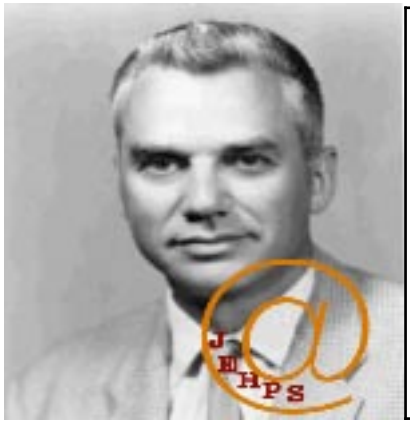

Journ@l Electronique d'Histoire des Probabilités et de la Statistique

Electronic Journ@l for History of Probability and Statistics

Vol 4, n°2; Décembre/December 2008

www.jehps.net

Origins and extensions of the k-means algorithm in cluster analysis

Hans-Hermann BOCK [1]

Abstract. This paper presents a historical view of the well-known k-means algorithm that aims at minimizing (approximately) the classical SSQ or variance criterion in cluster analysis . We show to which authors the different (discrete and continuous) versions of this algorithm can be traced back, and which were the underlying applications. Moreover, the paper describes a series of extensions and generalizations of this algorithm (for fuzzy clustering, maximum likelihood clustering, convexity-based criteria,...) that shows the importance and usefulness of the k-means approach and related alternating minimization techniques in data analysis.

1. Introduction

Cluster analysis emerged as a major topic in the 1960's and 1970's when the monograph 'Principles of numerical taxonomy' by Sokal and Sneath (1963) motivated world-wide research on clustering methods and thereby initiated the publication of a broad range of books such as 'Les bases de la classification automatique' (Lerman 1970), 'Mathematical taxonomy' (Jardine and Sibson 1971), 'Cluster analysis for applications' (Anderberg 1973), 'Cluster analysis' (Bijnen 1973), 'Automatische Klassifikation' (Bock 1974), 'Empirische Verfahren zur Klassifikation' (Sodeur 1974), 'Probleme und Verfahren der numerischen Klassifikation (Vogel 1975), 'Cluster-Analyse-Algorithmen (Späth 1975, 1985), and 'Clustering algorithms' (Hartigan 1975). With the consequence that the basic problems and methods of clustering became well-known

[1] Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany. A shorter version of this article has been published in the Festschrift for E. Diday published by Brito et al. (2007).

in a broad scientific community, in statistics, data analysis, and - in particular - in applications.

One of the major clustering approaches is based on the sum-of-squares (SSQ) criterion and on the algorithm that is today well-known under the name 'k-means'. When tracing back this algorithm to its origins, we see that it has been proposed by several scientists in different forms and under different assumptions. Later on, many researchers investigated theoretical and algorithmic aspects, and modifications of the method, e.g., when considering 'continuous' analogues of the SSQ criterion (Cox 1957, Fisher 1958, Engelman and Hartigan 1969, Bock 1974), by investigating the asymptotic behaviour under random sampling strategies (Hartigan 1975, Pollard 1982, Bock 1985), and by extending its domain of application to various new data types and probabilistic models. Later on, Diday's monograph (Diday et al. 1979), written with 22 co-authors, marks a considerable level of generalization of the basic idea and established its usage for model-based clustering.

This article surveys the origins and some important extensions of the k-means algorithm. In all situations the problem consists in partitioning a set of n objects or of n data points $x_1, ..., x_n$ (or even a space $\mathcal{X}$, e.g., $\mathbb{R}^p$) into a fixed known number k of non-empty disjoint classes (clusters) $C_1, ..., C_k$, say, that are 'as homogeneous as possible' with respect to some given data[2]. In Section 2 we formulate the SSQ clustering problem and the *k-means algorithm*. Section 3 describes the most early papers proposing the SSQ criterion and the k-means algorithm. Section 4 concentrates on extensions of the SSQ criterion that lead to so-called *generalized k-means algorithms*. Section 5 is devoted to the fuzzy k-means algorithm. Finally, Section 6 deals with one- and two-parameter criteria and shows how a 'convexity-based' clustering criterion can be minimized by a *k-tangent algorithm*.

2. k-means clustering for the SSQ criterion

There are two versions of the well-known SSQ clustering criterion: the 'discrete' and the 'continuous' case.

Discrete SSQ criterion for data clustering: Given n data points $x_1, ..., x_n$ in $\mathbb{R}^p$ and a k-partition $\mathcal{C} = (C_1, ..., C_k)$ of the set $\mathcal{O} = \{1, ..., n\}$ of underlying 'objects' with non-empty classes $C_i \subset \mathcal{O}$, the discrete SSQ criterion (also termed: variance criterion, inertia, or trace criterion) is given by

$$g_n(\mathcal{C}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - \overline{x}_{C_i}||^2 \ \rightarrow \ \min_{\mathcal{C}} \tag{1}$$

where $\overline{x}_{C_i}$ denotes the centroid of the data points x_ℓ 'belonging' to class C_i (i.e. with $\ell \in C_i$). We look for a k-partition of $\mathcal{O}$ with minimum criterion

[2] The determination of an appropriate number k of classes is beyond the scope of this article.

value $g_n(\mathcal{C})$. The one-parameter optimization problem (1) is related, and even equivalent, to the two-parameter optimization problem

$$g_n(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - z_i||^2 \ \rightarrow \ \min_{\mathcal{C}, \mathcal{Z}} \tag{2}$$

where minimization is also w.r.t. all systems $\mathcal{Z} = (z_1, ..., z_k)$ of k points $z_1, ..., z_k$ from $\mathbb{R}^p$ (class representatives, class prototypes). This results from part (i) of the following theorem:

Theorem 1:
(i) For any fixed k-partition $\mathcal{C}$ the criterion $g_n(\mathcal{C}, \mathcal{Z})$ is partially minimized w.r.t. $\mathcal{Z}$ by the system of class centroids $\mathcal{Z}^* = (\overline{x}_{C_1}, ..., \overline{x}_{C_k}) =: \mathcal{Z}(\mathcal{C})$:

$$g_n(\mathcal{C}, \mathcal{Z}) \geq g_n(\mathcal{C}, \mathcal{Z}^*) = \sum_{i=1}^{k} \sum_{k \in C_i} ||x_k - \overline{x}_{C_i}||^2 \ = g_n(\mathcal{C}) \quad \textit{for all } \mathcal{Z}. \tag{3}$$

(ii) For any fixed prototype system $\mathcal{Z}$ the criterion $g_n(\mathcal{C}, \mathcal{Z})$ is partially minimized w.r.t. $\mathcal{C}$ by any minimum-distance partition $\mathcal{C}^* := (C_1^*, ..., C_k^*) =: \mathcal{C}(\mathcal{Z})$ induced by $\mathcal{Z}$, i.e. with classes given by $C_i^* := \{\ell \in \mathcal{O} \mid d(x_\ell, z_i) = \min_{j=1,...,k} d(x_\ell, z_j)\}$ $(i = 1, ..., n)$ where $d(x, z) = ||x - z||^2$ is the squared Euclidean distance:

$$g_n(\mathcal{C}, \mathcal{Z}) \geq g_n(\mathcal{C}^*, \mathcal{Z}) = \sum_{\ell=1}^{n} \min_{j=1,...,k} \{\ ||x_\ell - z_j||^2 \ \} =: \gamma_n(\mathcal{Z}) \qquad \textit{for all } \mathcal{C}. \tag{4}$$

Remark 1: The previous theorem shows that the minimization problem (2) has three essentially equivalent formulations

(A) $g_n(\mathcal{C}, \mathcal{Z}) \rightarrow \min_{(\mathcal{C}, \mathcal{Z})}$, i.e., (2),
(B) $g_n(\mathcal{C}) := g_n(\mathcal{C}, \mathcal{Z}^*) \rightarrow \min_{\mathcal{C}}$, i.e., (1) and
(C) $\gamma_n(\mathcal{Z}) := g_n(\mathcal{C}^*, \mathcal{Z}) \rightarrow \min_{\mathcal{Z}}$ ('best location problem').

All three minimum values are equal, and any solution of one of the problems generates a solution of the two other ones. *Mutatis mutandis* this same remark applies also to the two-parameter clustering criteria presented below such that each optimization problem has three equivalent formulations (A), (B), (C).

A broad range of methods has been designed in order to minimize the discrete criteria (1) and (2), either exactly or approximately. They can be roughly grouped into enumeration methods, mathematical and combinatorial programming for exact minimization (Hansen and Jaumard 1997, Grötschel and Wakabayashi 1989), integer, linear, quadratic, and dynamic programming

(Jensen 1969, Vinod 1969, Rao 1971), van Os 2000), heuristical and branch & bound methods (see also Anderberg 1973, Mulvey and Crowder 1979).

The *k-means algorithm* tries to approximate an optimum k-partition by iterating the partial minimization steps (i) and (ii) from Theorem 1, in turn. It proceeds as follows[3]:

$t = 0$: Begin with an arbitrary prototype system $\mathcal{Z}^{(0)} = (z_1^{(0)}, ..., z_k^{(0)})$.

$t \to t+1$:

(i) Minimize the criterion $g_n(\mathcal{C}, \mathcal{Z}^{(t)})$ w.r.t. the k-partition $\mathcal{C}$, i.e., determine a minimum-distance partition $\mathcal{C}^{(t+1)} := \mathcal{C}(\mathcal{Z}^{(t)})$.

(ii) Minimize the criterion $g_n(\mathcal{C}^{(t+1)}, \mathcal{Z})$ w.r.t. the prototype system $\mathcal{Z}$, i.e., calculate the system of class centroids $\mathcal{Z}^{(t+1)} := \mathcal{Z}(\mathcal{C}^{(t+1)})$.

Stopping: Iterate the steps (i) and (ii) until stationarity.

By construction, this algorithm yields a sequence $\mathcal{Z}^{(0)}, \mathcal{C}^{(1)}, \mathcal{Z}^{(1)}, \mathcal{C}^{(2)}, ...$ of prototypes and partitions with decreasing values of the criteria (1) and (2) that converge to a (typically local) minimum value.

Remark 2: In mathematical terms, the k-means algorithm is a *relaxation method* for minimizing a function of several parameters by iterative partial minimization steps (see also Mulvey and Crowder 1979), and also called an *alternating optimization* method.

Remark 3: In psychometric contexts, the SSQ criterion can be considered as the approximation error in a linear factorial model $\mathbf{X} = \mathbf{WZ} + e$. Here $\mathbf{X} = (x_1, ..., x_n)'$ is the $n \times p$ data matrix, $\mathbf{Z} = (z_1, ..., z_k)'$ the $k \times p$ matrix of class prototypes, and $\mathbf{W} = (w_{\ell i})$ the binary $n \times k$ matrix that specifies the partition $\mathcal{C}$ with $w_{\ell i} = 1$ for $\ell \in C_i$, and $w_{\ell i} = 0$ else. In fact, we have $g_n(\mathcal{C}, \mathcal{Z}) = |||\mathbf{X} - \mathbf{WZ}|||^2 = |||e|||^2 := \sum_{\ell=1}^{n} \sum_{j=1}^{p} e_{\ell j}^2$ (where $|||e|||^2$ denotes the trace norm of matrices).

Continuous SSQ criterion for space dissection: Considering $x_1, ..., x_n$ as realizations of a random vector X with distribution P in $\mathbb{R}^p$, we may formulate the following 'continuous' analogues of (1) and (2): We look for a k-partition $\mathcal{B} = (B_1, ..., B_k)$ of $\mathbb{R}^p$ with minimum value

$$g(\mathcal{B}) := \sum_{i=1}^{k} \int_{B_i} ||x - E[X|X \in B_i]||^2 \, dP(x) \;\to\; \min_{\mathcal{B}} . \tag{5}$$

As before we can relate (5) to a two-parameter optimization problem:

$$g(\mathcal{B}, \mathcal{Z}) := \sum_{i=1}^{k} \int_{B_i} ||x - z_i||^2 \, dP(x) \;\to\; \min_{\mathcal{B}, \mathcal{Z}} \tag{6}$$

and formulate the analogue of Theorem 1:

[3] This is the *batch version* of the k-means algorithm; see *Remark 4*.

Theorem 2:
(i) For any fixed k-partition $\mathcal{B}$ of $\mathbb{R}^p$ the criterion $g(\mathcal{B}, \mathcal{Z})$ is partially minimized w.r.t. $\mathcal{Z}$ by the prototype system $\mathcal{Z}^* = (z_1^*, ..., z_k^*) =: \mathcal{Z}(\mathcal{B})$ given by the conditional expectations $z_i^* := E[X|X \in B_i]$ of B_i:

$$g(\mathcal{B}, \mathcal{Z}) \geq g(\mathcal{B}, \mathcal{Z}^*) = \sum_{i=1}^{k} \int_{B_i} || \, x - E[X|X \in B_i] \, ||^2 \; = g(\mathcal{B}) \quad \textit{for all } \mathcal{Z}. \tag{7}$$

(ii) For any fixed prototype system $\mathcal{Z}$ the criterion $g(\mathcal{B}, \mathcal{Z})$ is partially minimized w.r.t. $\mathcal{B}$ by any minimum-distance partition $\mathcal{B}^* = (B_1^*, ..., B_k^*) =: \mathcal{B}(\mathcal{Z})$ generated by $\mathcal{Z}$, i.e. with classes given by $B_i^* := \{x \in \mathbb{R}^p \mid d(x, z_i) = \min_{j=1,...,k}\{d(x, z_j)\} \}$ $(i = 1, ..., n)$:

$$g(\mathcal{B}, \mathcal{Z}) \geq g(\mathcal{B}^*, \mathcal{Z}) = \int_{\mathcal{X}} \min_{j=1,...,k} \{||x - z_j||^2\} \, dP(x) =: g(\mathcal{Z}) \quad \textit{for all } \mathcal{B} \tag{8}$$

It is obvious that Theorem 2 can be used to formulate, and justify, a continuous version of the k-means algorithm. However, in contrast to the discrete case, the calculation of the class centroids might be a computational problem.

3. First instances of SSQ clustering and k-means

The first formulation of the SSQ clustering problem I know has been provided by Dalenius (1950) and Dalenius and Gurney (1951) in the framework of optimum 'proportional' stratified sampling: For estimating the expectation $\mu = E[X]$ of a real-valued random variable X with distribution density $f(x)$ (e.g., the income of persons in a city), the domain $(-\infty, +\infty)$ of X is dissected into k contiguous intervals ('strata', 'classes') $B_i = (u_{i-1}, u_i]$ $(i = 1, ..., k+1$, with $u_0 = -\infty$ and $u_{k+1} = \infty)$ and from each stratum B_i a fixed number n_i of persons is sampled where $n_i = n \cdot P(B_i)$ is proportional to the probability mass of B_i. This yields n real data $x_1, ..., x_n$. The persons ℓ with income value x_ℓ in B_i build a class C_i with class average $z_i^* := \overline{x}_{C_i}$ $(i = 1, ..., k)$. The linear combination $\hat{\mu} := \sum_{i=1}^{k} (n_i/n) \cdot \overline{x}_{C_i}$ provides an unbiased estimator of μ with variance given by the SSQ criterion: $Var(\hat{\mu}) = g(\mathcal{B})/n$. Dalenius wants to determine a k-partition $\mathcal{B}$ with minimum variance, i.e., maximum accuracy for $\hat{\mu}$ – this means the continuous clustering problem (5).

Dalenius did not use a k-means algorithm for minimizing (5), but a 'shooting' algorithm that is based on the fact that for an optimum partition $\mathcal{B}$ of $\mathbb{R}^1$ the class boundaries u_i must necessarily lie midway between the neigbouring class centroids such that $u_i = (z_i^* + z_{i+1}^*)/2$ or $z_{i+1}^* = 2u_i - z_i^*$ must hold for $i = 1, ..., k-1$. Basically, he constructs a sequence $z_1 < u_1 < z_2 < u_2 < \cdots$ of centers and boundaries by

– choosing, for $i = 1$, an initial value $z_1 \in \mathbb{R}^1$
– determining, for $i = 1$, the upper boundary u_i of $B_i = (u_{i-1}, u_i]$ from the equation $E[X|X \in B_i] = [\int_{u_{i-1}}^{u_i} x f(x) dx]/[\int_{u_{i-1}}^{u_i} f(x) dx] \stackrel{!}{=} z_i$ (the expec-

tation is an increasing function of u_i)
– then calculating the next centroid by $z_{i+1} = 2u_i - z_i$
– and iterating for $i = 2, 3, ..., k$.

By trial and error, the initial value z_1 is adapted such that the iteration stops with k classes and the k-th upper boundary $u_k = \infty$. A 'data version' of this approach for minimizing (1) has been described, e.g., by Strecker (1957), Stange (1960), and Schneeberger (1967).

Steinhaus (1956) was the first to propose explicitly the k-means algorithm in the multidimensional case. His motivation stems from mechanics (even if he refers also to examples from anthropology and industry): to partition a heterogeneous solid $\mathcal{X} \subset I\!R^p$ with internal mass distribution $f(x)$ into k subsets $B_1, ..., B_k$ and to minimize (6), i.e., the sum of the partial moments of inertia with respect to k points $z_1, ..., z_k \in I\!R^p$ by a suitable choice of the partition $\mathcal{B}$ and the z_i's. He does not only describe the (continuous version of the) k-means algorithm, but also discusses the existence of a solution for (6), its uniqueness ('minimum parfait', examples and counterexamples), and the behaviour of the sequence of minimum SSQ values for $k \to \infty$.

The first to propose the discrete k-means algorithm for clustering data in the sense of minimizing (1), was Forgy (1965)[4]. In a published form this fact was first reported by Jancey (1966a) (see also Jancey 1966b). The k-means method became a standard procedure in clustering and is known under quite different names such as *dynamic clusters method* (Diday 1971, 1973, 1974a), *iterated minimum-distance partition method* (Bock 1974), *nearest centroid sorting* (Anderberg 1973), etc.

Remark 4: The name 'k-means algorithm' was first used by MacQueen (1967), but not for the 'batch algorithm' from Section 2. Instead he used it for his sequential, 'single-pass' algorithm for (asymptotically) minimizing the continuous SSQ criterion (5) on the basis of a sequence of data points $x_1, x_2, ... \in I\!R^p$ sampled from P[5]: The first k data (objects) define k initial singleton classes $C_i^{(k)} = \{i\}$ with class centroids $z_i^{(k)} := \overline{x}_{C_i^{(k)}} = x_i$ $(i = 1, ..., k)$. Then, for $\ell = k+1, k+2, ...$, the data x_ℓ were sequentially observed and assigned to the class $C_i^{(\ell-1)}$ with closest class centroid $z_i^{(\ell-1)} := \overline{x}_{C_i^{(\ell-1)}}$ and (only) its class centroid was updated: $z_i^{(\ell)} := \overline{x}_{C_i^{(\ell)}} = z_i^{(\ell-1)} + (x_\ell - \overline{x}_{C_i^{(\ell-1)}})/|C_i^{(\ell)}|$. When stopping at some 'time' T, the minimium-distance partition $\mathcal{B}(\mathcal{Z}^{(T)})$ of $I\!R^p$ induced by the last centroid system $\mathcal{Z}^{(T)} = (\overline{x}_{C_1^{(T)}}, ..., \overline{x}_{C_k^{(T)}})$ approximates a

[4] Forgy's abstract of his talk does not explicitly mention the k-means algorithm, but details of his lecture were described by Anderberg (1973), p. 161 and MacQueen (1967) p. 294. – The more or less informal paper by Thorndike (1953) describes a sequential relocation procedure that is, however, not directly linked to his clustering criterion.

[5] This procedure has been proposed by Sebestyen (1962) as well.

(local) solution of (5) if T is large. – In many texts, the term 'k-means algorithm' is used for this single-pass procedure, but refers often to some similar sequential clustering algorithms (see, e.g, Chernoff 1970). In Späth (1975) the batch-version of k-means is called HMEANS, whereas KMEANS denotes an algorithm that exchanges single objects between classes in order to decrease (1). Hartigan (1975) uses the term 'k-means' for various algorithms dealing with k class centroids, e.g. for Späth's exchange algorithm (on page 85/86), and k-means as described in our Section 2 is one of several options mentioned on page 102 of Hartigan (1975) (see also Hartigan and Wong 1979).

In computer science and pattern recognition communities the k-means algorithm is often termed *Lloyd's algorithm I*. In fact, Lloyd (1957) considers the continuous SSQ clustering criterion (6) in $\mathbb{R}^1$ in the context of pulse-code modulation: 'Quantization' means replacing a random (voltage) signal X by a discretized approximate signal $\hat{X}$ that takes a constant value z_i ('quantum') if X belongs to the i-th class B_i of the partition $\mathcal{B} = (B_1, ..., B_k)$ of $\mathbb{R}^1$ such that $\hat{X} = z_i$ iff $X \in B_i$ $(i = 1, ..., k)$. Optimum quantification means minimization of the criterion (6). Lloyd reports the optimality of the class centroids $z_i^* = E[X|X \in B_i]$ for a fixed partition $\mathcal{B}$ and describes the one-dimensional version of the k-means algorithm as his 'Method I' whereas his 'Method II' is identical to the 'shooting method' of Dalenius.

4. Generalized k-means algorithms

The two-parameter SSQ clustering criteria (2) and (6) have been generalized in many ways in order to comply with special data types or cluster properties, and work also in a probabilistic framework. In the discrete case, typical criteria for have the two-parameter form

$$g_n(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} d(\ell, z_i) \to \min_{\mathcal{C}, \mathcal{Z}} \tag{9}$$

where $d(\ell, z)$ measures the dissimilarity between an object ℓ and a class prototype z (sometimes written as $d(x_\ell, z)$ or $d_{\ell z}$ etc., depending on the context). There is much flexibility in this approach since

(1) there is almost no constraint on the type of underlying data (quantitative and/or categorical data, shapes, relations, weblogs, DNA strains, images)
(2) there are many ways to specify a family $\mathcal{P}$ of appropriate or admissible 'class prototypes' z to represent specific aspects of the clusters (points, hyperspaces in $\mathbb{R}^p$, subsets of $\mathcal{O}$, order relations),
(3) there exists a wealth of possibilities to choose the dissimilarity measure d, and we may, additionally, introduce weights w_ℓ for the objects $\ell \in \mathcal{O}$.

In all these cases, the following *generalized k-means algorithm* can be applied in order to attain a (locally or globally) optimum configuration $(\mathcal{C}, \mathcal{Z})$:

$t = 0$: Begin with an arbitrary prototype system $\mathcal{Z}^{(0)} = (z_1^{(0)}, ..., z_k^{(0)})$.

$t \to t+1$:

(i) Minimize the criterion $g_n(\mathcal{C}, \mathcal{Z}^{(t)})$ w.r.t. the k-partition $\mathcal{C}$ from $\mathcal{P}$. Typically, this yields a minimum-distance partition $\mathcal{C}^{(t+1)} = \mathcal{C}(\mathcal{Z}^{(t)})$ with k classes $C_i^{(t+1)} := \{\ell \in \mathcal{O} \mid d(\ell, z_i^{(t)}) = \min_{j=1,...,k} d(\ell, z_j^{(t)})\ \}$.

(ii) Minimize the criterion $g_n(\mathcal{C}^{(t+1)}, \mathcal{Z})$ w.r.t. the prototype system $\mathcal{Z}$. Often, this amounts to determining, for each class $C_i = C_i^{(t+1)}$, a 'most typical configuration' $z_i^{(t+1)}$ in the sense:

$$Q(C_i, z) := \sum_{\ell \in C_i} d(\ell, z) \quad \to \quad \min_{z \in \mathcal{P}}. \tag{10}$$

Stopping: Iterate the steps (i) and (ii) until stationarity.

The first paper to propose the general criterion (9) and its generalized k-means method is Maranzana (1963): He starts from a $n \times n$ dissimilarity matrix $(d_{\ell t})$ for n factories $\ell = 1, ..., n$ in an industrial network where $d_{\ell t}$ are the minimum road transportation costs between ℓ and t. He wants to partition the set of factories into k classes $C_1, ..., C_k$ and to find a selection $\mathcal{Z} = (z_1, ..., z_k)$ of k factories as 'supply points' such that when supplying all factories of the class C_i from the supply point $z_i \in \mathcal{O}$, the overall transport costs are minimized in the sense of (9) where $d(\ell, z_i) = d_{\ell, z_i}$ means the dissimilarity between the factory (object) ℓ and the factory (supply point) $z_i \in \mathcal{O}$ (where we have omitted object-specific weights from Maranzana's formulation). So the family $\mathcal{P}$ of admissible prototypes consists of all singletons from $\mathcal{O}$ and (ii) means determining the 'most cheapest supply point' in C_i. Kaufman and Rousseeuw (1987, 1990) termed this method 'partitioning around medoids' (the *medoid* or *centrotype* of a class C_i is the most typical object in C_i in the sense of (10); see also Gordon 2000).

Many authors, including Diday (1971, 1973, 1974a) and Diday et al. (1979), have followed the generalized clustering approach via (9) in various settings and numerous variations and thereby obtained a plethora of generalized k-means algorithms (see also Bock 1996b, 1996c). For example:

– We may use Mahalanobis-type distances $||x_\ell - z_i||_Q^2$ or $||x_\ell - z_i||_{Q_i}^2$ in (1) instead of the Euclidean one, eventually including constraints for Q (Diday and Govaert 1974, 1977: *méthode des distances adaptatives*; Späth 1985, chap. 3: *determinant criterion*)

– Similarly, a L_q or Minkowski distance measure may be used. In particular, the case of the L_1 distance has been considered by Vinod (1969), Massart et al. (1983), and Späth (1975), chap. 3.5 (*k-medians algorithm*).

– Each cluster may be represented by a prototype hyperplane (instead of a single point), resulting in *principal component clustering* (Bock 1974, chap. 17; Diday and Schroeder 1974a) and in *clusterwise regression* (Bock 1969, Charles 1977, Späth 1979). For fuzzy versions of this approach see Bezdek et al. (1981).

– In the case of high-dimensional data points (i.e., with a large number

of variables) we may conjecture that a potential cluster structure is essentially concentrated on a low-dimensional (s-dimensional) hyperplane and will therefore assume that all class centers are located on the same (unknown) s-dimensional hyperplane H in $\mathbb{R}^p$. The resulting criterion

$$g_n(\mathcal{C}, H, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - z_i||^2$$
$$= \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - \overline{x}_{C_i}||^2 + \sum_{i=1}^{k} |C_i| \cdot ||\overline{x}_{C_i} - z_i||^2 \;\rightarrow \min_{\mathcal{C}, H, \mathcal{Z} \; with \; z_i \in H}$$

is minimized by a k-means-like algorithm with three iterated partial minimization steps: (i) minimizing w.r.t. $\mathcal{Z}$ (resulting in $z_i^* := proj_H(\overline{x}_{C_i})$; (ii) minimizing w.r.t. H (resulting in the hyperplane spanned by the s eigenvectors of the scatter matrix $B(\mathcal{C}) := \sum_{i=1}^{k} |C_i| \cdot (\overline{x}_{C_i} - \overline{x})(\overline{x}_{C_i} - \overline{x})'$ that belong to the s largest eigenvalues of $B(\mathcal{C})$); (iii) minimizing w.r.t. the partition $\mathcal{C}$ yielding the minimum-distance partition $\mathcal{C}$ of the projected data points $y_\ell := proj_H(x_\ell)$ generated by the center system $\mathcal{Z}$ (*projection pursuit clustering*; see Bock 1987, 1996c;, Vichi 2005);

– Another option proceeds by characterizing a class by the 'most typical subset' (pair, triple,...) of objects from this class in an appropriate sense (Diday et al. 1979).

A major step with new insight was provided by Diday and Schroeder (1974a, 1974b, 1976) and Sclove (1977) who detected that under a probabilistic 'fixed-partition' clustering model, maximum-likelihood estimation of an unknown k-partition $\mathcal{C}$ leads to a clustering criterion of the type (9) and can therefore be handled by a k-means algorithm[6]. The *fixed-partition model* considers the data $x_1, ..., x_n$ as realizations of n independent random vectors $X_1, ..., X_n$ with distributions from a density family $f(\cdot; \vartheta)$ (w.r.t. the Lebesgue or counting measure) with parameter ϑ (e.g., a normal, van Mises, loglinear,... distribution). It assumes the existence of a fixed, but unknown k-partition $\mathcal{C} = (C_1, ..., C_k)$ of $\mathcal{O}$ together with a system $\theta = (\vartheta_1, ..., \vartheta_k)$ of class-specific parameters such that the distribution of the data is class-specific in the sense that

$$X_\ell \sim f(\cdot; \vartheta_i) \qquad \text{for all } \ell \in C_i \text{ and } \ell = 1, ..., n.$$

Then maximizing the likelihood of $(x_1, ..., x_n)$ is equivalent to

$$g_n(\mathcal{C}, \theta) \;:=\; \sum_{i=1}^{k} \sum_{\ell \in C_i} [-\log f(x_\ell; \vartheta_i)] \;\rightarrow \min_{\mathcal{C}, \theta}. \tag{11}$$

Obviously this returns the former criterion (9) with $z_i \equiv \vartheta_i, \mathcal{Z} \equiv \theta$, and $d(\ell, z_i) = -\log f(x_\ell; \vartheta_i)$. The minimum-distance assignment of an object ℓ in

[6] This fact was already known before, e.g., in the case of SSQ and the normal distribution, but these authors recognized its importance for more general cases.

(i) means maximum-likelihood assignment to a class C_i, and in (ii) optimum class prototypes are given by the maximum-likelihood estimate $\hat{\vartheta}_i$ of $z_i \equiv \vartheta_i$ in C_i. By assuming a normal distribution density f, we can find probabilistic models for most of the criteria and k-means options cited above and insofar sketch the domain of application of these criteria.

A major advantage of this probabilistic approach resides in the fact that we can design meaningful clustering criteria also in the case of qualitative or binary data, yielding, *entropy clustering* and *logistic clustering* methods (Bock 1986), or models comprising random noise or outliers (Windham 200, Gallegos 2002, Gallegos and Ritter 2005). – A detailed account of these approaches is given, e.g., in Bock (1974, 1996a, 1996b, 1996c) and Diday et al. (1979).

5. Fuzzy k-means clustering

Relatively early the k-means approach has also been applied to the determination of an optimum 'fuzzy' clustering of data points $x_1, ..., x_n$ where an object ℓ is not always assigned to a single class C_i, but eventually to several classes simultaneously, with appropriate degrees of membership. If $u_{i\ell}$ denotes the membership degree of object ℓ in the 'fuzzy class' $\widetilde{U}_i$ (with $0 \leq u_{i\ell} \leq 1$ and $\sum_{i=1}^{k} u_{i\ell} = 1$), the matrix $\mathcal{U} = (u_{i\ell})$ defines a 'fuzzy partition' of the set of objects (which is a classical or 'hard' partition whenever all $u_{i\ell}$ take values 0 or 1 only). In analogy to the SSQ criterion (2) an optimum fuzzy partition is commonly defined by the 'fuzzy variance criterion'

$$\widetilde{g}(\mathcal{U}, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{\ell=1}^{n} u_{i\ell}^{r} \cdot ||x_\ell - z_i||^2 \;\rightarrow \min_{(\mathcal{U},\mathcal{Z})} \tag{12}$$

where $r > 1$ is a given exponent[7] (Bezdek and Dunn 1974, Bezdek 1981).

The 'fuzzy k-means algorithm' for solving (12) starts with an initial set of prototypes $z_1, ..., z_k$ and iterates the two following partial minimization steps:

(i) Minimize (12) w.r.t. the fuzzy partition $\mathcal{U}$ (for a given prototype system $\mathcal{Z}$). The optimum fuzzy partition $\mathcal{U}^* \equiv \mathcal{U}(\mathcal{Z})$ is given by the membership values:

$$u_{i\ell}^{*} := \frac{||x_\ell - z_i||^{-2/(r-1)}}{\sum_{i=1}^{k} ||x_\ell - z_i||^{-2/(r-1)}} = \frac{d_\ell(\mathcal{Z})}{\sum_{i=1}^{k} ||x_\ell - z_i||^{-2/(r-1)}} \tag{13}$$

where $k \cdot d_\ell(\mathcal{Z})$ is the harmonic mean of the k transformed distances $||x_\ell - z_i||^{-2/(r-1)}$ for $i = 1, ..., k$ (a proof using Jensen's inequality is provided by Bock 1979).

(ii) Minimize (12) w.r.t. the prototype system $\mathcal{Z}$ (for a given fuzzy partition

[7] For $r = 1$ minimization of (12) always results in a hard partition such that 'fuzziness' plays no role in this case; see Fisher (1958).

$\mathcal{U}$). The solution $\mathcal{Z}^* \equiv \mathcal{Z}(\mathcal{U})$ is given by the weighted class centroids

$$z_i^* \;=\; \widetilde{x}_{U_i} \;:=\; \frac{\sum_{\ell=1}^n u_{ik}^r x_\ell}{\sum_{\ell=1}^n u_{i\ell}^r} \qquad\qquad i = 1, ..., k. \tag{14}$$

It is interesting to note that by substituting, for a given $\mathcal{Z}$, the solutions $\mathcal{U}^* = \mathcal{U}(\mathcal{Z})$ and $\mathcal{Z}^{**} := \mathcal{Z}(\mathcal{U}^*)$ into the criterion (12), we obtain

$$\widetilde{\gamma}(\mathcal{Z}) := \widetilde{g}(\mathcal{U}^*, \mathcal{Z}^{**}) \;=\; \sum_{k=1}^n d_k(\mathcal{Z})^{r-1} \quad \rightarrow \quad \min_{\mathcal{Z}}. \tag{15}$$

This shows that fuzzy clustering minimizes essentially the average of the (transformed) harmonic means of the transformed Euclidean distances form the data points x_ℓ to the prototypes in $\mathcal{Z}$, by an optimum choice of $\mathcal{Z}$.

6. Convexity-based criteria and the k-tangent method

The derivation of the k-means algorithm in Section 2 shows that it relies essentially on the fact that the intuitive SSQ optimization problem (1) involving only *one* parameter $\mathcal{C}$ has an equivalent version (2) where optimization is w.r.t. *two* parameters $\mathcal{C}$ and $\mathcal{Z}$. In order to extend the domain of applicability of the k-means algorithm we may ask, more generally, if for an arbitrary (e.g., intuitively defined) one-parameter clustering criterion there exists a two-parameter version such that both resulting optimization problems are equivalent and a k-means algorithm can be applied to the second one. This problem has been investigated and solved by Windham (1986, 1987) and Bryant (1988). In the following we describe a special situation where the answer is affirmative and leads to a new *k-tangent algorithm* that provides a solution to various non-classical clustering or stratification problems (Bock 1983, 1992, 2003, Pötzelberger and Strasser 2001).

We consider the following 'convexity-based' clustering criterion for data points $x_1, ..., x_n \in I\!R^p$ that should be maximized w.r.t. the k-partition $\mathcal{C} = (C_1, ..., C_k)$:

$$k_n(\mathcal{C}) \;:=\; \sum_{i=1}^k (|C_i|/n) \cdot \phi(\overline{x}_{C_i}) \;\rightarrow\; \max_{\mathcal{C}}. \tag{16}$$

Here $\phi(\cdot)$ is a prespecified smooth convex function. Obviously (16) is a generalization of the classical SSQ clustering problem (1) since for $\phi(x) := ||x||^2$ the problem (16) reduces to (1). We may also consider a continuous version of this problem that involves a probability distribution P on $I\!R^p$ and looks for an optimal k-partiton $\mathcal{B} = (B_1, ..., B_k)$ of $I\!R^p$ in the sense:

$$k(\mathcal{B}) \;:=\; \sum_{i=1}^k P(B_i) \cdot \phi(E[X|X \in B_i]) \;\rightarrow\; \max_{\mathcal{B}}. \tag{17}$$

For $\phi(x) := ||x||^2$ this is equivalent to (5). An even more general version of this problem is given by:

$$K(\mathcal{B}) \;=\; \sum_{i=1}^{k} P(B_i) \cdot \phi(E[\lambda(X)|X \in B_i])) \;\rightarrow\; \max_{\mathcal{B}} \qquad (18)$$

where X is a random variable in a space $\mathcal{X}$ (e.g., $\mathcal{X} = I\!R^p$) with distribution P, λ is a (quite arbitrary) function $\mathcal{X} \rightarrow I\!R^q$ and ϕ is a convex function $\mathcal{X} \rightarrow I\!R^q$ (for two special choices see below). The optimum k-partition $\mathcal{B} = (B_1, ..., B_k)$ of $\mathcal{X}$ can be approximated by a k-means type algorithm if we can find a equivalent 'dual' two-parameter criterion $G(\mathcal{B}, \mathcal{Z})$ with an appropriate 'prototype' system $\mathcal{Z} = (z_1, ..., z_m) \in I\!R^q$. In fact, it has been shown by Bock (1983, 1992, 2003) that maximization of $K(\mathcal{B})$ is equivalent to the minimization problem

$$G(\mathcal{B}, \mathcal{Z}) \;:=\; \sum_{i=1}^{k} \int_{B_i} [\phi(\lambda(x)) - t(\lambda(x); z_i)] \; dP_0(x) \;\rightarrow\; \min_{\mathcal{B}, \mathcal{Z}} \qquad (19)$$

where $\mathcal{Z} = (z_1, ..., z_k) \in I\!R^p$ and $t(u; z) := \phi(z) + \phi'(z)(u - z)$ (with $u \in I\!R^q$) is the tangent (support plane) of the manifold $y = \phi(u)$ in the support point $z \in I\!R^q$ ([....] is a weighted 'volume' between the manifold and the corresponding segments of the tangents such that (19) is termed the 'minimum volume problem'). Therefore we can apply the alternating partial minimization device. The resulting method is termed 'k-tangent algorithm' and comprizes the steps:

(i) For a given support point system $\mathcal{Z}$, determine the 'maximum-tangent partition' $\mathcal{B}$ with classes defined by maximum tangent values:

$$B_i \;:=\; \{\; x \in \mathcal{X} \mid t(\lambda(x); z_i) = \max_{j=1,...,k} t(\lambda(x); z_j) \;\}$$

(ii) For a given k-partition $\mathcal{B}$ of $\mathcal{X}$ determine the system $\mathcal{Z}$ of class-specific centroids:

$$z_i \;:=\; E[\; \lambda(X)] \mid X \in B_i \;] \qquad\qquad i = 1, ..., k.$$

Iteration of (i) and (ii) yields a sequence of partitions with decreasing values in (18) and (19). - The theoretical properties of the optimum partitions for (16) and (17) have been investigated by Pötzelberger and Strasser (2001). Related approaches (dealing with Bregman distance) were proposed by Dhillon et al. (2003b) and Banerjee et al. (2004a).

We conclude with two examples that show that the k-tangent algorithm can be applied to practical problems:

(a) *Optimum discretization* and *quantization:*
There are applications where a random vector X (with two alternative distribution P_0, P_1) must be discretized into a given number of k classes $B_1, ..., B_k$

in a way that maximizes the dissimilarity between the two resulting discrete distributions $\widetilde{P}_0 = (P_0(B_1), \cdots, P_0(B_k))$ and $\widetilde{P}_1 = (P_1(B_1), \cdots, P_1(B_k))$. We may measure this dissimilarity by the classical non-centrality parameter and obtain the maximization problem:

$$K(\mathcal{B}) = \sum_{i=1}^{k} \frac{(P_1(B_i) - P_0(B_i))^2}{P_0(B_i)} = \sum_{i=1}^{k} P_0(B_i) \cdot (1 - \frac{P_1(B_i)}{P_0(B_i)})^2 \rightarrow \max_{\mathcal{B}}.$$

Writing $\phi(u) := (1-u)^2$ we see that this amounts to maximizing Csiszár's ϕ-entropy:

$$K(\mathcal{B}) = \sum_{i=1}^{k} P_0(B_i) \cdot \phi(\frac{P_1(B_i)}{P_0(B_i)}) \rightarrow \max_{\mathcal{B}}. \tag{20}$$

Since $P_1(B_i)/P_0(B_i) = E_0[\lambda(X)|X \in B_i]$ with the likelihood ratio $\lambda(x) := (dP_1)/dP_0)(x)$ this latter problem has the general form (18) with $P \equiv P_0$ and so the optimum discretization can be found by the k-tangent algorithm (Bock 1983, 1992).

(b) *Simultaneous clustering of rows and columns of a contingency table:*
A second example is provided when considering a $a \times b$ contingency table (probability distribution) $\mathcal{N} = (p_{uv})$ for two qualitative variables U, V, both with a (large) set of categories $\mathcal{U} = \{1, ..., a\}$ and $\mathcal{V} = \{1, ..., b\}$, respectively, where p_{uv} is the relative frequency (probability) of observing the case $(U, V) = (u, v)$. We consider the problem of reducing the number of categories by clustering the a rows into a m-partition $\mathcal{A} = (A_1, ..., A_m)$ with classes $A_1, ..., A_m \subset \mathcal{U}$, and *simultaneously* the b columns into a ℓ-partition $\mathcal{B} = (B_1, ..., B_\ell)$ with classes $B_1, ..., B_\ell \subset \mathcal{V}$[8]. Clustering should be performed such that row clusters and column clusters are as much dependent as possible (such that knowing, e.g., the row cluster label of an item provides optimum information on the corresponding column cluster label, and conversely).

This idea may be formalized as looking for a pair $(\mathcal{A}, \mathcal{B})$ of partitions such that the ϕ-divergence between the aggregated distribution $P^* = (P(A_i \times B_j))_{m \times \ell}$ on $\mathcal{U} \times \mathcal{V}$ with

$$P(A_i \times B_j) := P(U \in A_i, V \in B_j) = \textstyle\sum_{u \in A_i} \sum_{v \in B_j} p_{uv}$$

$(i = 1, ..., m, j = 1, ..., \ell)$ and the corresponding distribution $Q = (q_{ij})_{m \times \ell}$ under independence, i.e. with

$$q_{ij} := P(U \in A_i) \cdot P(V \in B_j) =: P_1(A_i) \cdot P_2(B_j)$$

$(i = 1, ..., m, j = 1, ..., \ell)$, is maximized:

$$g(\mathcal{A}, \mathcal{B}) := \sum_{i=1}^{m} \sum_{j=1}^{l} P_1(A_i) P_2(B_j) \cdot \phi(\frac{P(A_i \times B_j)}{P_1(A_i) P_2(B_j)}) \rightarrow \max_{\mathcal{A}, \mathcal{B}} \tag{21}$$

[8] Simultaneous clustering of rows and columns is commonly called two-way clustering, bi-clustering, co-clustering etc.

Typical minimization algorithms for simultaneous clustering proceed by starting from two initial partitions $\mathcal{A}, \mathcal{B}$ and then partially maximizing w.r.t. $\mathcal{A}$ and $\mathcal{B}$ in turn, until stationarity is obtained. In the case of (21) both steps can be performed by using the k-tangent algorithm above (Bock 2003). In fact, if we fix $\mathcal{A}$ and maximize w.r.t. $\mathcal{B}$ we can write $g(\mathcal{A}, \mathcal{B})$ in the form

$$\begin{aligned} g(\mathcal{A}, \mathcal{B}) &= \sum_{j=1}^{l} P_2(B_j) \sum_{i=1}^{m} P_1(A_i) \phi(\frac{P(B_j|A_i)}{P_2(B_j)}) \\ &= \sum_{j=1}^{l} P_2(B_j) \phi_{\mathcal{A}}(E_{P_2}[\lambda(V|\mathcal{A})]|V \in B_j) \rightarrow \max_{\mathcal{A},\mathcal{B}} \end{aligned} \tag{22}$$

where P_1, P_2 are the marginal distributions of U and V, and where we have introduced the convex function $\phi_(\mathcal{A})(x) := \sum_{i=1}^{m} P_1(A_i) \cdot \phi(x_i)$ for $x \in I\!R^m$, and the vector $\lambda(v|\mathcal{A}) := (\lambda(v|A_1), ..., \lambda(v|A_m))'$ whose components are the likelihood ratios $\lambda(v|A_i) := P(V = v|U \in A_i)/P_(V = v)$ $(i = 1, ..., m)$. This expression has the form (18) such that the k-tangent algorithm can be applied here.

A similar approach has been proposed by Dhillon (2003a) and Banerjee et al. (2004b) and applied to simultaneously clustering the documents and keywords in a database.

7. Conclusions

In this paper we have given a survey of k-means type algorithms form its origins to the modern state-of-the-art, thereby emphasizing the historical (and not the technical) aspects. It appears that this method has been proposed for solving (approximately) many specific optimization problems in cluster analysis, often within a quite non-classical framework. This underpins the importance of the k-means approach in spite of various deficiencies such a finding typically only a local (and not a global) optimum, the critical role of the initial partition, and the case of empty classes or ties. For more details on technical issues and further generalizations we point, e.g., to the articles by Steinley (2003, 2006a, 2006b).

References

Anderberg, M.R. (1973): *Cluster analysis for applications.* Academic Press, New York.

Banerjee, A., Merugu, S., Dhillon, I., Ghosh, J. (2004a): Clustering with Bregman divergences. SIAM Intern. Conf. on Data Mining (SDM) 2004, and *J. of Machine Learning Research* 6(2005) 1705-1749.

Banerjee, A., Dhillon, I., Ghosh, J., Merugu S., Modha, D.S. (2004b): A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. Conference on Knowledge Discovery and Data Mining (KDD) 2004 and *J. of Machine Learning Research* 8 (2007) 1919-1986.

Bezdek, J.C. (1981): *Pattern recognition with fuzzy objective function algorithms.* Plenum Press, New York.

Bezdek, J.C., Dunn, J.C. (1974): Optimal fuzzy partitions: a heuristic for estimating the parameters in a mixture of normal distributions. *IEEE Transactions on Computers* C-24, 835-837.

Bezdek, J.C., Coray, C., Gunderson, R., Watson, J. (1981): Detection and characterization of cluster substructure. I: Linear structure: Fuzzy c-lines. II: Fuzzy c-varieties and convex combinations thereof. *SIAM Journal on Applied Mathematics* 40, 339-357, 358-372.

Bijnen, E.J. (1973): *Cluster analysis.* Tilburg University Press, Tilburg, Netherlands.

Bock, H.-H. (1969): *The equivalence of two extremal problems and its application to the iterative classification of multivariate data.* Paper presented at the Workshop 'Medizinische Statistik', February 1969, Forschungsinstitut Oberwolfach.

Bock, H.-H. (1974): *Automatische Klassifikation. Theoretische und praktische Methoden zur Strukturierung von Daten (Clusteranalyse).* Vandenhoeck & Ruprecht, Göttingen.

Bock, H.-H. (1979): Clusteranalyse mit unscharfen Partitionen. In: H.-H. Bock (ed.): *Klassifikation und Erkenntnis III: Numerische Klassifikation.* Gesellschaft für Klassifikation, Frankfurt, 137-163.

Bock, H.-H. (1985): On some significance tests in cluster analysis. *Journal of Classification 2, 77-108.*

Bock, H.-H. (1983): *A clustering algorithm for choosing optimal classes for the chi-square test.* Bull. 44th Session of the International Statistical institute, Madrid, Contributed Papers, Vol 2, 758-762.

Bock, H.-H. (1986): Loglinear models and entropy clustering methods for qualitative data. In: W. Gaul, M. Schader (eds.): *Classification as a tool of research.* North Holland, Amsterdam, 19-26.

Bock, H.-H. (1987): On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In: H. Bozdogan, A.K. Gupta (eds.): *Multivariate statistical modeling and data analysis.* Reidel, Dordrecht, 17-34.

Bock, H.-H. (1992): A clustering technique for maximizing ϕ-divergence, noncentrality and discriminating power. In: M. Schader (ed.): *Analyzing and modeling data and knowledge.* Springer, Heidelberg, 19-36.

Bock, H.-H. (1996a): Probability models and hypotheses testing in partitioning cluster analysis. In: P. Arabie, L.J. Hubert, G. De Soete (eds.): *Clustering and classification.* World Scientific, Singapore, 377-453.

Bock, H.-H. (1996b): Probabilistic models in partitional cluster analysis. *Computational Statistics and Data Analysis 23*, 5-28.

Bock, H.-H. (1996c): Probabilistic models in cluster analysis. In: A. Ferligoj, A. Kramberger (eds.): *Developments in data analysis.* Proc. Intern. Conf. on 'Statistical data collection and analysis', Bled, 1994. FDV, Metodoloski zvezki, 12, Ljubljana, Slovenia, 3-25.

Bock, H.-H. (2003): Convexity-based clustering criteria: theory, algorithms, and applications in statistics. *Statistical Methods & Applications 12, 293-317.*

Brito, P., Bertrand, P., Cucumel, G., de Carvalho, F. (2007): *Selected contributions in data analysis and classification.* A Festschrift for E. Diday. Springer, Heidelberg.

Bryant, P. (1988): On characterizing optimization-based clustering methods. *Journal of Classification 5, 81-84.*

Charles, C. (1977): *Regression typologique.* Rapport de Recherche no. 257. IRIA, Le Chesnay.

Chernoff, H. (1970): Metric considerations in cluster analysis. In: Proc. 6^{th} Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 621-629.

Cox, D.R. (1957) Note on grouping. *J. Amer. Statist. Assoc. 52, 543-547.*

Dalenius, T. (1950): The problem of optimum stratification I. *Skandinavisk Aktuarietidskrift 1950, 203-213.*

Dalenius, T., Gurney, M. (1951): The problem of optimum stratification. II. *Skandinavisk Aktuarietidskrift 1951, 133-148.*

Dhillon, I.S., Mallela, S., Modha, D.S. (2003a): Information-theoretic co-clustering. In: Proc. 9^{th} SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005, 89-98.

Dhillon, I.S., Mallela, S., MKumar (2003b): A divisive information-theoretic feature clustering algorithm for text classification. *J. of Machine Learning Research* 3, 1265-1287.

Diday, E. (1971): Une nouvelle méthode de classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. *Revue de Statistique Appliqué* XIX (2), 1970, 19-33.

Diday, E. (1973): The dynamic clusters method in nonhierarchical clustering. *Intern. Journal of Computer and Information Sciences 2 (1)*, 61-88.

Diday, E. et al. (1979): *Optimisation en classification automatique. Vol. I, II.* Institut National der Recherche en Informatique et en Automatique (INRIA), Le Chesnay, France.

Diday, E., Govaert, G. (1974): Classification avec distance adaptative. *Comptes Rendus Acad. Sci. Paris 278 A*, 993-995.

Diday, E., Govaert, G. (1977): Classification automatique avec distances adaptatives. *R.A.I.R.O. Information/Computer Science 11 (4)*, 329-349.

Diday, E., Schroeder, A. (1974a): The dynamic clusters method in pattern recognition. In: J.L. Rosenfeld (ed.): *Information Processing 74.* Proc. IFIP Congress, Stockholm, August 1974. North Holland, Amsterdam, 691-697.

Diday, E., Schroeder, A. (1974b): *A new approach in mixed distribution detection.* Rapport de Recherche no. 52, Janvier 1974. INRIA, Le Chesnay.

Diday, E., Schroeder, A. (1976): A new approach in mixed distribution detection. *R.A.I.R.O. Recherche Opérationelle 10 (6)*, 75-106.

Engelman, L., Hartigan, J.A. (1969): Percentage points of a test of clusters. *J. American Statistical Association* 64, 1647-1648.

Fisher, W.D. (1958): On grouping for maximum heterogeneity. *J. Amer. Statist. Assoc. 53*, 789-798.

Forgy, E.W. (1965): Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometric Society Meeting, Riverside, California, 1965. Abstract in *Biometrics 21 (1965) 768.*

Gallegos, M.T. (2002): Maximum likelihood clustering with outliers. In: K. Jajuga, A. Sokolowski, H.-H. Bock (eds.): *Classification, clustering, and data analysis.* Springer, Heidelberg, 248-255.

Gallegos, M.T., Ritter, G. (2005): A robust method for cluster analysis. *Annals of Statistics 33*, 347-380.

Gordon, A. (2000): An iterative relocation algorithm for classifying symbolic data. In: W. Gaul, O. Opitz, M. Schader (eds.): *Data analysis.* Festschrift for H.-H. Bock. Springer, Heidelberg, 17-23.

Grötschel, M., Wakabayashi, Y. (1989): A cutting plane algorithm for a clustering problem. *Mathematical Programming 45*, 59-96.

Hansen, P., Jaumard, B. (1997): Cluster analysis and mathematical programming. *Mathematical Programming* 79, 191-215.

Hartigan, J.A. (1975): *Clustering algorithms.* Wiley, New York.

Hartigan, J.A., Wong, M.A. (1979): A k-means clustering algorithm. *Applied Statistics 28*, 100-108.

Jancey, R.C. (1966a): Multidimensional group analysis. *Australian J. Botany 14*, 127-130.

Jancey, R. C. (1966b): The application of numerical methods of data analysis to the genus Phyllota Benth. in New South Wales. *Australian J. Botany 14*, 131-149.

Jardine, N., Sibson, R. (1971): *Mathematical taxonomy.* Wiley, New York.

Jensen, R.E. (1969): A dynamic programming algorithm for cluster analysis. *Operations Research 17*, 1034-1057.

Kaufman, L., Rousseeuw, P.J. (1987): Clustering by means of medoids. In: Y. Dodge (ed.): *Statistical data analysis based on the L_1-norm and related methods.* North Holland, Amsterdam, 405-416.

Kaufman, L., Rousseeuw, P.J. (1990): *Finding groups in data.* Wiley, New York.

Lerman, I.C. (1970): *Les bases de la classification automatique.* Gauthier-Villars, Paris.

Lloyd, S.P. (1957): Least squares quantization in PCM. Bell Telephone Labs Memorandum, Murray Hill, NJ. Reprinted in: *IEEE Trans. Information Theory IT-28 (1982), vol. 2, 129-137.*

MacQueen, J. (1967): Some methoods for classification and analysis of mulivariate observations. In: L.M. LeCam, J. Neyman (eds.): *Proc. 5th Berkely Symp. Math. Statist. Probab. 1965/66.* Univ. of California Press, Berkely, vol. I, 281-297.

Maranzana, F.E. (1963): On the location of supply points to minimize transportation costs. *IBM Systems Journal 2*, 129-135.

Massart, D.L., Plastria, E., Kaufman, L. (1983): Non-hierarchical clustering with MASLOC. *Pattern Recognition* 16, 507-516.

Mulvey, J.M., Crowder, H.P. (1979): Cluster analysis: an application of Lagrangian relaxation. *Mangement Science 25*, 329-340.

Pötzelberger, K., Strasser, H. (2001): Clustering and quantization by MSP partitions. *Statistics and Decision 19*, 331-371.

Pollard, D. (1982): A central limit theorem for k-means clusering. *Annals of Probability 10*, 919-926.

Rao, M.R. (1971): Cluster analysis and mathematical programming. *J. Amer. Statist. Assoc. 66*, 622-626.

Schneeberger, H. (1967): Optimale Schichtung bei proportionaler Aufteilung mit Hilfe eines iterativen Analogrechners. *Unternehmensforschung 11*, 21-32.

Sclove, S.L. (1977): Population mixture models and clustering algorithms. *Commun. in Statistics, Theoretical Methods, A6*, 417-434.

Sebestyen, G.S. (1962): *Decision making processes in pattern recognition.* Macmillan, New York.

Sodeur, W. (1974): *Empirische Verfahren zur Klassifikation.* Teubner, Stuttgart.

Sokal, R.R., Sneath, P. H. (1963): *Principles of numerical taxonomy.* Freeman, San Francisco - London.

Späth, H. (1975): *Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion.* Oldenbourg Verlag, München - Wien. English translation: Cluster analysis algorithms for data reduction and classification of objects. Ellis Horwood Ltd., Chichester, UK, 1980.

Späth, H. (1979): Algorithm 39: Clusterwise linear regression. *Computing 22*, 367-373. Correction in *Computing 26 (1981)*, 275.

Späth, H. (1985): *Cluster dissection and analysis.* Wiley, Chichester.

Stange, K. (1960): Die zeichnerische Ermittlung der besten Schätzung bei proportionaler Aufteilung der Stichprobe. *Zeitschrift für Unternehmensforschung 4*, 156-163.

Steinhaus, H. (1956): Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe III, vol. IV, no. 12*, 801-804.

Steinley, D. (2003): Local optima in k-means clustering: what you don't know may hurt you. *Psychological Methods* 8, 294-304.

Steinley, D. (2006a): K-means clustering: a half-century synthesis. *British J. on Mathematical and Statistical Psychology* 59, 1-34.

Steinley, D. (2006b): Profiling local optima in k-maens clustering: developing a diagnostic technique. *Psychological Methods* 11, 178-192.

Strecker, H. (1957): *Moderne Methoden in der Agrarstatistik.* Physica, Würzburg, p. 80 etc.

Thorndike, R.L. (1953): Who belongs to the family? *Psychometrika* 18, 267-276.

van Os, B.J. (2000): *Dynamic programming for partitioning in multivariate data analysis.* Leiden University Press, Leiden, The Netherlands.

Vichi, M. (2005): Clustering including dimensionality reduction. In: D. Baier, R. Decker, L. Schmidt-Thieme (eds.): *Data analysis and decision support.* Springer, Heidelberg, 149-156.

Vinod, H.D. (1969): Integer programming and the theory of grouping. *J. Amer. Statist. Assoc. 64*, 506-519.

Vogel, F. (1975): *Probleme und Verfahren der numerischen Klassifikation.* Vandenhoeck & Ruprecht, Göttingen.

Windham, M.P. (1986): A unification of optimization-based clustering algorithms. In: W. Gaul, M. Schader (eds.): *Classification as a tool of research.* North Holland, Amsterdam, 447-451.

Windham, M.P. (1987): Parameter modification for clustering criteria. *J. of Classification 4*, 191-214.

Windham, M.P. (2000): Robust clustering. In: W. Gaul, O. Opitz, M. Schader (eds.): *Data analysis.* Festschrift for H.-H. Bock. Springer, Heidelberg, 385-392.