EVALUATION OF CLUSTERS STABILITY BASED ON MINKOWSKI ULTRAMETRICS

Sergio Scippacercola

Dipartimento di Matematica e Statistica, Università degli studi di Napoli "Federico II", Via Cinthia, Napoli – e-mail: *SS@UNINA.IT*

Abstract

The main purpose of this paper is to suggest a family of Minkowski distances as a tool for measuring the stability of clusters yielded by a Cluster Analysis. Preliminarly, we recall the main properties of the Minkowski distances family and we introduce the Minkowski ultrametric distances. Latter a proof of the Theorem on the subdominant Minkowski ultrametric distances is given. Since the ultrametric distance between two units decreases whenever the Minkowski parameter increases, we suggest to evaluate the clusters stability by the identification of the Minkowski parameter that measures the convergence to stable clusters. The validity of the methodology is confirmed by an example of Cluster Analysis on a set of real data. This methodology can be applied to compare the cluster stability among various grouping criteria.

Keywords: *Minkowski Metrics, Ultrametric distance, Clusters Stability, Cluster Analysis.*

1. INTRODUCTION

Cluster Analysis classifies the statistical units, characterized with a pattern of variables, in classes of equivalence by metrics that evaluate the distance or the similarity among statistical units (Scepi, 1997). We define a partition $\wp_s$ of a population $\wp$ of n statistical units in s, non-empty, classes $C_1, C_2, \ldots, C_s$ (clusters), such that (Rizzi, 1985):

$$C_i \neq \varnothing \quad \forall i;$$

$$C_i \bigcap C_j = \varnothing \quad i \neq j = 1,2,\ldots,s;$$

$$\wp = \bigcup_{i=1}^{s} C_i \ .$$

In a cluster, the statistical units must have a high degree of similarity. Every cluster must be well distinct from the others. Each statistical unit must belong to a single cluster. The clustering methods are distinguished in *non-hierarchical* and *hierarchical* ones. The non-hierarchical clustering methods lead to a partition of the n statistical units into k classes defined *a priori*. Hierarchical methods produce a sequence of partitions (from 1 to n clusters) that can be ordered by nested increasing levels until a single cluster is reached (Rizzi, 1985). Each step of the Cluster Analysis is critical for the singleness and the stability of the solutions. Firstly, the choice of the metric, as a measure of the similarity among statistical units, is essential in creating the grouping clusters. There are no systematic studies to forecast the conditions under which the Euclidean metric is robust (Borg, Lingoes, 1987) and it is preferable the choice of one metric rather than another. Secondly, different grouping strategies often produce similar results; the clustering results depend on the followed strategy, on the chosen options and on the type of adopted distance (Fabbris, 1990). The main purpose of this paper is to suggest a methodology that allows to obtain a measure of the clusters stability by the Minkowski metrics. In Section 2, we recall the main properties of the Minkowski distances family and we introduce the ultrametric Minkowski distances. In Section 3, we turn our attention to the methodology that permits the measure of clusters stability. In Section 4, we underline the validity of the methodology by an application on a set of real data.

2. THE MINKOWSKI ULTRAMETRIC DISTANCES

The most widely used distance for data collected in a matrix $\boldsymbol{X}$, of p quantitative variables observed on n statistical units, is the Minkowski distances family. Let d be a real-valued function. The function d is metric if it has the following properties (Gower, Legendre, 1986; Fabbris, 1990):

$$d_{ij} \geq 0 \ (i, j = 1,\ldots,n);$$
$$d_{ii} = 0 \ (i = 1,\ldots,n);$$
$$d_{ij} = d_{ji} (i \neq j = 1,\ldots,n);$$
$$d_{ij} \leq d_{ih} + d_{jh} \ \forall \ i, j, k \ (i \neq j \neq h = 1,\ldots,n).$$

The $\mathbf{D} = (d_{ij})$ distance matrix is symmetric and positive semi-definite (Mardia *et al.*, 1989). The λ order Minkowski distance defined by the function (Borg, Lingoes, 1987):

$$ {}_{\lambda}d_{ij} = \left[\sum_{h=1}^{p} \left| x_{ih} - x_{jh} \right|^{\lambda} \right]^{1/\lambda} \qquad (i, j = 1, \ldots, n; \quad \lambda \text{ integer } \geq 1) \tag{1} $$

is commonly used in applications. Let ${}_{\lambda}\mathbf{D} = ({}_{\lambda}d_{ij})$ be the distance matrix. For each $\lambda \geq 1$ we have different distances. If $\lambda = 1$, the (1) is the Manhattan distance or city-block metric; if $\lambda = 2$, the (1) is the Euclidean distance. Further, we have for $\lambda \to \infty$, the Lagrange-Tchebychev distance or the dominant metric or the L_{∞} metric or the superior norm (Rizzi, 1987):

$$ {}_{\infty}d_{ij} = \lim_{\lambda \to \infty} {}_{\lambda}d_{ij} = \underset{h=1,2,\ldots,p}{Max} \left\{ \left| x_{ih} - x_{jh} \right| \right\}. \tag{2} $$

By the Jensen inequality (Hardy, et. al., 1964), the following inequalities are valid for the Minkowski metrics (Rizzi, 1985):

$$ {}_{1}d_{ij} \geq {}_{2}d_{ij} \geq \ldots \geq {}_{\infty}d_{ij} \qquad (i, j = 1, \ldots, n). \tag{3} $$

If λ increases, the distance between i and j decreases. Also, by a hierarchical Cluster method we can obtain the ${}_{\lambda}u_{ij}$ subdominant ultrametric distance (Scippacercola, 2000) subject to the ultrametric inequality:

$$ {}_{\lambda}u_{ij} \leq Max\,({}_{\lambda}u_{ih}, {}_{\lambda}u_{jh}) \quad \forall i, j, h \tag{4} $$

Let ${}_{\lambda}\mathbf{U} = \left({}_{\lambda}u_{ij} \right)$ be the Minkowski ultrametric distances $(\lambda \geq 1)$.

3. THE MEASURE OF CLUSTERS STABILITY BY THE MINKOWSKI PARAMETER

In order to verify the resulting stability of a hierarchical Cluster Analysis, the problem is to determine the more meaningful partition $\wp(C_1, C_2, \ldots)$ among all possible partitions (Scepi, 1997). The basic criterion (Sneath, Sokal, 1963) to judge a good grouping of the data is the evident difference among two consecutive levels of a dendrogram showing a good grouping of the data before the last aggregation. Other methods, proposed in literature, are based on bootstrap techniques (Jaine, Moreau, 1987) or on perturbation techniques of the data to search the influent observations on the clusters stability (Jolliffe et. al., 1988).

According to the basic criterion (Sneath, Sokal, 1963), we consider the matrix X and we select a grouping criterion of two clusters, for instance single linkage,

Ward method, etc. We set $\lambda = 1$ and by (1) we obtain ${}_1\mathbf{D}$ and the corresponding ${}_1\mathbf{U}$. Let ${}_1\wp = \wp({C_1}, {C_2}, ...)$ be the obtained partition. By $\lambda = 2$, we have ${}_2\mathbf{D}$, ${}_2\mathbf{U} = ({}_2u_{ij})$ and ${}_2\wp = \wp({C_1}, {C_2}, ...)$.

Theorem. The sequence of ultrametric matrices $\{{}_\lambda\mathbf{U}: \lambda = 1, 2, ...\}$ converges to the matrix ${}_\infty\mathbf{U} = ({}_\infty u_{ij})$.

Proof: Indeed, by (3) and (4) since ${}_\lambda u_{ij} \leq {}_\lambda d_{ij}$ for each λ, it follows that:

$$ {}_1u_{ij} \geq {}_2u_{ij} \geq ... \geq {}_\infty u_{ij} \quad (i, j = 1, ..., n). \tag{5}$$

The sequence of scalars $\{{}_\lambda u_{ij}: \lambda = 1, 2, ...\}$ converges to $({}_\infty u_{ij})$ for each $i, j = 1, 2, ..., n$. Therefore, the sequence of the ${}_\lambda\mathbf{U}$ converges to the matrix ${}_\infty\mathbf{U}$.

Following the basic criterion, the convergence allows to identify (*stopping criterion* of the algorithm) when the $\wp({C_1}, {C_2}, ...)$ partition is judged stable: *at λ-th aggregation, the clusters are stable when two successive partitions have the same elements in the classes* $\left(\wp^{(\lambda-1)} = \wp^{(\lambda)}\right)$. Therefore the λ value becomes a measure of stability of a grouping criterion. Also, we can apply this methodology to compare the clusters stability among various grouping criteria.

4. AN APPLICATION TO REAL DATA

In this Section, we briefly describe the results obtained by applying this methodology to real data. We are interested in obtaining groups from a cluster analysis of 20 Italian Districts with 12 observed variables (on criminality, unemployment, businesses starts, difficulty to reach essential services) in 1998 (Istat, 1998). The data are standardized. We choose a hierarchical cluster analysis and the Ward's criterion. We obtain six classes (C_1, C_2, ...,C_6). By increasing λ ($3 \leq \lambda \leq 10$) the results change (Tab. 1). Some Districts, in italic (Tab. 1), change the cluster and are stable in the same cluster for $\lambda \geq 6$. Other Districts, in bold (Tab. 1), always remain in the same cluster proving to be strong forms. We deduce, for this application, that $\lambda = 6$ is the measure of stability. We judge a good and stable grouping of the data when $\lambda = 6$. We have an evident difference (Tab. 1) between two consecutive levels ($\lambda = 5$ and $\lambda = 6$).

Tab. 1: Clusters of the Italian Districts.

	Partitions							
Districts	$\wp^{(\lambda=3)}$	$\wp^{(\lambda=4)}$	$\wp^{(\lambda=5)}$	$\wp^{(\lambda=6)}$	$\wp^{(\lambda=7)}$	$\wp^{(\lambda=8)}$	$\wp^{(\lambda=9)}$	$\wp^{(\lambda=10)}$
Piemonte	C_4	C_5	C_4	$\mathbf{C_5}$	C_5	C_5	C_5	C_5
Valle d'Aosta	C_3	C_3	C_3	$\mathbf{C_3}$	C_3	C_3	C_3	C_3
Lombardia	C_6	C_4	C_6	$\mathbf{C_4}$	C_4	C_4	C_4	C_4
Trentino	C_3	C_3	C_3	$\mathbf{C_3}$	C_3	C_3	C_3	C_3
Veneto	C_6	C_6	C_3	$\mathbf{C_3}$	C_3	C_3	C_3	C_3
Friuli	C_6	C_6	C_3	$\mathbf{C_6}$	C_6	C_6	C_6	C_6
Liguria	C_4	C_4	C_6	$\mathbf{C_4}$	C_4	C_4	C_4	C_4
Emilia	C_6	C_4	C_6	$\mathbf{C_4}$	C_4	C_4	C_4	C_4
Toscana	C_6	C_4	C_6	$\mathbf{C_4}$	C_4	C_4	C_4	C_4
Umbria	C_6	C_6	C_3	$\mathbf{C_6}$	C_6	C_6	C_6	C_6
Marche	C_3	C_3	C_5	$\mathbf{C_3}$	C_3	C_3	C_3	C_3
Lazio	C_4	C_4	C_4	$\mathbf{C_4}$	C_4	C_4	C_4	C_4
Abruzzo	C_6	C_6	C_3	$\mathbf{C_6}$	C_6	C_6	C_6	C_6
Molise	C_5	C_5	C_5	$\mathbf{C_5}$	C_5	C_5	C_5	C_5
Campania	C_2	C_2	C_2	$\mathbf{C_2}$	C_2	C_2	C_2	C_2
Puglia	C_2	C_2	C_2	$\mathbf{C_2}$	C_2	C_2	C_2	C_2
Basilicata	C_5	C_5	C_5	$\mathbf{C_5}$	C_5	C_5	C_5	C_5
Calabria	C_1	C_1	C_1	$\mathbf{C_1}$	C_1	C_1	C_1	C_1
Sicilia	C_1	C_1	C_1	$\mathbf{C_1}$	C_1	C_1	C_1	C_1
Sardegna	C_5	C_5	C_5	$\mathbf{C_5}$	C_5	C_5	C_5	C_5

5. CONCLUSIONS

In this paper we introduce a new measure of clusters stability. In this field the on-going present study gives the first results. The present approach introduces the ultrametric Minkowski distances and uses the λ variations as a "metal ring for focusing" on a camera lens. "In focus" symbolically coincides with the stability of the units to stay in the same clusters. This approach, different from others, does not

judge the stability *ex-post* and does not intervene on the data, but proposes *ex-ante* a family of metrics to measure a stable grouping. Also, in perspective, this tool can be used to evaluate the speed of convergence among various clustering methods applied on the same data.

ACKNOWLEDGE

The present paper is financially supported by a grant of the *Dipartimento di Matematica e Statistica* (Università degli Studi di Napoli "Federico II" – Italy).

REFERENCES

BORG I., LINGOES J. (1987) *Multidimensional Similarity Structure Analysis*, Springer-Verlag.

FABBRIS L. (1990) *Analisi esplorativa di dati multidimensionali*, Cleup ed.

GOWER J. C., LEGENDRE P. (1986) Metric and Euclidean Properties of Dissimilarity Coefficients, *Journal of Classification*, 3, 5-48.

HARDY G.H., LITTLEWOOD J. E., POLYA G. (1964) *Inequalities*, Cambridge at the University Press.

ISTAT (1998) *http://www.istat.it*

JAINE A.K., MOREAU J.V. (1987) Bootstrap Technique in Cluster Analysis, *Pattern Recognition*, 2.

JOLIFFE I.T., JONES B., MORGAN J.T. (1988) Stability and influence in cluster analysis, *Data Analysis and Informatics,* V, Diday ed.

MARDIA K.V., KENT J.T., BIBBY J.M. (1989) *Multivariate Analysis*, Academic Press.

RIZZI A. (1985) *Analisi dei dati*, La Nuova Italia Scientifica.

RIZZI A. (1987) Measures of distance and dissimilarity, *Methods for Multidimensional Data Analysis*, ECAS European Courses in Advanced Statistics, Dipartimento di Matematica e Statistica, Università degli studi di Napoli, 9-28.

SCEPI G. (1997) *Stabilità e validazione nei metodi fattoriali*, Dipartimento di Matematica e Statistica, Università degli studi di Napoli "Federico II", Serie Ricerca, Rocco Curto ed.

SCIPPACERCOLA S. (2000) Il confronto dei vincoli ultrametrici di criteri classificatori mediante l' Analisi in Componenti principali rispetto ad un sottospazio di riferimento, *XL Riunione scientifica della Società italiana di Statistica*, Firenze.

SNEATH P. H., SOKAL R.R. (1963) *Numerical Taxonomy*, Freeman.

LE ULTRAMETRICHE DI MINKOWSKI PER LA VALUTAZIONE DELLA STABILITÀ DI UNA CLASSIFICAZIONE DEI DATI

Riassunto

Nel presente approccio viene considerata la famiglia delle distanze di Minkowski per misurare la stabilità dei raggruppamenti di una classificazione dei dati. Dopo aver richiamato le principali proprietà delle distanze di Minkowski, si introducono le relative ultrametriche. Viene dimostrato un Teorema sulle distanze ultrametriche sottodominanti il cui uso consente il riconoscimento e la misura della stabilità delle partizioni. Variando il parametro delle distanze di Minkowski si ottiene una successione di distanze metriche e di corrispondenti distanze ultrametriche. Ad ogni incremento del parametro di Minkowski le distanze ultrametriche tra gli oggetti da classificare diminuiscono ed aumenta la stabilità dei gruppi: gruppi coesi diventano più coesi; gruppi isolati rimangono tali. Si propone, quindi, un metodo che permette l'identificazione del valore del parametro di Minkowski che rappresenta una misura di stabilità dei gruppi a rimanere nella stessa classe. Un'applicazione della metodologia ad un caso reale mette in evidenza la validità del metodo. La metodologia può essere, inoltre, applicata per confrontare varie analisi di classificazione sullo stesso insieme di dati e per valutare quale tra i vari criteri produce una più veloce convergenza alla stabilità.