

Clustering methods for convexity-based criteria with applications

Metodi di clustering con criteri basati sulla convessità: teoria e applicazioni

Hans-Hermann Bock
Institute of Statistics
Aachen Technical University
D-52056 Aachen, Germany

Riassunto: L'articolo presenta metodi di clustering e di segmentazione per dati quantitativi o distribuzioni di probabilità p -dimensionali. Essi si basano su un criterio di clustering generale che considera una arbitraria funzione convessa ϕ dei centroidi, e generalizzano il classico criterio di varianza (SSQ). Si dimostra che l'algoritmo classico delle k -medie può essere modificato per massimizzare questo criterio basato sulla convessità, introducendo partizioni con massimo supporto piano di R^p per ϕ . Diversi problemi di ottimizzazione coinvolgono un criterio di questo tipo: stratificazione ottima, massimizzazione del parametro di centralità del χ^2 , piani di proiezione di Kohonen e clustering bidimensionali per tavole di contingenza.

Keywords: Clustering, optimum partition, generalized k -means, convexity, two-way clustering, ϕ -divergence

1. Classical SSQ clustering

This paper deals with the *clustering problem* for a set of n data vectors $x_1, \dots, x_n \in R^p$ which describe the properties of a set $\mathcal{O} = \{1, \dots, n\}$ of n objects in terms of p quantitative variables. More specifically, for a given integer m , we look for a partition $\mathcal{C} = (C_1, \dots, C_m)$ of $\mathcal{O}$ with m classes $C_1, \dots, C_m \subset \mathcal{O}$ which is 'optimum' in some sense. A common clustering approach is based on the *variance (SSQ) criterion* defined by

$$g_n(\mathcal{C}) := \sum_{i=1}^m \sum_{k \in C_i} \|x_k - \bar{x}_{C_i}\|^2 = \sum_{k=1}^n \|x_k\|^2 - \sum_{i=1}^m |C_i| \cdot \|\bar{x}_{C_i}\|^2 \rightarrow \min_{\mathcal{C}} \quad (1)$$

where $\bar{x}_{C_i} := (\sum_{k \in C_i} x_k) / |C_i|$ denotes the centroid of the observations x_k belonging to the class C_i (here $|C_i|$ is the number of objects in C_i). An *approximate* optimization can be attained by an iterative improvement of an initial partition $\mathcal{C}^{(0)}$ by exchanging or reassigning objects among classes in an appropriate way. In particular, the classical *k-means algorithm* (dynamic clustering) starts from the remark that the *one-parameter* minimization problem (1) is equivalent to the *two-parameter* minimization problem

$$g_n(\mathcal{C}, \mathcal{Z}) := \frac{1}{n} \sum_{i=1}^m \sum_{k \in C_i} \|x_k - z_i\|^2 \rightarrow \min_{\mathcal{C}, \mathcal{Z}} \quad (2)$$

with two variables $\mathcal{C}$ and $\mathcal{Z}$ where $\mathcal{Z} = (z_1, \dots, z_m)$ is a system of *centers* or *prototypes* $z_i \in R^p$ for the m classes. The *k-means algorithm* proceeds by (a) first minimizing

$g_n(\mathcal{C}, \mathcal{Z})$ with respect to the center system $\mathcal{Z}$; this yields the class centroids $z_i^* = \bar{x}_{C_i}$, (b) then minimizing $g_n(\mathcal{C}, \mathcal{Z})$ with respect to the m -partition $\mathcal{C}$; this yields a *minimum-distance partition* $\mathcal{C}^*$ where each object k is assigned to the closest centroid z_i^* , i.e. with minimum Euclidean distance $\|x_k - z_i^*\|$, and (c) then to iterate this alternating optimization procedure until stationarity.

The previous criteria and the algorithm can easily be transferred to the 'continuous' case where P is the distribution of an arbitrary random vector X in R^p and we look for an optimum m -partition $\mathcal{B} = (B_1, \dots, B_m)$ of R^p . Then (1) and (2) take the form of the *segmentation (quantization, stratification) problem* (Bock 1974):

$$\begin{aligned} g(\mathcal{B}) &:= \sum_{i=1}^m \int_{B_i} \|x - E[X|X \in B_i]\|^2 dP(x) \\ &= E[\|X\|^2] - \sum_{i=1}^m P(X \in B_i) \cdot \|E[X|X \in B_i]\|^2 \rightarrow \min_{\mathcal{B}} \end{aligned} \quad (3)$$

$$g(\mathcal{B}, \mathcal{Z}) := \sum_{i=1}^m \int_{B_i} \|x - z_i\|^2 dP(x) \rightarrow \min_{\mathcal{B}, \mathcal{Z}}. \quad (4)$$

2. A convexity-based clustering criterion

The minimization problem (1) is equivalent to the maximization problem

$$G_n(\mathcal{C}) := \sum_{i=1}^m \frac{|C_i|}{n} \cdot \|\bar{x}_{C_i}\|^2 \rightarrow \max_{\mathcal{C}}. \quad (5)$$

We generalize this clustering criterion by replacing the squared Euclidean distance $\|x\|^2$ by an arbitrary *convex* function $\phi(x)$ on R^p and obtain the optimization problem

$$G_n(\mathcal{C}) = \sum_{i=1}^m \frac{|C_i|}{n} \cdot \phi(\bar{x}_{C_i}) \rightarrow \max_{\mathcal{C}} \quad (6)$$

with its 'continuous' counterpart:

$$G(\mathcal{B}) = \sum_{i=1}^m P(X \in B_i) \cdot \phi(E[X|X \in B_i]) \rightarrow \max_{\mathcal{B}}. \quad (7)$$

We will show in section 5 that this type of optimization problems arises in various domains of statistics and data analysis. In fact, criteria of the kind (7) have been introduced and investigated by Bock (1983, 1991, 1994) in order to maximize the power of statistical tests by a suitable discretization (stratification). Pötzelberger and Strasser (2001) have considered the p -dimensional case and presented a detailed mathematical investigation.

3. Formulation of an equivalent two-parameter criterion

In order to maximize (approximately) the criteria (6) and (7) we will design an analogue to the classical k -means algorithm. This requires an appropriate *two-parameter* criterion – similar to (2) – which is equivalent to (6) or (7). This analogue can be obtained

in the following way.

Recall that, for each $z \in R^p$, the function $u = \phi(x)$ has a *support hyperplane* $u = t(x; z, a) := a^{tr}(x - z) + \phi(z)$ (with a coefficient vector $a \in R^p$) with the properties

$$\phi(x) \geq t(x; z, a) \quad \text{for all } x \in R^p. \quad (8)$$

and $\phi(z) = t(z; z, a)$ in the special case $x = z$. If the function ϕ is differentiable and $a = a(z) := \text{grad}_x \phi(x)|_{x=z}$, this hyperplane is just the *tangent hyperplane* to the surface $u = \phi(x)$ at the support point $x = z$. For ease of presentation, differentiability will be assumed in the sequel and the direction vector a is typically omitted.

In this notation, we introduce for all classes $B_i \subset R^p$ an arbitrary support point $z_i \in R^p$ and compile them in the *support system* $\mathcal{Z} = (z_1, \dots, z_m)$. We also consider the tangent (support) hyperplane H_i for ϕ at the support point z_i which is described by $u = t(x; z_i) := a_i^{tr}(x - z_i) + \phi(z_i)$ with $a_i := a(z_i)$. Finally we consider the piecewise linear function $u = p(x; \mathcal{B}, \mathcal{Z})$ which is given by the hyperplane H_i in the domain B_i , i.e., $p(x) := t(x; z_i)$ for $x \in B_i$. Obviously $p(x; \mathcal{B}, \mathcal{Z}) \leq \phi(x)$ for all x . The following theorem specifies the required two-parameter criterion:

Theorem: The 'continuous' maximization problem (7) is equivalent to the 'minimum volume problem' defined by:

$$K(\mathcal{B}, \mathcal{Z}) := E[\phi(X) - p(X; \mathcal{B}, \mathcal{Z})] = \sum_{i=1}^m \int_{B_i} [\phi(x) - t(x; z_i)] dP(x) \rightarrow \min_{\mathcal{B}, \mathcal{Z}}. \quad (9)$$

Similarly, the clustering problem (6) is equivalent to:

$$K_n(\mathcal{C}, \mathcal{Z}) := \frac{1}{n} \sum_{i=1}^m \sum_{k \in C_i} [\phi(x_k) - t(x_k; z_i)] \rightarrow \min_{\mathcal{C}, \mathcal{Z}}. \quad (10)$$

Note that the criterion $K(\mathcal{B}, \mathcal{Z})$ can be interpreted as a weighted volume between the surfaces $u = \phi(x)$ and $u = p(x)$ in R^{p+1} .

It appears that we can *partially* minimize these two-parameter criteria with respect to each *single* parameter, i.e., $\mathcal{B}$ (or $\mathcal{C}$) and $\mathcal{Z}$, in an *explicit* form (see below). Therefore, we obtain an analogue of the classical k -means algorithm for the new clustering or quantization criteria (7) and (6) just by iteratively minimizing the criteria (9) and (10) with respect to the partition $\mathcal{B}$ (or $\mathcal{C}$) and then with respect to the support system $\mathcal{Z}$ in an alternating way. This *MSP algorithm* is described in the following section.

4. The 'maximum-support-plane' algorithm

We describe here the *maximum-support-plane (MSP) algorithm* for the 'discrete' case, i.e., for finding a sub-optimal m -partition $\mathcal{C}$ for n data points $x_1, \dots, x_n \in R^p$ with the clustering criterion (6). Since we assume a differentiable function ϕ this will in fact be a *maximum-tangent-plane algorithm* which proceeds as follows for $t = 0, 1, 2, \dots$:

- $t = 0$: We select an arbitrary m -partition $\mathcal{C}^{(0)} = (C_1^{(0)}, \dots, C_m^{(0)})$ of the set $\mathcal{O}$ of objects.
- $t \rightarrow t + 1$:
- (i) We determine the system $\mathcal{Z}^{(t)} = (z_1^{(t)}, \dots, z_m^{(t)})$ of support points that minimizes the

criterion $K_n(\mathcal{C}^{(t)}, \mathcal{Z})$ with respect to $\mathcal{Z}$. This system is given by the centroids $z_i^{(t)} := \bar{x}_{C_i^{(t)}}$ of the classes $C_i^{(t)}$, i.e., by $\mathcal{Z}^{(t)} := (\bar{x}_{C_1^{(t)}}, \dots, \bar{x}_{C_m^{(t)}})$.

(ii) We determine the partition $\mathcal{C}^{(t+1)} = (C_1^{(t+1)}, \dots, C_m^{(t+1)})$ of the n objects that minimizes the criterion $K_n(\mathcal{C}, \mathcal{Z}^{(t)})$ with respect to $\mathcal{C}$. We can show that $\mathcal{C}^{(t+1)}$ is essentially given by the *maximum-support-plane partition* (here: a *maximum-tangent-plane partition*) that is generated by the support system $\mathcal{Z}^{(t)}$, i.e., with classes

$$C_i^{(t+1)} := \{k \in \mathcal{O} \mid t(x_k; z_i^{(t)}) = \max_{j=1, \dots, m} t(x_k; z_j^{(t)})\} \quad \text{for } i = 1, \dots, m \quad (11)$$

where $u = t(x; z_i^{(t)})$ is the tangent hyperplane H_i for ϕ in the support point $z_i^{(t)}$.

• *Stopping criterion:*

The two previous steps are iterated for $t = 1, 2, \dots$ until the partitions $\mathcal{C}^{(t)}$ do not change any more or until the support points $\bar{x}_{C_i^{(t)}}$ attain an (approximately) stationary state.

5. Applications and special cases

5.1 Maximizing the non-centrality parameter of the χ^2 goodness-of-fit test

If Y is a random vector in R^q with distribution P , the well-known χ^2 goodness-of-fit test checks the hypothesis $H_0 : P = P_0$ for a given distribution P_0 . Based on a sample of size n it uses the test statistic $\chi^2 := n \sum_{i=1}^m (N_i/n - P_0(D_i))^2 / P_0(D_i)$ where $\mathcal{D} = (D_1, \dots, D_m)$ is a given m -partition of R^q and N_i the number of observations in the class D_i . The problem of optimizing the test performance by a suitable choice of $\mathcal{D}$ has been investigated by various authors (e.g., Mann and Wald 1942, Spruill 1977). Bock (1983) optimizes the asymptotic (Pitman or Bahadur) efficiency of the χ^2 -test when testing against $P = P_1$. This amounts to maximizing the non-centrality parameter

$$\begin{aligned} \tilde{G}(\mathcal{D}) &:= \sum_{i=1}^m \frac{(P_1(D_i) - P_0(D_i))^2}{P_0(D_i)} \\ &= \sum_{i=1}^m P_0(D_i) \lambda(D_i) = \sum_{i=1}^m P_0(D_i) \cdot \phi(\lambda(D_i)) \end{aligned} \quad (12)$$

where $\lambda(D_i) := P_1(D_i)/P_0(D_i)$ is the discretized likelihood ratio and ϕ the convex function $\phi(\lambda) := (\lambda - 1)^2$. We can show that maximizing this parameter is equivalent to the convexity-based 'clustering' problem (7). In fact, if we introduce the likelihood ratio $\lambda(y) = f_1(y)/f_0(y)$ for the densities of P_1 and P_0 and consider the random variable $X := \lambda(Y)$ (of dimension $p = 1$), (12) is equivalent to finding an optimum partition $\mathcal{B} = (B_1, \dots, B_m)$ of R^1 in the sense:

$$G(\mathcal{B}) := \sum_{i=1}^m P_0^X(B_i) \cdot \phi(E_0[X|X \in B_i]) \rightarrow \max_{\mathcal{B}} \quad (13)$$

where the partitions $\mathcal{D}$ of R^q and $\mathcal{B}$ of R^1 are related by $D_i := \{y \in R^q \mid \lambda(y) \in B_i\}$ and $E_0[X|X \in B_i] = E_0[\lambda(Y)|Y \in D_i] = P_1(D_i)/P_0(D_i) = \lambda(D_i)$. Bock (1983) obtains numerical results, e.g., for p -dimensional normal distributions.

5.2 Maximizing the ϕ -divergence between discretized distributions

More generally, we may look for a partition $\mathcal{D}$ which maximizes the performance of other tests between $H_0 : P = P_0$ and $H_1 : P = P_1$ based on the counts N_i . Corresponding performance measures have often the form of Csiszár's ϕ -divergence:

$$\begin{aligned}\tilde{G}(\mathcal{D}) &:= \sum_{i=1}^m P_0(D_i) \cdot \phi(\lambda(D_i)) \\ &= \sum_{i=1}^m P_0^X(B_i) \cdot \phi(E_0[X|X \in B_i]) =: G(\mathcal{B}) \rightarrow \max_{\mathcal{B}}.\end{aligned}\quad (14)$$

with a convex ϕ (Bock 1983). Special cases include the *Kullback-Leibler discriminating information Chernoff* and *Bhattacharyya distances*, and the *variation distance*.

5.3 Kohonen's iterative projection method

Pötzelberger and Strasser (2001) have shown that for an arbitrary convex function ϕ the maximization problem (7) is equivalent to the problem

$$F_m(a_1, \dots, a_m) := \int_{R^p} \max_{j=1, \dots, m} \{a_j^{tr} x - \phi^c(a_j)\} dP(x) \rightarrow \max_{a_1, \dots, a_m \in R^p} \quad (15)$$

where $\phi^c(a) := \sup_{x \in R^p} \{a^{tr} x - \phi(x)\}$ (with $a \in R^p$ a direction vector) is the conjugate convex function of ϕ . This implies that for the Euclidean norm $\phi(x) := \|x\|$ with $\phi^c(a) = 0 (= \infty)$ for $\|a\| \leq 1$ (for $\|a\| > 1$) the problem (7), i.e.:

$$G(\mathcal{B}) = \sum_{i=1}^m P(X \in B_i) \cdot \|E[X|X \in B_i]\| \rightarrow \max_{\mathcal{B}} \quad (16)$$

and the problem (17), i.e.:

$$\begin{aligned}F_m(a_1, \dots, a_m) &:= \int_{R^p} \max_{j=1, \dots, m} \{a_j^{tr} x\} dP(x) \\ &= \sum_{i=1}^m \int_{B_i^*} a_i^{tr} x dP(x) \rightarrow \max_{a_1, \dots, a_m \text{ with } \|a_j\| \leq 1 \text{ for all } j}\end{aligned}\quad (17)$$

are equivalent. Here $B_i^* := \{x \in R^p | a_i^{tr} x = \max_j \{a_j^{tr} x\}\}$ for $i = 1, \dots, m$. This latter problem corresponds to a problem investigated by Kohonen (1984): Consider an m -partition $\mathcal{B} = (B_1, \dots, B_m)$ of R^p and for each class B_i a direction $a_i \in R^p$. Project all $x \in B_i$ on the direction a_i . Look for a $\mathcal{B}$ and m directions $a_1, \dots, a_m$ such that the 'average projection length' is maximized and, insofar, the subspaces spanned by the a_i discriminate best the classes of $\mathcal{B}$. Whereas Kohonen has presented a stochastic approximation method for solving (17) we can use the MSP algorithm here for resolving the equivalent problem (16).

5.4 Simultaneous clustering of the rows and columns of a contingency table

Consider two random variables U and V with domains $\mathcal{U}$ and $\mathcal{V}$ and a joint distribution $P = P^{(U,V)}$ with a density $f(u, v)$. For ease of presentation we assume $\mathcal{U}$ and $\mathcal{V}$ to be finite here such that P is a discrete distribution with counting density $f(u, v) = P(U = u, V = v)$. Robardet (2001) considers the problem of finding an m -partition $\mathcal{A} = (A_1, \dots, A_m)$ of $\mathcal{U}$ and an l -partition $\mathcal{B} = (B_1, \dots, B_l)$ of $\mathcal{V}$ such that the dependence structure of (U, V) is optimally reflected by the two partitions. We interpret this idea in the sense of maximizing the Kullback-Leibler information

$$G_1(\mathcal{A}, \mathcal{B}) := \sum_{i=1}^m \sum_{j=1}^l P(A_i)P(B_j) \cdot \left[-\log \frac{P(A_i \times B_j)}{P(A_i)P(B_j)} \right] \rightarrow \max_{\mathcal{A}, \mathcal{B}}. \quad (18)$$

We can use an alternating version of the MSP algorithm for improving an a given initial pair $\mathcal{A}^{(0)}$ and $\mathcal{B}^{(0)}$ by minimizing first with respect to $\mathcal{B}$, then with respect to $\mathcal{A}$ and iterating these steps in turn. In fact, for a fixed $\mathcal{A}$, we denote

- by P_0 the marginal distribution of V with density $f^V(v) = P(V = v)$,
- by P_{1i} the conditional distribution of V under $U \in A_i$ with density $f_{1i}(v) := P(V = v|U \in A_i) = \sum_{u \in A_i} f(u, v)/P(A_i)$

such that $\lambda_i(v) := f_{1i}(v)/f^V(v) = (\sum_{u \in A_i} f(u, v))/[P(A_i)f^V(v)]$ is a class-specific likelihood ratio and $\lambda(v) := (\lambda_1(v), \dots, \lambda_m(v))^{tr}$ the vector of all likelihood ratios.

Then we can write the criterion $G_1(\mathcal{A}, \mathcal{B})$ in the form

$$\begin{aligned} G_1(\mathcal{A}, \mathcal{B}) &= \sum_{j=1}^l P(B_j) \cdot \sum_{i=1}^m P(A_i) \cdot [-\log \frac{P(B_j|A_i)}{P(B_j)}] \\ &= \sum_{j=1}^l P(B_j) \cdot \sum_{i=1}^m P(A_i) \cdot [-\log E_0[\lambda_i(V)|B_j]] \\ &= \sum_{j=1}^l P(B_j) \cdot \phi_{\mathcal{A}}(E_0[\lambda(V)|B_j]) \end{aligned} \quad (19)$$

where the function $\phi_{\mathcal{A}}(\lambda) := \sum_{i=1}^m P(A_i) \cdot [-\log \lambda_i]$ is obviously convex on R^m . Therefore (19) has just the form of the convexity-based basic criterion (7) and we can maximize (19) (here: with respect to $\mathcal{B}$) with the MSP algorithm.

References

- Bock, H.H. (1974). *Automatische Klassifikation*. Vandenhoeck & Ruprecht, Göttingen.
- 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.
- Kohonen, T. (1984). *Self-organization and associative memory*. Springer, Berlin.
- Mann, H.B. and Wald, A. (1942). On the choice of the number of class intervals in the application of the chi-square test. *Ann. Math. Statist.* 16, 305-320.
- Pötzlberger, K. and Strasser, H. (2001). Clustering and quantization by MSP-partitions. *Statistics and Decisions* 19 (2001), 331-371.
- Robardet, C. and Feschet, F. (2001). *Une mesure d'association en classification simultanée non-supervisée*. Actes du 8ème Congrès de la Société Francophone de Classification, Pointe-à-Pitre, 287-294.