Contribution to the Classification of Multidimensional Data

Eva OCELÍKOVÁ[1] , Dana KLIMEŠOVÁ[2]

[1]Technical University of Košice, Department of Cybernetics and Artificial Intelligence, 041 20 Košice,
Letná 9/B, The Slovak Republic e-mail: Eva.Ocelikova@.tuke.sk
[2] Dept. of Image Processing, Institute of Information Theory and Automation, Czech Academy of Sciences,
Pod vodarenskou věží 4, 182 08 Praha 8, Czech Republic klimes@utia.cas.cz
Dept. of Information Engineering, Faculty of Economics and Management, Czech University of Agriculture, Kamycka 129, 165 21 Prague 6 - Suchdol, Czech Republic e-mail: klimesova@pef.czu.cz

***Abstract**: The contribution deals with a problem of creation of multidimensional data classes by cluster analysis. It presents methods of hierarchical and non-hierarchical clustering to create the clusters of similar objects. It describes experience and results obtained during the clustering of multidimensional data which described real status of respondents from a coronary heart disease point of view.*

***Keywords**: Multicriterial classification, cluster analysis, similarity coefficients, clusters*

1 Introduction

Decomposition of an input set of objects into definite system of classes is the result of a classification process. This paper deals with the methods of cluster analysis which are one of many applications of the classification process.

Cluster analysis is defined [1, 3] as a general logical approach formulated as a procedure that groups the objects into classes - clusters. Clustering is based on the similarity and the unsimilarity of the objects. Cluster analysis achieves correct results mainly in situations where the researched set of objects separates into the classes really; e.g. the objets have tendency to group into natural classes. Creation of classes causes a radical dimension reduction of work because one characteristic substitutes the subset of objects. This characteristic expresses the competence into the class defined this way. Cluster analysis formalises and generalises well-known procedure used at the multicriterial classification in daily life.

2 Basis Problems of Cluster analysis

One of the basic problems of cluster analysis is the conception of the mutual similarity among the objects and its quantitative expression. The commonest way to express this similarity is through the metrics.

Because the objects are characterised by *m*-dimensional vectors that are the points of *m*-dimensional Euclidean space, it is possible to use Euclidean distance as a measure of the similarity. The shorter distance between objects in the space, the more similar objects are. Besides Euclidean metrics there are also another metrics suitable for judging of the mutual similarity between the objects, e.g. Sokal metrics, Minkowski metrics, sup-metrics (Manhattan), etc. [4].

Let $\boldsymbol{x} = (x_1, x_2, ..., x_m)$ and $\boldsymbol{y} = (y_1, y_2, ..., y_m)$ be two objects characterised by *m*-dimensional vectors. The coefficients of their similarity are evaluated by following formulas of metrics mentioned before:

- Euclidean metrics

$$d_E(\boldsymbol{x},\boldsymbol{y}) = \left[\sum_{i=1}^{m}(x_i - y_i)^2\right]^{1/2} \tag{1}$$

- Sokal metrics

$$d_S(\boldsymbol{x},\boldsymbol{y}) = \left[\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2\right]^{1/2} \tag{2}$$

- Minkowski metrics

$$d_M(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{i=1}^{m}|x_i - y_i|^r\right)^{1/r} \tag{3}$$

- Sup-metrics

$$d_M(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m}|x_i - y_i| \tag{4}$$

- Chebychev metrics

$$d_\infty(\boldsymbol{x},\boldsymbol{y}) = \max_i \left(|x_i - y_i|\right) \ . \tag{5}$$

3 Methods of Clustering

There are many methods of clustering analysed in variety of scientific contributions. Methods are divided into two groups: hierarchical and non-hierarchical methods of cluster analysis [4].

The methods of hierarchical clustering create a system of mutually different non-empty subsets from the original set of objects where the intersection of each of two subsets is one of them or the empty set. Non-hierarchical clustering makes such a system of subsets from the set of objects where none of the intersections of each of two subsets belongs to that system.

There are two main types of hierarchical clustering agglomerative and division based. Agglomerative methods can create a hierarchical system of decompositions of the set of objects by unifying the sets. At the beginning of clustering each object represents one cluster. On the other hand, division methods gradually divide already existing clusters. At the beginning of this clustering all objects represent one cluster and at the end each object already belongs to separate cluster.

The unifying or dividing of clusters depends on values of coefficients of unsimilarity of two μ-coherent clusters. The distance between compared clusters expresses it.

4 Cluster Similarity Coefficients

The cluster A_x is μ -coherent with the cluster A_y for the defined μ if there is such a sequence of the clusters $A_x = A_1, A_2, ..., A_m = A_y,\ m > 1$ where $D(A_k, A_{k+1}) \leq \mu$ for $k = 1, 2, ..., m-1$. $D(A_i, A_j)$ is the value of similarity of the clusters A_i and A_j and μ is the clustering threshold.

Let $P_1, P_2, ..., P_t$ and $L_1, L_2, ..., L_s$ be two groups of clusters creating two μ -coherent clusters. There are many methods for the calculation of similarity coefficients $D(P_1 \cup P_2 \cup ... \cup P_t, Q)$ and $D(P_1 \cup P_2 \cup ... \cup P_t,\ L_1 \cup L_2 \cup ... \cup L_s)$, where Q is any cluster of the decomposition, which gets to the next decomposition with no change.

Three methods are of interest here:

1. Method of the nearest neighbour - coefficient $D_{nn/d}$

$$D_{nn/d}(P,Q) = \min\{D(P_1,Q), D(P_2,Q), \ldots, D(P_t,Q)\} \qquad (6)$$

$$D_{nn/d}(P,L) = \min\{D(P_1,L_1), D(P_1,L_2), \ldots, D(P_t,L_s)\} \qquad (7)$$

where and $L = (L_1 \cup L_2 \cup \ldots \cup L_s)$.

2. Method of the farthest neighbour - coefficient $D_{fn/d}$

$$D_{fn/d}(P,Q) = \max\{D(P_1,Q), D(P_2,Q), \ldots, D(P_t,Q)\} \qquad (8)$$

$$D_{fn/d}(P,L) = \max\{D(P_1,L_1), D(P_1,L_2), \ldots, D(P_t,L_s)\} \qquad (9)$$

3. Method of average similarity - coefficient $D_{ag/d}$

$$D_{ag/d}(P,Q) = 1/|P| \sum_{i=1}^{t} |P_i| D(P_i,Q) \qquad (10)$$

$$D_{ag/d}(P,Q) = 1/(|P|\ |L|) \sum_{(i,j)}^{txs} |P_i||L_i| D(P_i,L_j)\ , \qquad (11)$$

where $|P|$, $|L|$ are the numbers of objects in the clusters *P*, *L*. Symbol *d* signifies that any similarity coefficients of the objects can be used in the calculation of the similarity coefficients of clusters.

5 Structure of Experiments for Realization of Clustering

The mentioned methods of clustering were applied to the creation of the risk classes of coronary heart disease. The objects, respondents, were represented by seven dimensional vectors with risk factors of atherosclerotic vascular disease: sex, age, smoking, cholesterol, systolic blood pressure, diastolic blood pressure and body weight [6,8].

The input set contained multidimensional data of 107 respondents. The results from clustering were compared with results from the table of European society of cardiology of risk of coronary heart disease. The statistical data show five risk levels with percent change of a coronary event in the next 10 years:

Table 1 Risk levels

5	- very high ... > 40 %
4	- high ... 20-40 %
3	- moderate ... 10-20%
2	- mild ... 5-10 %
1	- low ... < 5 %

These risk levels are based on the risk function derived from the Framingham study [2]. Mentioned risk levels are derived from five risk factors: sex, age, smoking, systolic blood pressure and cholesterol. At the beginning each respondent was assigned the risk level according to the Table 1. This was important for the comparing of risk classes obtained by clustering.

Different threshold values were used to terminate the clustering process for each similarity coefficient. The results were compared with the actual distribution of the input vectors into risk classes. The results were compared with real risk level obtained from Table 1.

Table 2 Results of risk classes created by agglomerative clustering

Similarity coefficients	Classification into the same risk classes							
	the same class		class difference = 1		class difference = 2		class difference = 3	
	number	%	number	%	number	%	number	%
$D_{ag/E}$	93	86.9	14	13.1	0	0	0	0
$D_{ag/S}$	87	78.5	19	17.8	1	0.9	0	0
$D_{ag/\infty}$	89	83.2	18	16.8	0	0	0	0
$D_{nn/E}$	85	79.4	22	20.6	0	0	0	0
$D_{nf/E}$	88	82.2	19	17.8	0	0	0	0

Conclusion

The results from the clustering calculations show a loss of resolution inherent in the creation of clusters. Members of the statistical population will migrate to clusters with “wrong” class characteristics. It is important to note that the clustering effected did not lose move accuracy than a class difference of value 1. The choice of suitable metrics, similarity coefficients of objects and clusters is important. This should if possible, reflect the distribution of the dominant characterising factors.

REFERENCES

[1] ANDERBERG, M.R: Cluster Analysis for Applications. Academic Press, New York, 1973.

[2] ANDERSON, K. at al: An updated coronary risk profile: A statement for health professionals. Circulation 83:356-362, 1991.

[3] HARTIGAN, J.A.: Clustering Algorithms. J.Wiley, New York – London – Sydney - Toronto, 1975.

[4] JARDINE, N. and SIBSON, R.: The Construction of Hierarchic and Non - Hierarchic Classifications. *Computer J.*, No. 11, pp. 177-185, 1968.

[5] LANDRYOVÁ, L. - ZOLOTOVÁ, I.: Integrating Methods of Artificial Intelligence into Control, In: Proceedings of International Carpathian Control Conference, ICCC `2000, Hight Tatras, Podbanské, Slovak Republic, May 23-26, 2000, pp. 447-450, ISBN 80-7099-510-6.

[6] LUKASOVA, A .- SARMANOVA, J.: Metody shlukové analyzy (Metods of cluster analysis - in Czech), SNTL, Praha, 1985.

[7] OCELÍKOVÁ, E.: Significance of Multicriterial Classification for the Selection of a Suitable Control Strategy of Complex Systems. *Elektrotechnický časopis* 44, No. 12, 1992, pp. 372-374.

[8] OCELÍKOVÁ, E.: Multicriterial valuation risk factors. Lékař a technika 24, 1993, No.1, pp.10-12.

[9] OCELÍKOVÁ, E. - KLIMEŠOVÁ, D.: Clustering by Boundary Detection. In: *Proc. of the 4th International Scientific Technical Conference "Process control – ŘÍP 2000*", Pardubice , June 2000, pp. 108, ISBN 80-7194-271-5.

[10] OCELÍKOVÁ, E.-KRIŠTOF, J. : Classification of Multispectral Data. *Journal of Information and Organizational Sciences,* Vol. 25, Number 1, pp. 35-41,Varaždín, 2001, ISSN 0351-1804

[11] OCELÍKOVÁ, E.-ZOLOTOVÁ, I.-KLIMEŠOVÁ, D. : Multikriteriálna klasifikácia zdravotného stavu.. *Lékař a technika,* No. 6, pp. 17-21, 2002*,* ISSN 0301-5491

[12] OCELÍKOVÁ, E. - MADARÁSZ, L.: Contribution to Creation of Macrosituations of Complex System. *Journal of Advanced Computional Intelligence,* FUJI Technology Press, Ltd. Tokyo, Japan 2002*,* Vol. 6, No.2, pp. 79-83, ISSN 1343-0130

[13] TURČAN,A.-OCELÍKOVÁ,E.-MADARÁSZ,L.: Fuzzy C-means Algorithms in Remote Sensing. In: *Proc. of the 1st Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence "SAMI 2003",* Herľany, Slovakia, February 12-14, 2003, pp.207-216, ISBN 963 7154 140