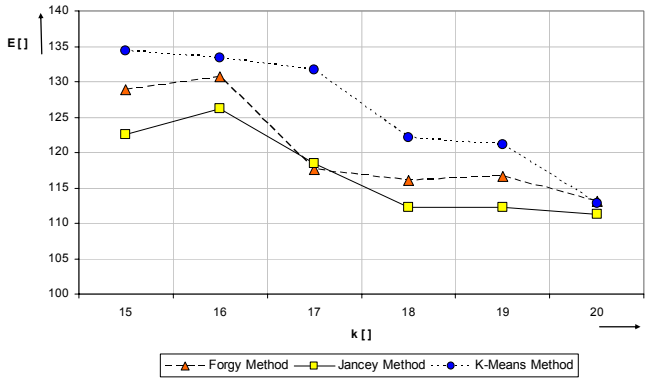

Fig. 3 The graphical representation of the dependence of *E* on every *k*

4. Conclusions

By results in Fig. 3 the clustering of objects in 18 clusters is possible. This number of clusters is determined on the basis of the values that are achieved by the Jancey method. The results of clustering by this method with the fixed number of clusters for $k = 18$ are in Table 5.

Table 5: Results of the clustering by methods: Forgy, Jancey and K-Means

The method	Forgy	Jancey	K-Means
The number of clusters *k*	18	18	18
The general sum of square deviations *E*	116,00	112,30	122,18

On the basis of the clustering of *n* municipalities in *k* clusters it is possible to create a model of classification of municipalities by means of decision trees [5, 13]. According to the rules from the decision trees it will be possible to classify new municipalities automatically on the basis of characteristics values in the cluster for next processing in PM.

References:

[1] Kašparová, M. Prediction Models Analysis of Financing of Basic Transport Services. *WSEAS Transactions on Systems*, January 2006, vol. 5, no. 1, s. 211 - 218. ISSN: 1109-2777.

[2] Freitas, A. A. *Data Mining and Knowledge Discovery with Evolutionary Algorithms,* Berlin: Springer, 2002.

[3] Guidici, P. *Applied Data Mining: Statistical Methods for Business and Industry,* West Sussex: Wiley, 2003.

[4] Lukasová, A. - Šarmanová, J. *Metody shlukové analýzy,* SNTL Nakladatelství technické literatury v Praze, 1985.

[5] Han, J. - Kamber, M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Press, 2001.

[6] *Zákon č. 129/2000 Sb., o krajích (krajské zřízení).*

[7] Petr, P. *Terciální zpracování radiolokační informace v heterogenní síti čidel.* [dissertation], VVTŠ, Liptovský Mikuláš, 1987.

[8] SPSS Inc. *Clementine® 7.0 User's Guide*, 2002.

[9] Berry, M. J. A. - Linoff, G. S. *Data Mining Techniques: For Marketing, Sales, and Customers Relationship Management,* Wiley, 2004.

[10] Pyle, D., *Business Modeling and Data Mining*, Morgan Kaufmann Publishers, 2003.

[11] Forgy, E. W., Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications. In: Biometrics Soc. Meetings, Riverside, 1965.

[12] Jancey, R. C. Multidimensional Group Analysis, Austral J. Botany, 14, 1966.

[13] Rusell, S. J. - Norvig, P., *Artificial Intelligence: A Modern Approach,* Prentice Hall, 2002.