Clustering Algorithms for Convexity-Based Clustering Criteria

Hans-Hermann Bock
Institut für Statistik, Technische Hochschule Aachen,
D-52056 Aachen, Germany, bock@stochastik.rwth-aachen.de

Clustering methods are designed in order to find homogeneous subgroups in a (typically large) set $\mathcal{O}$ of n objects which are described by data which determine the mutual similarities and dissimilarities among objects. In many situations these data consists in a matrix $\underline{X} = (x_{kj})_{n\times p}$ based on measurements on p quantitative variables, resulting in n data vectors $x_k = (x_{k1}, ..., x_{kp})' \in R^p$. A very common clustering strategy is to search for an 'optimum' m-partition $\mathcal{C} = (C_1, ..., C_m)$ of the set $\mathcal{O} = \{1, ..., n\}$ with a specified number of classes $C_1, ..., C_m \subset \mathcal{O}$. The underlying optimality criterion might be, e.g., the classical variance or SSQ criterion

$$g_n(\mathcal{C}) = \sum_{i=1}^{m} \sum_{k\in C_i} ||x_k - \overline{x}_{C_i}||^2 = \sum_{k=1}^{n} ||x_k||^2 - \sum_{i=1}^{m} |C_i| \cdot ||\overline{x}_{C_i}||^2 \to \min_{\mathcal{C}} \tag{1}$$

or some suitable generalization.

This paper deals with discrete clustering criteria of the type

$$G_n(\mathcal{C}) = \sum_{i=1}^{m} |C_i| \cdot \phi(\overline{x}_{C_i}) \to \max_{\mathcal{C}} \tag{2}$$

where $\phi(\cdot)$ is a convex function on R^p (in fact, (2) is equivalent to (1) for the special case $\phi(x) = ||x||^2$), and with their continuous counterpart given by

$$G(\mathcal{B}) = \sum_{i=1}^{m} P(X \in B_i) \cdot \phi(E[X|X \in B_i]) \to \max_{\mathcal{B}} \tag{3}$$

where X denotes a random vector in R^p and maximization is over all m-partitions $\mathcal{B} = (B_1, ..., B_m)$ of the entire sample space R^p. Such criteria can be met in various statistical and data-analytic problems (see below).

It is well-known that the variance criterion can be (approximately) minimized by considering the two-variables criterion

$$\gamma_n(\mathcal{C}, \mathcal{Z}) = \sum_{i=1}^{m} \sum_{k\in C_i} ||x_k - z_i||^2 \to \min_{\mathcal{C},\mathcal{Z}} \tag{4}$$

with a system $\mathcal{Z} = (z_1, ..., z_m)$ of 'class representatives' $z_i \in R^p$, and minimizing recursively with respect to the system $\mathcal{Z}$ (resulting in *class centroids* $z_i = \overline{x}_{C_i}$) and $\mathcal{C}$ (resulting in a *minimum-distance partition*) in turn. This *k-means algorithm* cannot be directly transferred to the new criterion $G_n(\mathcal{C})$ or $G(\mathcal{B})$ because there is no obvious two-variable formulation.

This paper shows that such a formulation with a suitable criterion $\gamma(\cdot)$ can be found by considering, to each class, a support-plane (tangent) hyperplane H_i to the function $\phi(x)$ (for a class-specific support point $x = z_i$), and to minimize the area γ_n between the function ϕ and the resulting 'maximum support plane surface'. This approach leads to the idea of a *maximum-support-plane partition* (MSP) and yields the desired analogue to the classical k-means algorithm for minimizing G_n or G (*maximum-support-planes algorithm*).

Various criteria of the new type G_n or G have been considered in the literature: Bock (1983, 1991, 1994) has used them for optimum discretization (stratified sampling design, quantization)

of R^p and for the optimization of statistical tests based on the discretized data (χ^2 goodness-of-fit test, discrimination). This involves likelihood ratios and leads, e.g., to the maximization of the *χ^2-noncentrality parameter*, of the *Kullback-Leibler's information*, of *Chernoff's distance*, and, more generally, of Csizar's *ϕ-divergence*. Quite recently, Strasser (2000a, 2000b) and Pötzelberger and Strasser (2001) have generalized the method, embedded it into a general strategy of data compression in inferential statistics, and applied it, e.g., to a special version of Kohonen maps and to special similarity data.

REFERENCES

Bock, H.H. (1983): *A clustering algorithm for choosing optimal classes for the chi-squared test.* Bull. Intern. Statist. Inst., 44th Session, Madrid 1983, Vol. II: Contributed papers, 758-762.

Bock, H.H. (1991): A clustering algorithm for maximizing ϕ-divergence, non-centrality and discriminating power. In: M. Schader (ed.): *Analyzing and modeling data and knowledge.* Springer-Verlag, Heidelberg, 1991, 19-36.

Bock, H.H. (1994): Information and entropy in cluster analysis. In: H. Bozdogan et al. (eds.): *The Frontiers of Statistical Modeling: An Informational Approach.* Proc. First US/Japan Conference on Statistical Modeling, Knoxville, Tennessee, May 1992. Kluwer Academic Press, Dordrecht, 1994, Vol. II, 115-147.

Bock, H.H. (2001): *Convexity-based clustering criteria: a new approach.* Rector's lecture, Academy of Economics, Krakow (in press).

Pötzelberger, K., and H. Strasser (2001): Clustering and quantization by MSP-partitions. *Statistics and Decisions.* (In press.)

Strasser, H. (2000a): Reduction of complexity. In: J.A. Mazanec and H. Strasser (eds.): *A nonparametric approach to perceptions-based market segmentation: foundations.* Interdisciplinary Studies in Economics and Management, Vol. 1. Springer-Verlag, Berlin.

Strasser, H. (2000b): Towards a statistical theory of optimal quantization. In: W. Gaul, O. Opitz, and M. Schader (eds.): *Data Analysis. Scientific modeling and practical application.* Springer-Verlag, Heidelberg, 2000, 369-383.

RESUME

Cette contribution présente un algorithme de transfert du type k-means (nuées dynamiques) pour arranger un ensemble d'objets dans un nombre fixe de classes non disjointes telles qu'un critère de classification soit (aproximativement) optimisé. Ce critère généralise le critère classique de la 'variance dans les classes' d'une manière qui introduit une fonction convexe des centroides des classes. La convexité permet de formuler un problème dual d'optimisation et de proposer un algorithme iteratif qui utilise des hyperplans maximaux de support, d'une manière comme on utilise la partition de distance minimale dans le cas classique.