Minimization subproblems and heuristics for an applied clustering problem

E. G. Birgin [*] J. M. Martínez [†] D. P. Ronconi [‡]

December 18, 2001

Abstract

A practical problem that requires the classification of a set of points of $\mathbb{R}^n$ using a criterion not sensitive to bounded outliers is studied in this paper. A fixed-point (k-means) algorithm is defined that uses an arbitrary distance function. Finite convergence is proved. A robust distance defined by Boente, Fraiman and Yohai is selected for applications. Smooth approximations of this distance are defined and suitable heuristics are introduced to enhance the probability of finding global optimizers. A real-life example is presented and commented.

Key words: Nonlinear programming, heuristics, clustering, classification, fixed points.

[*] Department of Computer Science IME-USP, University of São Paulo, Rua do Matão 1010, Cidade Universitária, 05508-900, São Paulo SP, Brazil. This author was supported by PRONEX-Optimization 76.79.1008-00 and FAPESP (Grants 99-08029-9 and 01-04597-4). Corresponding author. FAX: +55(11)3818-6134. e-mail: egbirgin@ime.usp.br

[†] Department of Applied Mathematics, IMECC-UNICAMP, University of Campinas, CP 6065, 13081-970 Campinas SP, Brazil. This author was supported by PRONEX-Optimization 76.79.1008-00, FAPESP (Grant 01-04597-4), CNPq and FAEP-UNICAMP. e-mail: martinez@ime.unicamp.br

[‡] Department of Production Engineering, EP-USP, University of São Paulo, Av. Prof. Almeida Prado, 128, Cidade Universitária, 05508-900, São Paulo SP, Brazil. This author was supported by FAPESP (Grant 00-01715-3). e-mail: dronconi@usp.br

1 Introduction

This research has a practical motivation. We wish to classify students into different groups regarding their application to different training programs. The parameters used for classification are the scores in a set of tests and exams. We observed that "outliers" among these scores are frequent. They are probably due, on one side, to occasional faking and, on the other side, to illness or stress disorders. Outliers cause misclassifications when the ordinary Euclidean distance is used in the context of a fixed-point procedure. Since all the observations are scores between 0 and 10 there is a severe restriction to the distribution of outliers. The probability of their occurrence outside a fixed interval is null. This motivated us to seek a different "robust" distance, which should be less sensitive to this type of outliers. With that purpose, we chose a distance introduced by Boente, Fraiman and Yohai in [9]. The new distance seems to fit our objectives better than other L_1-like alternatives. See [29]–[32].

Fixed-point procedures for classification are largely known in classical literature. They are generally known as k-means algorithms (see [18, 23] and many others). The main properties of these algorithms can be found in modern literature. See, for example, [1, 2, 8, 11, 15] and references therein.

Suppose that we want to classify m points of $I\!R^n$ into q different clusters. Given q arbitrary subsets, the center of gravity of each group is computed. Then, the clusters are reorganized in such a way that each point belongs to the set defined by the closest center. This procedure is repeated until a repetition in classification occurs. In k-means algorithms, the center of gravity is the point that minimizes the sum of squared Euclidean distances and, therefore, is quite sensitive to the presence of outliers. Using a continuous arbitrary distance function, the fixed-point algorithm can be generalized and it can be proved that finite convergence to a local minimizer of an adequate merit function takes place under a regularity assumption. Regularity is essential for obtaining convergence.

The general algorithm will be analyzed in connection to the Boente-Fraiman-Yohai (BFY) distance function (see [9]). In order to smooth the function we use a half-Gaussian approximation of Heaviside step functions. We will prove that, when the smoothing parameter tends to its limit, the smoothed problem produces the same results as the original BFY function.

In order to illustrate the advantages of using the BFY distance for clustering in the presence of outliers let us give an example. Assume that $P_1 = (0,0,0,0,0), P_2 = (0,0,0,0,10), P_3 = (0,1,2,3,4)$ and that we wish to classify these points into two groups. Suppose that the points represent the scores of three students in a course with five exams. So, P_i^j is the score of student i in exam j. The "reasonable" solution is $C_1 = \{P_1, P_2\}, C_2 =$

$\{P_3\}$. In fact, the fifth score of P_2 is, very likely, an outlier due, perhaps, to fake. This is the solution found by the algorithm that we are going to present here, that uses the BFY distance. On the other hand, k-means and k-median algorithms with the Euclidean and the 1-norm, respectively, classify $C_1 = \{P_1\}, C_2 = \{P_2, P_3\}$. It is worth mentioning that the efficiency of the BFY approach is not independent of scaling considerations. In fact, in this case it is essential that all the measurements lie between 0 and 10. If the scaling of the variables is very different, the BFY function must be scaled as well.

The fact that the fixed-point method stops only at local minimizers motivated us to find suitable heuristics to determine an initial classification, previous to the centering cycles. The final algorithm combines heuristics and fixed-point iterations (in the outer stage) with centering steps based on BFY minimizations (in the inner stage). For finding the centers we used an algorithm recently introduced in [6]. See, also [24, 25]. Gradient algorithms for clustering problems have also been proposed in [22], where deterministic annealing plays the role of our heuristic for improving global properties.

The algorithm was applied to practical situations where it turned out to be efficient. We describe one of these situations in the present research.

The development of this paper follows the sequence sketched above. Section 2 contains the convergence proof of the generalized fixed-point method. Section 3 is devoted to the approximation of the BFY distance function. In Section 4 we describe the heuristics for the initial classification and the application is presented in Section 5. The last section contains final remarks.

2 The fixed-point procedure

The results of this section seem to be well known in the classical literature. See, for example, [1] and the references of this book. We give a brief survey of them for future reference.

Assume that $P_1, \ldots, P_m$ are points in $I\!R^n$ which we wish to classify into q groups. The idea is to determine $C_1, \ldots, C_q$, the "centers" of the groups, in an optimal way. The point P_i will be assigned to the group whose center is C_j if a distance-like continuous function $\varphi(P_i, C)$ takes its minimum value at C_j. We assume that $\varphi(P, C) \geq 0$ for all $P, C, \in I\!R^n$. Therefore, the goal is to find $C_1, \ldots, C_q$ that solves the optimization problem

$$\text{Minimize } f(C_1, \ldots, C_q) \equiv \sum_{i=1}^{m} \text{ minimum } \{\varphi(P_i, C_j), j = 1, \ldots, q\}. \tag{1}$$

Problem (1) is nonsmooth and nonconvex. So, its resolution by means of standard optimization algorithms can be very hard. The general k-means method is an algorithm of fixed-point type for solving it. Given $(C_1, \ldots, C_q) \in I\!R^{n \times q}$ we define $F(C_1, \ldots, C_q) = (C'_1, \ldots, C'_q)$ by means of:

(a) For all $j = 1, \ldots, q$, define $\mathcal{F}_j$ saying that $P_i \in \mathcal{F}_j$ if j is the smallest index k in $\{1, \ldots, q\}$ such that

$$\varphi(P_i, C_k) \leq \varphi(P_i, C_\ell) \text{ for all } \ell = 1, \ldots, q. \tag{2}$$

(b) Compute, for all $j = 1, \ldots, q$ such that $\mathcal{F}_j \neq \emptyset$,

$$C'_j = \operatorname{Arg}\min_C \sum_{P_k \in \mathcal{F}_j} \varphi(C, P_k). \tag{3}$$

If $\mathcal{F}_j = \emptyset$ we define $C'_j = C_j$.

The set $\mathcal{F}_j$ contains the points P_i that have C_j as its closest "center", deciding for the one with smallest index in case of equal "distances". $\mathcal{F}_j$ will be called the "influence set" of C_j. Clearly, the $\mathcal{F}_j$'s are disjoint and

$$\mathcal{F}_1 \cup \ldots \cup \mathcal{F}_q = \{P_1, \ldots, P_m\}. \tag{4}$$

The computation of C'_j involves the solution of the optimization problem

$$\text{Minimize} \sum_{P_k \in \mathcal{F}_j} \varphi(C, P_k). \tag{5}$$

We assume that this problem is always solvable. In fact, in some specific situations it can be very simple. For example, if $\varphi(P, C) = \|P - C\|^2$ and $\|\cdot\|$ is the Euclidean norm, it is easy to see that

$$C'_j = \frac{1}{n_j} \sum_{P_k \in \mathcal{F}_j} P_k, \tag{6}$$

where n_j is the number of elements of $\mathcal{F}_j$.

Given the current approximation $(C_1^k, \ldots, C_q^k)$ to the solution of (1), the algorithm computes $(C_1^{k+1}, \ldots, C_q^{k+1})$ by means of

$$(C_1^{k+1}, \ldots, C_q^{k+1}) = F(C_1^k, \ldots, C_q^k), \tag{7}$$

where F is defined by (a) and (b).

Lemma 1 states that the objective function f is nonincreasing at the successive iterations of the fixed-point algorithm.

Lemma 1. *If* $(C'_1,\ldots,C'_q)=F(C_1,\ldots,C_q)$ *then* $f(C'_1,\ldots,C'_q)\leq f(C_1,\ldots,C_q)$.

Lemma 2 says that the fixed-point algorithm necessarily finishes after a finite number of iterations repeating always the same functional value.

Lemma 2. *There exists* $k_0\in\{0,1,2,\ldots\}$ *such that*

$$f(C_1^k,\ldots,C_q^k)=f(C_1^{k_0},\ldots,C_q^{k_0})$$

for all $k\geq k_0$.

It is natural to ask whether this implies that $(C_1^k,\ldots,C_q^k)=(C_1^{k_0},\ldots,C_q^{k_0})$ for all $k\geq k_0$. Clearly, if problem (5) admits more than one minimizer, this can be false. However, it can be proved that the $q-$uple of centers necessarily repeats for all k large enough under the assumption that (5) admits a unique solution.

Theorem 1. *Assume that, for all* $\mathcal{F}\subset\{P_1,\ldots,P_m\}$, *the solution of (5) is unique. Then, there exists* $k\in\{0,1,2,\ldots\}$ *such that* $(C_1^k,\ldots,C_q^k)$ *is a fixed point of* F.

We saw that the fixed-point algorithm converges, starting from any initial point and in a finite number of iterations, to a fixed point of F. It is interesting to observe that neither continuity nor positivity of φ need to be used for that purpose. In the following theorem, we characterize the fixed points of F. In this case the continuity of φ will be used. A previous definition will be necessary: A fixed point $(C_1,\ldots,C_q)$ of F will be said to be *regular* if its influence sets are given by

$$\mathcal{F}_j=\{P_i\mid\varphi(P_i,C_j)<\varphi(P_i,C_k)\quad\text{for all}\quad k=1,\ldots,q,k\neq j\}\tag{8}$$

for $j=1,2,\ldots,q$. In other words, the fixed point is regular if the lower-index decision is not necessary when some P_i is assigned to some center C_j.

Theorem 2. *If* $(C_1,\ldots,C_q)$ *is a regular fixed point of* F, *then it is a local minimizer of* f.

The property stated in Theorem 2 is not true, in general, if the fixed point $(C_1,\ldots,C_q)$ is not regular. In fact, consider the following counter-example: $m=3$, $n=1$, $q=2$, $P_1=0$, $P_2=2$, $P_3=-1$, $C_1=1$, $C_2=-1$. By the rule of the lower index, with $\varphi(C,P)=|C-P|^2$, $\mathcal{F}_1=\{P_1,P_2\}$ and $\mathcal{F}_2=\{P_3\}$. Clearly (C_1,C_2) is a non-regular fixed point of F since $C_1=(P_1+P_2)/2$ and $C_2=P_3$. We have $f(C_1,C_2)=2$. However, taking $\varepsilon>0$ we obtain $f(C_1+\varepsilon,C_2)=(1-\varepsilon)^2+1$. Therefore, (C_1,C_2) is not a local minimizer.

3 Subproblems with the BFY distance function

The implementation of the fixed-point algorithm requires the solution of q subproblems (5) at each iteration. For simplicity, and without loss of generality, let us write (5) as

$$\text{Minimize } g(C) \tag{9}$$

where

$$g(C) = \sum_{i=1}^{m} \varphi(C, P_i). \tag{10}$$

The difficulty of (9) depends on the definition of $\varphi(C, P)$. If $\varphi(C, P)$ is the squared Euclidean distance between C and P, subproblem (9) is trivial. However, that choice of φ is not satisfactory, and it is better to consider a function less sensitive to outliers. With that purpose, we consider here the distance introduced by Boente, Fraiman and Yohai [9]:

$$\varphi(C, P) = \text{ Infimum } \{\varepsilon \mid \#\{k \mid |C^k - P^k| \geq \varepsilon\} \leq n\varepsilon\}, \tag{11}$$

where C^k and P^k denote the $k-$th coordinate of C and P respectively and $\#$ denotes the number of elements of a finite set. This function is continuous and satisfies

$$\varphi(C, P) \leq 1 \quad \text{for all} \quad C, P \in I\!R^n. \tag{12}$$

According to this definition, the distance between C and P is small if all the coordinates of $C - P$ except a small fraction are close to zero (see [9]). It is important to mention that scaling considerations are implicit in the assertion above. This can be understood even in the one-dimensional case. In this case $\varphi(C, P) = |C - P|$ if $|C - P| \leq 1$ and $\varphi(C, P) = 1$ otherwise. This way of measuring distances might not be reasonable in many situations but it is under the range of scaling of the (bounded) variables that we consider in our applications. It is easy to see that, when the coordinates of $C - P$ are integers, the function $\varphi(C, P)$ is $\nu(C, P)/n$, where

$$\nu(C, P) = \#\{k \mid C^k \neq P^k\}.$$

From now on, we consider always $g(C)$ associated to the distance (11). The following results are directed to justify the resolution of (9) by means of standard optimization techniques.

Theorem 3

$$\text{Infimum } \{g(C) \mid C \in I\!R^n\}$$

$$= \text{Infimum } \{\sum_{i=1}^{m} z_i \mid \#\{k \mid |C^k - P_i^k| \geq z_i\} \leq n z_i, i = 1, \ldots, m\}$$

Proof. Assume that $C \in I\!R^n$ and $z_1, \ldots, z_m$ are such that $\#\{k \mid |C^k - P_i^k| \geq z_i\} \leq n z_i\}$. Therefore, by (11), $\varphi(C, P_i) \leq z_i$ for all $i = 1, \ldots, m$. So, $g(C) = \sum_{i=1}^{m} \varphi(C, P_i) \leq \sum_{i=1}^{m} z_i$ and, consequently,

$$\text{Infimum } \{g(C) \mid C \in I\!R^n\} \leq$$

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \#\{k \mid |C^k - P_i^k| \geq z_i\} \leq n z_i, i = 1, \ldots, m\}. \tag{13}$$

Now, suppose that $C \in I\!R^n$ and $\varepsilon > 0$. Define, for $i = 1, \ldots, m$, $z_i = \varphi(C, P_i) + \varepsilon/m$. The definition (11) implies that $\#\{k \mid |C^k - P_i^k| \geq z_i\} \leq n z_i$ and, clearly, $\sum_{i=1}^{m} z_i \leq \sum_{i=1}^{m} \varphi(C, P_i) + \varepsilon = g(C) + \varepsilon$. This implies that

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \#\{k \mid |C^k - P_i^k| \geq z_i\} \leq n z_i, i = 1, \ldots, m\} \leq$$

$$\text{Infimum } \{g(C) \mid C \in I\!R^n\} + \varepsilon. \tag{14}$$

Since (14) holds for arbitrary $\varepsilon > 0$, the desired result follows from this inequality and (13). □

Theorem 3 justifies the definition of the following constrained optimization problem which, in fact, has been proved to be equivalent to (9) for the estimation purposes.

$$\text{Minimize } \sum_{i=1}^{m} z_i \quad \text{subject to} \quad \#\{k \mid |C^k - P_i^k| \geq z_i\} \leq n z_i,\ i = 1, \ldots, m. \tag{15}$$

Let us define now the classical Heaviside step function:

$$H(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to see that (15) can be written as

$$\text{Minimize } \sum_{i=1}^{m} z_i \quad \text{subject to} \quad \sum_{k=1}^{n} H(|C^k - P_i^k| - z_i) \leq n z_i, i = 1, \ldots, m. \tag{16}$$

Now, let H_ℓ be a sequence of bounded, nondecreasing and non-negative continuous functions that converges to H in the sense that

$$\lim_{\ell \to \infty} H_\ell(x) = H(x) \quad \text{for all } x \in I\!R$$

and

$$H_\ell(x) \geq H(x) \quad \text{for all } x \in I\!R.$$

Theorem 4

$$\lim_{\ell\to\infty} \text{ Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \leq nz_i, i = 1,\ldots,m\} =$$

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \#\{k \mid |C^k - P_i^k| \geq z_i\} \leq nz_i, i = 1,\ldots,m\}.$$

Proof. Assume that $C, z_1, \ldots, z_m$ are such that

$$\sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \leq nz_i, i = 1,\ldots,m.$$

Since $H_\ell(x) \geq H(x)$ for all $x \in I\!R$, we have that

$$\sum_{k=1}^{n} H(|C^k - P_i^k| - z_i) \leq nz_i, i = 1,\ldots,m,$$

so,

$$\#\{k \mid |C^k - P_i^k| \geq z_i\} \leq nz_i, i = 1,\ldots,m\}.$$

Therefore,

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \#\{k \mid |C^k - P_i^k| \geq z_i\} \leq nz_i, i = 1,\ldots,m\} \leq$$

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \leq nz_i, i = 1,\ldots,m\} \tag{17}$$

for all $\ell = 0, 1, 2, \ldots$.

Now, let us define

$$s = \text{ Infimum } \{\sum_{i=1}^{m} z_i \mid \#\{k \mid |C^k - P_i^k| \geq z_i\} \leq nz_i, i = 1,\ldots,m\}$$

and let ε be an arbitrary positive number. (Since $g(C) \geq 0$ for all $C \in I\!R^n$ the infimum s is not $-\infty$.) Let $C, z_1, \ldots, z_m$ be such that

$$\#\{k \mid |C^k - P_i^k| \geq z_i\} \leq nz_i, i = 1,\ldots,m$$

and

$$\sum_{i=1}^{m} z_i \leq s + \varepsilon/2.$$

Define

$$\bar{z}_i = z_i + \varepsilon/(2m) \text{ for all } i = 1, \dots, m.$$

Therefore, from $\sum_{i=1}^{m} H(|C^k - P_i^k| - z_i) \le n z_i$ it follows that

$$\sum_{i=1}^{m} H(|C^k - P_i^k| - \bar{z}_i) < n\bar{z}_i.$$

Since $H_\ell(|C^k - P_i^k| - \bar{z}_i) \to H(|C^k - P_i^k| - \bar{z}_i)$, there exists $\ell_0 \in \{0, 1, 2, \dots\}$ such that for all $\ell \ge \ell_0$,

$$\sum_{i=1}^{m} H_\ell(|C^k - P_i^k| - \bar{z}_i) < n\bar{z}_i.$$

But

$$\sum_{i=1}^{m} \bar{z}_i = \sum_{i=1}^{m} z_i + \varepsilon/2 \le s + \varepsilon,$$

therefore, by (17), for all $\ell \ge \ell_0$,

$$s \le \text{ Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \le n z_i, i = 1, \dots, m\} \le s + \varepsilon. \tag{18}$$

Since $\varepsilon > 0$ was arbitrary, this implies the thesis of the theorem. □

The theorems above justify the consideration of the family of subproblems

$$\text{Minimize } \sum_{i=1}^{m} z_i \quad \text{subject to} \quad \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \le n z_i, i = 1, \dots, m. \tag{19}$$

However, we are going to prove an additional result that shows that the resolution of (9) by means of auxiliary continuous problems admits further simplifications.

Theorem 5

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \le n z_i, i = 1, \dots, m\} =$$

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) = n z_i, i = 1, \dots, m\}$$

for all $\ell \in \{0, 1, 2, \dots\}$.

Proof. Clearly,

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \le n z_i, i = 1, \dots, m\} \le$$

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) = nz_i, i = 1, \ldots, m\}.$$

Now, let us assume that $C, z_1, \ldots, z_m$ are such that

$$\sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \le nz_i, i = 1, \ldots, m$$

and that, for some i, $H_\ell(|C^k - P_i^k| - z_i) < nz_i$. So, defining

$$\beta(z) = \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z) - nz$$

we have that β is continuous, $\beta(z_i) < 0$ and $\beta(0) \ge 0$. This implies that there exists $y_i \in [0, z_i)$ such that $\beta(z) = 0$. Repeating this reasoning for all i such that $\sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) < nz_i$ and defining $y_i = z_i$ in the remaining cases, we see that

$$\sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - y_i) = ny_i$$

for all $i = 1, \ldots, m$ and $\sum_{i=1}^{m} y_i < \sum_{i=1}^{m} z_i$. So,

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) \le nz_i, i = 1, \ldots, m\} \ge$$

$$\text{Infimum } \{\sum_{i=1}^{m} z_i \mid \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) = nz_i, i = 1, \ldots, m\}$$

and the proof is complete. □

Therefore, by the results proved above, the original problem (9) can be approximated by the problems defined below:

$$\text{Minimize } \sum_{i=1}^{m} z_i \quad \text{subject to} \quad \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) - nz_i = 0, i = 1, \ldots, m. \tag{20}$$

Now, given $C \in I\!R^n, i \in \{1, \ldots, m\}$, consider, as in the proof of Theorem 5, the function $\beta(z) = \sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z) - nz$. Clearly, $\beta(0) \ge 0$, $\beta(z) < 0$ if z is large enough and, finally, $\beta(z)$ is strictly decreasing. Therefore, for all $C \in I\!R^n$, $i \in \{1, \ldots, m\}$, there exists a unique z_i such that $\sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z_i) - nz_i = 0$. This z_i is easy to compute using standard numerical procedures. Assuming that H_ℓ is differentiable, a safeguarded Newton's method surely finds z_i in few steps (see [12, 21]). Let us call $\xi_{i,\ell}(C)$ the unique value of z that verifies $\sum_{k=1}^{n} H_\ell(|C^k - P_i^k| - z) - nz = 0$. Then, (20) reduces to

$$\text{Minimize } \sum_{i=1}^{m} \xi_{i,\ell}(C). \tag{21}$$

Unfortunately, the objective function of (21) is not differentiable at the points C such that $C^k = P_i^k$ for some i, k. For overcoming this problem, let us assume from now on that H_ℓ is differentiable and $|H_\ell'(t)| \leq c_\ell$ for all $t \in I\!R$. Moreover, assume that $\alpha(t)$ is a differentiable function such that $|\alpha(t) - |t|| \leq \varepsilon$ for all $t \in I\!R$. The theorem below allows us to consider a differentiable problem.

Theorem 6

Under the assumptions above, there exists $\eta = \eta_\ell$ such that $|\eta| \leq nc_\ell\varepsilon$ and

$$\text{Infimum } \{\sum_{i=1}^m z_i \mid \sum_{k=1}^n H_\ell(|C^k - P_i^k| - z_i) = nz_i, i = 1, \ldots, m\} =$$

$$\text{Infimum } \{\sum_{i=1}^m z_i \mid \sum_{k=1}^n H_\ell(\alpha(C^k - P_i^k) - z_i) = nz_i, i = 1, \ldots, m\} + \eta.$$

Proof. Let C, z_i, w_i be such that

$$\sum_{k=1}^n H_\ell(|C^k - P_i^k| - z_i) = nz_i, i = 1, \ldots, m$$

and

$$\sum_{k=1}^n H_\ell(\alpha(C^k - P_i^k) - w_i) = nw_i, i = 1, \ldots, m.$$

The existence and unicity of w_i is guaranteed by the same arguments that guarantee the existence and unicity of z_i. Moreover, the equation

$$\sum_{k=1}^n H_\ell(x_k - z) = nz, i = 1, \ldots, m. \tag{22}$$

defines z as a function of x. Differentiation with respect to x_j gives

$$\sum_{k \neq j} H_\ell'(x_k - z)(-\frac{\partial z}{\partial x_j}) + H_\ell'(x_j - z)(1 - \frac{\partial z}{\partial x_j}) = n\frac{\partial z}{\partial x_j}.$$

So,

$$\sum_{k=1}^n H_\ell'(x_k - z)(-\frac{\partial z}{\partial x_j}) + H_\ell'(x_j - z) - n\frac{\partial z}{\partial x_j} = 0.$$

Therefore,

$$(-\frac{\partial z}{\partial x_j})[\sum_{k=1}^n H_\ell'(x_k - z) + n] = -H_\ell'(x_j - z).$$

and

$$\frac{\partial z}{\partial x_j} = \frac{H_\ell'(x_j - z)}{\sum_{k=1}^n H_\ell'(x_k - z) + n}.$$

Since $H'_\ell(t) \geq 0$ and $|H'_\ell(t)| \leq c_\ell$ for all $t \in I\!R$, this implies that

$$|\frac{\partial z}{\partial x_j}| \leq c_\ell / n,$$

for all $j = 1, \ldots, n$. Therefore, since $||C_i^k - P_i^k| - \alpha(C_i^k - P_i^k)| \leq \varepsilon$, the Mean Value theorem guarantees that

$$|z_i - w_i| \leq c_\ell \varepsilon.$$

This implies that the thesis holds. □

Therefore, the solution of (20) can be approximated by the solution of the smooth problem

$$\text{Minimize } \sum_{i=1}^{m} z_i \quad \text{subject to} \quad \sum_{k=1}^{n} H_\ell(\alpha(C^k - P_i^k) - z_i) - n z_i = 0, i = 1, \ldots, m. \tag{23}$$

or, as in the deduction of (21),

$$\text{Minimize } \sum_{i=1}^{m} \kappa_{i,\ell}(C) \tag{24}$$

where $\kappa_{i,\ell}(C)$ is the the unique value of z_i that verifies

$$\sum_{k=1}^{n} H_\ell(\alpha(C^k - P_i^k) - z_i) - n z_i = 0, i = 1, \ldots, m.$$

The problem that we are going to solve to find the centers is the one given by (24). The functions H_ℓ and α will be chosen in such a way that the hypotheses of the theorems of this section hold. So, we choose

$$H_\ell(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ e^{(-x/\ell)^2}, & \text{otherwise,} \end{cases}$$

and

$$\alpha(x) = \sqrt{x^2 + \varepsilon}.$$

4 Heuristics for the initial classification

We have two motivations to define an adequate heuristics for an initial classification. On one hand, the fixed-point method converges to local minimizers and, very likely, the heuristics can help to avoid some local-nonglobal minimal points. On the other hand, the general functions used for finding centers need not to be convex and, so, false centers can arise from the application of gradient-like algorithms.

Our heuristic approach gives an initial clustering based on classical methods for a well known production problem: construction of families of jobs and machines based on their path of production. This problem comes from the application of Technology Group ideas and was addressed using similarity-coefficient techniques by many authors. See [16, 17, 20, 28, 26, 27]. Following [17], the re-assignment of data points in the algorithm is done in a serial fashion.

The heuristic approach has two phases: (i) initial clustering and (ii) improvement by relocation. The number of points m, the set of points $\{P_1, P_2, ..., P_m\}$ and the number of desired clusters q are given. In the algorithm described below the functions D and Φ are arbitrary and several possibilities will be defined latter.

Phase 1: *Initial clustering*

Step 0: *Initialization*

Consider the m clusters $\mathcal{F}_i = \{P_i\}, i = 1, 2, \ldots, m$, and
set $dist(\mathcal{F}_i, \mathcal{F}_j) \leftarrow \varphi(P_i, P_j)$ for all $1 \leq i, j \leq m$.

Step 1: *Stopping criteria*

If the number of clusters is equal to q stop.

Step 2: *Shrinking clusters*

Find the pair of nearest clusters $(\mathcal{F}_r, \mathcal{F}_s), r \neq s$,
create (deleting $\mathcal{F}_r$ and $\mathcal{F}_s$) a new cluster $\mathcal{F}_t = \mathcal{F}_r \cup \mathcal{F}_s, t = \min\{r, s\}$, and
for all $i \neq t$ compute $dist(\mathcal{F}_t, \mathcal{F}_i) \leftarrow D(\mathcal{F}_s, \mathcal{F}_r, \mathcal{F}_i)$.

Step 3: Go to Step 1.

Phase 2: *Re-assignment of data points*

Step 1: *Local reduction*

For each point P_i, let $\mathcal{F}_r$ be the cluster to which P_i belongs.

Step 1.1:

For each $\ell = 1, \ldots, q$, move P_i from $\mathcal{F}_r$ to $\mathcal{F}_\ell$, compute $\gamma(\ell) = \sum_{j=1}^{q} \Phi(\mathcal{F}_j)$ after this change and return P_i to $\mathcal{F}_r$. Let $s \in \{1, \ldots, q\}$ be the minimum index minimizer of $\{\gamma(1), \ldots, \gamma(q)\}$.

Step 1.2:

If $s \neq r$, redefine $\mathcal{F}_r \leftarrow \mathcal{F}_r - \{P_i\}$ and $\mathcal{F}_s \leftarrow \mathcal{F}_s \cup \{P_i\}$.

Step 2: *Stopping criteria*

If Step 1 did not produce any change of cluster, stop.

Step 3: Go to Step 1.

Different algorithms come from different definitions of D (at Phase 1) and Φ (at Phase 2). Some possibilities (D_1, D_2, D_3) for the definition of D are given below.

$$D_1(\mathcal{F}_s, \mathcal{F}_r, \mathcal{F}_i) = [(\#\mathcal{F}_r) dist(\mathcal{F}_r, \mathcal{F}_i) + (\#\mathcal{F}_s) dist(\mathcal{F}_s, \mathcal{F}_i)]/[(\#\mathcal{F}_r) + (\#\mathcal{F}_s)], \tag{25}$$

$$D_2(\mathcal{F}_s, \mathcal{F}_r, \mathcal{F}_i) = (dist(\mathcal{F}_r, \mathcal{F}_i) + dist(\mathcal{F}_s, \mathcal{F}_i))/2. \tag{26}$$

$$D_3(\mathcal{F}_s, \mathcal{F}_r, \mathcal{F}_i) =$$

$$\begin{cases} \min\{dist(\mathcal{F}_s, \mathcal{F}_i), dist(\mathcal{F}_r, \mathcal{F}_i)\}, & \text{if } dist(\mathcal{F}_s, \mathcal{F}_i) \le 0.5 \text{ and } dist(\mathcal{F}_r, \mathcal{F}_i) \le 0.5 \\ \max\{dist(\mathcal{F}_s, \mathcal{F}_i), dist(\mathcal{F}_r, \mathcal{F}_i)\}, & \text{if } dist(\mathcal{F}_s, \mathcal{F}_i) > 0.5 \text{ and } dist(\mathcal{F}_r, \mathcal{F}_i) > 0.5 \\ 0.5, & \text{otherwise.} \end{cases} \tag{27}$$

Alternative definitions of $\Phi(c)$ are:

$$\Phi_1(\mathcal{F}) = \sum_{P_i, P_j \in \mathcal{F}} dist(P_i, P_j), \tag{28}$$

$$\Phi_2(\mathcal{F}) = \frac{\sum_{P_i, P_j \in \mathcal{F}} dist(P_i, P_j)}{\#\mathcal{F}(\#\mathcal{F} - 1)/2}, \tag{29}$$

$$\Phi_3(\mathcal{F}) = \sum_{P_i \in \mathcal{F}} dist(P_i, E), \tag{30}$$

where E is the Euclidean center of cluster $\mathcal{F}$.

The output of the heuristic algorithm is used as an initial point for the fixed-point (k-means) method that uses the BFY function. The definition of the functions (Φ and D) used in the heuristics might look as rather disconnected with the original problem. However, according our experience, in many global optimization problems very good results come from using initial approximations for local optimization procedures by means of criteria that do not seem connected with the original objective function. The reason is that, when one tries to avoid local minimizers, it is sensible to "forget" the true objective function for a while. Many "climbing" methods for global optimization are based on this principle.

5 Examples

In Table 1 the scores of 48 students in five different tests corresponding to the subject "Complements of Mathematics" at the University of Campinas during the first semester of 1997 are displayed. This is a small example. Large real-life problems can involve all the students that enter the university (about 2500) and as much as 11 tests.

We applied the algorithm to the classification of the students into four groups. The result was compared with the clustering obtained by the classical k-means and k-median algorithms. All possible combinations of D_i and Φ_j $(1 \leq i, j \leq 3)$ were tested in the heuristic to generate the initial clustering for the 3 algorithms. The initial solution which leads the algorithms to the best clustering were given in all cases by D_1 combined with Φ_2 for k-means and k-median algorithms and with Φ_1 for the algorithm using the BFY distance. The two classical algorithms obtained the same classification. In all cases, the local minimizers found starting from a point given by the different heuristics were better than local minimizers found in several experiments with trivial initial points. The minimization subproblems whose solutions are the centers of the clusters were solved using the spectral projected gradient method [6, 7]. Table 2 shows the obtained clusterings.

We mentioned in the introduction that the use of the BFY function is motivated by the specific need of classification in a situation where severe outliers are expected. Let us comment briefly here the way in which the characteristics of the function influenced the classification in the presented example. Notably, the students 19 and 22 did not appear in the first group of the BFY classification. The reason is that the score 10, obtained by these students in the fourth test was not considered meaningful by the clustering procedure. Clearly, the decision about the correctness of this classification depends of external considerations which, in this case, involve the specific environment under which the fourth test was applied. Further knowledge about the behavior of the involved students in other tests is also useful. In this case, one of the teachers of the course told us that, in fact, there are good reasons to disregard outstanding records in the mentioned test. Other discrepancies between the two classifications are due to the fact that BFY tends to put together points with the maximum number of similar coordinates. Again, the expert opinion in this specific case revealed that the taken decisions were quite adequate and more suitable for the specific purpose of the classification than the one obtained by the classical k-means method.

The fact that BFY tends to classify according to the "number of similarities" has important consequences when the purpose of the classification is to form several groups for different training programs.

6 Final remarks

The practical results obtained with the application of the approximate Boente-Fraiman-Yohai distance to the real problem considered in this paper in the fixed-point classification

Student	Scores				
P_1	7.80	3.50	8.50	9.00	8.90
P_2	6.90	2.00	7.00	7.00	5.00
P_3	9.40	6.50	7.00	9.80	9.20
P_4	0.00	0.00	0.00	0.00	0.00
P_5	1.90	0.50	6.50	1.00	2.00
P_6	9.70	4.00	8.50	7.00	7.80
P_7	2.50	0.00	4.00	6.30	0.00
P_8	7.20	1.00	6.00	5.30	6.40
P_9	6.50	8.50	6.00	6.50	7.70
P_{10}	6.50	0.00	7.00	5.30	1.20
P_{11}	9.70	6.00	9.00	9.50	8.50
P_{12}	5.00	3.00	7.50	8.30	5.70
P_{13}	3.70	0.00	1.50	3.00	0.00
P_{14}	3.50	0.50	6.50	3.80	2.20
P_{15}	3.90	0.00	4.50	2.30	0.80
P_{16}	4.50	3.00	4.00	1.50	3.30
P_{17}	5.40	0.00	1.50	4.30	0.00
P_{18}	0.00	0.00	0.00	0.00	0.00
P_{19}	7.40	1.50	7.50	10.00	6.90
P_{20}	3.30	1.50	6.50	7.20	3.20
P_{21}	5.90	4.50	8.00	6.80	8.30
P_{22}	7.60	5.50	7.50	10.00	6.90
P_{23}	8.50	6.00	6.50	10.00	7.30
P_{24}	6.70	7.00	8.50	7.50	7.70
P_{25}	3.70	3.50	6.00	7.00	6.80
P_{26}	1.00	0.00	0.50	0.50	0.00
P_{27}	6.70	1.00	8.00	6.00	4.90
P_{28}	3.90	0.00	4.00	2.80	0.00
P_{29}	7.00	1.50	5.50	2.00	4.40
P_{30}	7.50	0.00	2.00	4.00	0.00
P_{31}	4.00	3.00	6.50	5.30	7.70
P_{32}	8.00	2.00	7.50	9.80	8.20
P_{33}	4.20	4.50	5.00	7.50	3.50
P_{34}	0.00	0.00	0.50	0.00	0.00
P_{35}	3.70	0.50	6.50	6.30	1.50
P_{36}	4.20	0.00	0.00	1.50	0.00
P_{37}	1.90	0.00	4.50	0.50	0.00
P_{38}	6.90	1.50	5.50	5.00	3.40
P_{39}	0.00	0.00	2.50	1.50	0.00
P_{40}	4.70	0.50	5.50	5.80	5.70
P_{41}	7.20	3.50	7.00	8.30	7.50
P_{42}	0.00	0.00	0.00	0.00	0.00
P_{43}	5.70	0.50	6.00	5.50	5.00
P_{44}	5.20	0.50	1.50	2.30	2.90
P_{45}	5.90	2.00	8.00	6.50	7.50
P_{46}	1.10	0.50	6.00	1.80	1.80
P_{47}	7.70	0.50	7.70	6.30	6.70
P_{48}	10.00	7.00	8.50	10.00	9.80

Table 1: Scores of 48 students

k-means					
C_1	8.02	5.04	7.69	8.78	8.05
P_1	7.8	3.5	8.5	9.0	8.9
P_3	9.4	6.5	7.0	9.8	9.2
P_6	9.7	4.0	8.5	7.0	7.8
P_9	6.5	8.5	6.0	6.5	7.7
P_{11}	9.7	6.0	9.0	9.5	8.5
P_{19}	7.4	1.5	7.5	10.0	6.9
P_{21}	5.9	4.5	8.0	6.8	8.3
P_{22}	7.6	5.5	7.5	10.0	6.9
P_{23}	8.5	6.0	6.5	10.0	7.3
P_{24}	6.5	7.0	8.5	7.5	7.7
P_{32}	8.0	2.0	7.5	9.8	8.2
P_{41}	7.2	3.5	7.0	8.3	7.5
P_{48}	10.0	7.0	8.5	10.0	9.8
C_2	5.57	1.63	6.51	6.02	4.91
P_2	6.9	2.0	7.0	7.0	5.0
P_8	7.2	1.0	6.0	5.3	6.4
P_{10}	6.5	0.0	7.0	5.3	1.2
P_{12}	5.0	3.0	7.5	8.3	5.7
P_{20}	3.3	1.5	6.5	7.2	3.2
P_{25}	3.7	3.5	6.0	7.0	6.8
P_{27}	6.7	1.0	8.0	6.0	4.9
P_{29}	7.0	1.5	5.5	2.0	4.4
P_{31}	4.0	3.0	6.5	5.3	7.7
P_{33}	4.2	4.5	5.0	7.5	3.5
P_{35}	3.7	0.5	6.5	6.3	1.5
P_{38}	6.9	1.5	5.5	5.0	3.4
P_{40}	4.7	0.5	5.5	5.8	5.7
P_{43}	5.7	0.0	6.0	5.5	5.0
P_{45}	5.9	2.0	8.0	6.5	7.5
P_{47}	7.7	0.5	7.7	6.3	6.7
C_3	3.78	0.35	3.58	2.7	1.0
P_5	1.9	0.5	6.5	1.0	2.0
P_7	2.5	0.0	4.0	6.3	0.0
P_{13}	3.7	0.0	1.5	3.0	0.0
P_{14}	3.5	0.0	6.5	3.8	2.2
P_{15}	3.9	0.0	4.5	2.3	0.8
P_{16}	4.5	3.0	4.0	1.5	3.3
P_{17}	5.4	0.0	1.5	4.3	0.0
P_{28}	3.9	0.0	4.0	2.8	0.0
P_{30}	7.5	0.0	2.0	4.0	0.0
P_{36}	4.2	0.0	0.0	1.5	0.0
P_{37}	1.9	0.0	4.5	0.5	0.0
P_{44}	5.2	0.5	1.5	2.3	2.9
P_{46}	1.1	0.5	6.0	1.8	1.8
C_4	0.17	0.0	0.58	0.33	0.0
P_4	0.0	0.0	0.0	0.0	0.0
P_{18}	0.0	0.0	0.0	0.0	0.0
P_{26}	1.0	0.0	0.5	0.5	0.0
P_{34}	0.0	0.0	0.5	0.0	0.0
P_{39}	0.0	0.0	2.5	1.5	0.0
P_{42}	0.0	0.0	0.0	0.0	0.0

BFY					
C_1	9.70	6.72	8.62	9.80	7.79
P_3	9.4	6.5	7.0	9.8	9.2
P_6	9.7	4.0	8.5	7.0	7.8
P_{11}	9.7	6.0	9.0	9.5	8.5
P_{23}	8.5	6.0	6.5	10.0	7.3
P_{24}	6.5	7.0	8.5	7.5	7.7
P_{32}	8.0	2.0	7.5	9.8	8.2
P_{48}	10.0	7.0	8.5	10.0	9.8
C_2	7.34	3.35	7.65	6.58	7.15
P_1	7.8	3.5	8.5	9.0	8.9
P_2	6.9	2.0	7.0	7.0	5.0
P_9	6.5	8.5	6.0	6.5	7.7
P_{12}	5.0	3.0	7.5	8.3	5.7
P_{19}	7.4	1.5	7.5	10.0	6.9
P_{21}	5.9	4.5	8.0	6.8	8.3
P_{22}	7.6	5.5	7.5	10.0	6.9
P_{25}	3.7	3.5	6.0	7.0	6.8
P_{27}	6.7	1.0	8.0	6.0	4.9
P_{29}	7.0	1.5	5.5	2.0	4.4
P_{41}	7.2	3.5	7.0	8.3	7.5
P_{45}	5.9	2.0	8.0	6.5	7.5
P_{47}	7.7	0.5	7.7	6.3	6.7
C_3	3.86	0.30	6.30	5.39	3.20
P_5	1.9	0.5	6.5	1.0	2.0
P_8	7.2	1.0	6.0	5.3	6.4
P_{10}	6.5	0.0	7.0	5.3	1.2
P_{14}	3.5	0.0	6.5	3.8	2.2
P_{15}	3.9	0.0	4.5	2.3	0.8
P_{16}	4.5	3.0	4.0	1.5	3.3
P_{20}	3.3	1.5	6.5	7.2	3.2
P_{31}	4.0	3.0	6.5	5.3	7.7
P_{33}	4.2	4.5	5.0	7.5	3.5
P_{35}	3.7	0.5	6.5	6.3	1.5
P_{38}	6.9	1.5	5.5	5.0	3.4
P_{40}	4.7	0.5	5.5	5.8	5.7
P_{43}	5.7	0.0	6.0	5.5	5.0
P_{44}	5.2	0.5	1.5	2.3	2.9
P_{46}	1.1	0.5	6.0	1.8	1.8
C_4	0.0	0.0	0.0	0.05	0.0
P_4	0.0	0.0	0.0	0.0	0.0
P_7	2.5	0.0	4.0	6.3	0.0
P_{13}	3.7	0.0	1.5	3.0	0.0
P_{17}	5.4	0.0	1.5	4.3	0.0
P_{18}	0.0	0.0	0.0	0.0	0.0
P_{26}	1.0	0.0	0.5	0.5	0.0
P_{28}	3.9	0.0	4.0	2.8	0.0
P_{30}	7.5	0.0	2.0	4.0	0.0
P_{34}	0.0	0.0	0.5	0.0	0.0
P_{36}	4.2	0.0	0.0	1.5	0.0
P_{37}	1.9	0.0	4.5	0.5	0.0
P_{39}	0.0	0.0	2.5	1.5	0.0
P_{42}	0.0	0.0	0.0	0.0	0.0

Table 2: Four groups clustering.

framework were quite satisfactory. In this research we did not consider the problem of determining the optimum number of clusters. See [13, 14]. In fact, this was not relevant in our case, in which the number of clusters is predetermined, but can be important in other applications.

The analysis of classification results is, many times, much easier under a graphical computer environment. We are planning to adapt our codes to this type of environment, along the lines of [33]. Another promising line of research is to investigate the robust capabilities of neural network classifiers [10].

Finally, we would like to mention that the problem presented in this paper provides a new example of the application of a novel optimization technique, based on the spectral gradient philosophy [24, 25, 19, 6]. This family of algorithms has proved to be very efficient in several practical nontrivial situations ([3, 4, 5]) where more sophisticated methods failed. In the case study presented in this paper, we feel that the nonmonotonic strategy used by spectral methods has been useful to avoid some local nonglobal minimizers.

Acknowledgements

We are indebted to Mario A. Gneri, for having introduced the BFY function to us and for useful comments on this paper. We are also indebted to Nataša Krejić and to two anonymous referees whose comments helped us a lot for improving the paper.

References

[1] J. C. Bezdek, J. Keller, R. Krisnapuram, M. Pal (editors), *Fuzzy models and algorithms for pattern recognition and image processing*, Kluwer Academic Publishers, 1999.

[2] J. C. Bezdek, S. K. Pal (editors), *Fuzzy models for pattern recognition: methods that search for structures in data*, IEEE Press, Piscataway, NJ, 1991.

[3] E.G. Birgin, R. Biloti, M. Tygel, L.T. Santos, Restricted optimization: a clue to a fast and accurate implementation of the Common Reflection Surface stack method, Journal of Applied Geophysics 42 (1999) 143–155.

[4] E. G. Birgin, I. Chambouleyron, J.M. Martínez, Estimation of the optical constants and the thickness of thin films using unconstrained optimization, Journal of Computational Physics 151 (1999) 862–880.

[5] E. G. Birgin, Y.G. Evtushenko, Automatic differentiation and spectral projected gradient methods for optimal control problems, Optimization Methods and Software 10 (1998) 125–146.

[6] E.G. Birgin, J.M. Martínez, M. Raydan, Nonmonotone spectral projected gradient methods on convex sets, SIAM Journal on Optimization 10 (2000) 1196–1211.

[7] E. G. Birgin, J. M. Martínez, M. Raydan, SPG: Software for convex-constrained optimization, to appear in ACM Transactions on Mathematical Software (2001).

[8] L. Bobrowski, C. Bezdek, c-Means clustering with the ℓ_1 and ℓ_∞ norm, IEEE Transactions on Systems, Management and Cybernetics 21 (1991) 545–554.

[9] G. Boente, R. Fraiman, V.J. Yohai, Qualitative robustness for stochastic processes, Annals of Statistics 15 (1987) 1293–1312.

[10] W. Chang, H.S. Soliman, Image coding by a neural net classification process, Applied artificial intelligence 11 (1997) 33–57.

[11] R.N. Davé, R. Krishnapuran, Robust clustering methods: a unified view, IEEE Transactions on Fuzzy Systems 5 (1997) 270–293.

[12] J.E. Dennis Jr., R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall, Englewood Cliffs, 1983.

[13] H. Frigui, R. Krishnapuran, A robust algorithm for automatic extraction of an unknown number of clusters from noisy data, Pattern Recognition Letters 12 (1996) 1223–1232.

[14] H. Frigui, R. Krishnapuran, A robust competitive clustering algorithm with applications in computer vision, IEEE Transactions on Pattern Analysis 21 (1999) 450–465.

[15] K. E. Gowda, E. Diday, Symbolic clustering using a new similarity measure, IEEE Transactions on Systems, Management and Cybernetics 22 (1992) 368–378.

[16] T. Gupta, H. Seifoddini, Production data based similarity coefficient in the design of a cellular manufacturing, a comparative study, International Journal of Production Research 28 (1990) 1247–1269.

[17] G. Harhalakis, R. Nagis, J. M. Proth, An Efficient Heuristic in Manufacturing Cell Formation for Group Technology Applications, International Journal of Production Research 28 (1990) 185–198.

[18] M. Kendall, Multivariate Analysis, Charles Griffin and Company, London, 1975.

[19] F. Luengo, M. Raydan, W. Glunt, T.L. Hayden, Preconditioned spectral gradient method, Technical Report R.T. 96-08, Computer Science Department, Universidad Central de Venezuela, Venezuela, 1996. Submitted with corrections to SIAM Journal on Scientific Computing.

[20] J. McAuley, Machine grouping for efficient production, The Production Engineer 52 (1972) 53–57.

[21] J.M. Ortega, W.G. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[22] G. Qiu, M. R. Varley, T.J. Terrell, Improved clustering using deterministic annealing with a gradient descent technique, Pattern Recognition Letters 15 (1994) 607–610.

[23] M. Randerberg, Cluster Analysis for Applications, Academic Press, New York, 1973.

[24] M. Raydan, On the Barzilai and Borwein choice of steplength for the gradient method, IMA Journal of Numerical Analysis 13 (1993) 321–326.

[25] M. Raydan, The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem, SIAM Journal on Optimization 7 (1997) 26–33.

[26] H. Seifoddini, Single linkage versus average linkage clustering in machine cells formation applications, Computers and Industrial Engineering 16 (1989) 419–426.

[27] H. Seifoddini, B. Tjahjana, Part-family formation for cellular manufacturing: a case study at Harnischfeger, International Journal of Production Research 37 (1999) 3263–3273.

[28] H. Seifoddini, P.M. Wolfe, Application of similarity coefficient method in group technology cells, International Journal of Production Research 28 (1986) 293–300.

[29] H. Späth, L_1 Cluster-Analysis, Computing 16 (1976) 379–387.

[30] H. Späth, Computational experiences with the exchange method, European Journal of Operational Research 1 (1977) 23–32.

[31] H. Späth, Numerical experiences with heuristic solution methods for minimum variance criterion within cluster-analysis, Angew. Inform. 2 (1977) 67–72.

[32] H. Späth, Clusterwise linear-regression, Computing 22 (1979) 367–373.

[33] A. Struyf, M. Hubert, P.J. Rousseeuw, Integrating robust clustering techniques in S-PLUS, Computer Statistics and Data Analysis 26 (1997) 17–37.