ELSEVIER Journal of Computational and Applied Mathematics 64 (1995) 81–90

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Fuzzy clustering with high contrast

P.J. Rousseeuw[a,*], E. Trauwaert[b], L. Kaufman[c]

[a] *Department of Mathematics and Computing, University of Antwerp (UIA), Universiteitsplein 1, B-2610 Antwerp, Belgium*
[b] *Belgoprocess, Gravenstraat 73, B-2480 Dessel, Belgium*
[c] *VUB, Pleinlaan 2, B-1050 Brussels, Belgium*

Received 9 November 1993; revised 25 November 1994

Abstract

In a fuzzy clustering an object typically receives strictly positive memberships to all clusters, even when the object clearly belongs to one particular cluster. Consequently, each cluster's estimated center and scatter matrix are influenced by many objects that have small positive memberships to it. This effect may keep the fuzzy method from finding the true clusters. We analyze the cause and propose a remedy, which is a modification of the objective function and the corresponding algorithm. The resulting clustering has a high contrast in the sense that outlying and bridging objects remain fuzzy, whereas the other objects become crisp. The enhanced version of fuzzy k-means is illustrated with an example, as well as the enhanced version of the fuzzy minimum volume method.

Keywords: Algorithm; Classification; Cluster analysis; k-means

1. Introduction

For partitioning a data set into groups of similar objects it has been argued by many authors [2, 3, 6, 9, 10] that fuzzy approaches often work better than crisp ones. This is the case for many iterative algorithms which converge to a local minimum of the objective function, without any assurance of its proximity to the global minimum. In this situation a fuzzy method evolves more smoothly to the global minimum whereas a crisp method bears more risk to get stuck in a local minimum. Moreover, fuzzy methods are better able to cope with marginal objects such as outliers and bridges [5].

However, fuzzy clustering methods also have their drawbacks. One of the shortcomings is that nearly all objects receive positive membership values to all clusters. Therefore the cluster centers, which are weighted means of all objects, are influenced by objects that clearly do not belong to these clusters. This effect can be observed when a small cluster with few objects is close to a large

* Corresponding author.

SSDI 0377-0427(95)00008-9

cluster with many objects. The center of the small cluster will then be biased towards the large cluster, because the many objects of the large cluster all have a nonzero membership to the small cluster and hence contribute to its center. Since these centers (and often scatter matrices as well) need to be computed *inside* the clustering algorithm, this may seriously affect the clustering itself.

These effects would be avoided or at least largely reduced if only the outlying objects and the bridges remain fuzzy, whereas the other objects become crisp. In this way the memberships would no longer all be grey: quite a number of them would be white ($u_{it} = 1$) whereas most others would be black ($u_{it} = 0$), and only a few would remain grey ($0 < u_{it} < 1$). The overall picture would therefore have a much higher contrast.

2. General formulation of nonhierarchial clustering methods

Notwithstanding their different objectives, most nonhierarchial clustering methods can be described by the same generic elements as follows.

Suppose we have multivariate objects x_i for $i = 1, \ldots, n$, each described by p attributes, such that x_i is the column vector

$$x_i = (x_{i1} \ldots x_{ij} \ldots x_{ip})' \quad \text{for all } i = 1, \ldots, n. \tag{1}$$

The general purpose is to group the objects into k clusters, each of which is characterized by a center μ_t (with $t = 1, \ldots, k$) and possibly a scatter matrix. The unknowns of the problem are the membership functions u_{it}, together with the centers μ_t and the scatter matrices. A membership u_{it} with value 1 indicates that object i belongs completely to cluster t; the value 0 means that it does not belong to this cluster. Two very different approaches are possible. If u_{it} is defined as a Boolean variable, with the sole values 0 or 1, the clustering is said to be *crisp*. If however u_{it} can take on all values between 0 and 1, we have a *fuzzy* clustering. In both cases the partition constraint must be verified:

$$\sum_t u_{it} = 1 \quad \text{for all } i = 1, \ldots, n, \tag{2}$$

where

$$u_{it} \geqslant 0 \quad \text{for all } i \text{ and } t. \tag{3}$$

All these elements are used to compute the objective function, which is of the type:

$$F = F(x_i, u_{it}). \tag{4}$$

It always involves the n objects x_i and the $n \times k$ membership functions u_{it}. The actual form taken by this objective function varies according to the type of method and the purpose of the clustering, as will be seen below.

A well-known clustering method is *fuzzy k-means* [3] where (4) becomes

$$F = \sum_t \sum_i u_{it}^{\alpha} (x_i - \mu_t)'(x_i - \mu_t) \tag{5}$$

for arbitrary vectors μ_t and memberships u_{it} (subject to (2) and (3)). The general formulation given above also covers several other fuzzy clustering methods which will be described in Section 5.

3. The role of the exponent α

We will look more closely at the effect of the exponent α on the optimal solution. Let us first consider the case of the fuzzy k-means method (5) in a situation with two clusters.

A solution of this optimization problem is a collection of memberships $\hat{u}_{it}$ (for $1 \leqslant i \leqslant n$ and $1 \leqslant t \leqslant k$) and cluster centers $\hat{\mu}_1, \ldots, \hat{\mu}_k$ which together minimize (5). By fixing everything at its optimal value except for $\hat{u}_{i1}$ and $\hat{u}_{i2}$ we see that this implies that $\hat{u}_{i1}$ and $\hat{u}_{i2}$ minimize the expression

$$F_i = u_{i1}^{\alpha}(x_i - \hat{\mu}_1)'(x_i - \hat{\mu}_1) + u_{i2}^{\alpha}(x_i - \hat{\mu}_2)'(x_i - \hat{\mu}_2)$$
$$= F_{i1} + F_{i2} \tag{6}$$

subject to the constraint

$$u_{i1} + u_{i2} = 1. \tag{7}$$

Here, $F_{it} := u_{it}^{\alpha}(x_i - \hat{\mu}_t)'(x_i - \hat{\mu}_t)$. Fig. 1 shows how the first term of (6) depends on the membership u_{i1} for different values of α.

Although the relation between u_{i1} and F_{i1} depends on α, we note that F_{i1} is always zero for $u_{i1} = 0$, and that for $u_{i1} = 1$ it becomes

$$\delta_{i1} := (x_i - \hat{\mu}_1)'(x_i - \hat{\mu}_1). \tag{8}$$

Therefore, the minimum of F_{i1} is always obtained for $u_{i1} = 0$. Furthermore, for any $\alpha > 1$ it has a zero derivative at $u_{i1} = 0$, whereas at $u_{i1} = 1$ the derivative equals

$$\alpha(x_i - \hat{\mu}_1)'(x_i - \hat{\mu}_1) = \alpha\delta_{i1}. \tag{9}$$

For $\alpha = 2$ the function F_{i1} is a parabola. For $\alpha = 1$, the curve degenerates to a straight line.

As the second term of (6) has the same form as the first, the same conclusions hold for F_{i2} except that δ_{i2} may differ from δ_{i1}. We will assume that $\delta_{i1} \neq 0$ and $\delta_{i2} \neq 0$.

Plotting the whole function F_i subject to (7) amounts to superimposing Fig. 1 and a reversed graph for the second term, as in Fig. 2. The function F_i takes on the value δ_{i2} at $u_{i1} = 0$ and δ_{i1} at $u_{i1} = 1$.

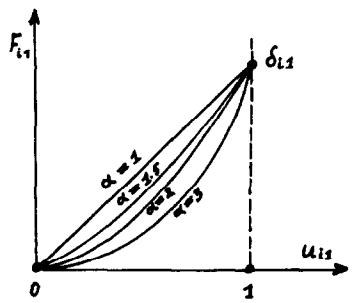

Fig. 1. Plot of F_{i1} as a function of u_{i1} for various exponents α.

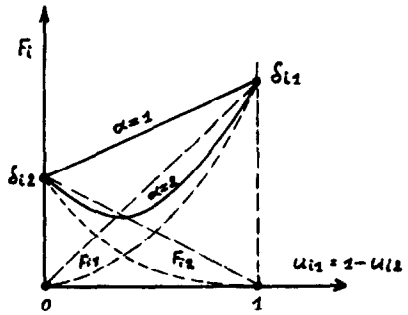

Fig. 2. Plot of F_i as a function of u_{i1} for $\alpha = 1$ and $\alpha = 2$.

For $\alpha > 1$, the derivative of F_i at $u_{i1} = 0$ is strictly negative, and equal to the derivative of F_{i2} there. At $u_{i1} = 1$ the derivative has the same positive value (9) as before. Hence, the minimum of F_i must be attained for memberships u_{it} that lie inside $]0, 1[$. By differentiation we find that the minimum is attained for

$$u_{i1} = \frac{\delta_{i2}^{\beta}}{\delta_{i1}^{\beta} + \delta_{i2}^{\beta}} \quad \text{with } \beta = 1/(\alpha - 1). \tag{10}$$

Hence, if $\delta_{i1} = \delta_{i2}$ we have $u_{i1} = u_{i2} = \frac{1}{2}$. Moreover, for $\delta_{i1} > \delta_{i2}$ we have $u_{i1} < \frac{1}{2}$, and vice versa.

For $\alpha = 1$, the picture is quite different. The derivative of F_i then equals $\delta_{i1} - \delta_{i2}$ for all $0 \leqslant u_{i1} \leqslant 1$. Therefore the minimum of F_i is attained for $u_{i1} = 0$ if $\delta_{i2} < \delta_{i1}$, and for $u_{i1} = 1$ if $\delta_{i2} > \delta_{i1}$.

The above reasoning explains why choosing $\alpha > 1$ yields a *strictly fuzzy* result (that is, always $0 < u_{it} < 1$), whereas $\alpha = 1$ yields a *crisp* result (i.e., $u_{it} = 0$ or $u_{it} = 1$) even though the u_{it} are in principle allowed to take any value in the continuous range $[0, 1]$.

Note that the above reasoning can be generalized to situations with more than 2 clusters. Moreover, we will see in Section 5 that the same effect also exists when implementing other fuzzy clustering methods, for instance based on minimizing the total volume of tolerance ellipsoids.

The challenge is now to find a way to improve the objective function such that "clear-cut" objects will be classified in a crisp manner, whereas "doubtful" objects are still classified in a fuzzy manner. The next section indicates how this can be done.

4. Fuzzy clustering with high contrast

We have seen in the previous section why the objective function (6) yields strictly fuzzy memberships for nearly all objects. This results directly from the fact that for $\alpha > 1$, each F_{it} has a zero derivative at $u_{it} = 0$. This would no longer be the case if this derivative could be made positive. This can be done by replacing u_{it}^{α} by another function $f(u_{it})$.

Let us take a parabola which passes through the points (0, 0) and (1, 1) but which has a positive derivative at any $u_{it} \geqslant 0$. Since $f(0) = 0$ and $f(1) = 1$ we obtain

$$f(u_{it}) = cu_{it} + (1 - c)u_{it}^2 \tag{11}$$

where $c < 1$ because the coefficient of u_{it}^2 must be positive. Moreover, the derivative of f at $u_{it} = 0$ equals c, hence we have to take $c > 0$. The constant $0 < c < 1$ will be called the *contrast factor*. When $c \to 0$ we see that (11) reduces to u_{it}^2 which yields a strictly fuzzy clustering (no contrast). For $c \to 1$ formula (11) becomes u_{it} which yields a crisp clustering (maximal contrast). For $0 < c < 1$ the function f lies between these extremes, and yields a fuzzy clustering with enhanced contrast.

Remark. A more general parabola-like function can be obtained as follows. Fix $0 \leqslant \varepsilon \leqslant 1$, and put

$$f(u_{it}) = cu_{it} + (1 - c)/[\varepsilon(2 - \varepsilon)]u_{it}^2 \quad \text{if } u_{it} \leqslant \varepsilon, \tag{12}$$

$$= [(2 - c\varepsilon)u_{it} - (1 - c)\varepsilon]/(2 - \varepsilon) \quad \text{if } u_{it} \geqslant \varepsilon. \tag{13}$$

For $\varepsilon = 1$, (12) reduces to (11); for $\varepsilon = 0$, (12) is inoperational and (13) reduces to

$$f(u_{it}) = u_{it} \tag{14}$$

which is the equation of a straight line. For $0 < \varepsilon < 1$, Eqs. (12) and (13) represent a curve extended by a straight line, which is a generalization of (11). Moreover, at $u_{it} = \varepsilon$ we have

$$f(\varepsilon -) = f(\varepsilon +) \quad \text{and} \quad f'(\varepsilon -) = f'(\varepsilon +), \tag{15}$$

which ensures the continuity of the function f and its first derivative.

5. General algorithm

Apart from the fuzzy k-means method we will list several other fuzzy clustering techniques, and then present an algorithm by which their contrast can be improved as well.

1. The *adaptive distances* method [4] has

$$F = \sum_t \sum_i u_{it}^{\alpha}(x_i - \mu_t)' G_t (x_i - \mu_t). \tag{16}$$

Here G_t is an unknown positive-definite matrix, which is estimated by the optimization of the objective function (16). However, in order to prevent any G_t from becoming (nearly) singular, this matrix must somehow be constrained. Gustafson and Kessel proposed the set of constraints

$$|G_t| = \theta_t \quad \text{for all } t = 1, \ldots, k, \tag{17}$$

with θ_t having a fixed value for each cluster. For this technique to produce the natural clusters, their relative volumes must be known in advance.

2. The *minimum determinant* method [9] has

$$F = |S| \tag{18}$$

in which S is defined for all clusters simultaneously by

$$S = \sum_t \sum_i u_{it}^{\alpha}(x_i - \mu_t)(x_i - \mu_t)'/n. \tag{19}$$

This method is based on the assumption that all clusters have similar shapes.

3. The *product of determinants* method [9] is based on the maximum likelihood criterion. Here

$$F = \prod_t |S_t|^{n_t} \tag{20}$$

in which S_t and n_t are defined as

$$S_t = \sum_i u_{it}^{\alpha}(x_i - \mu_t)(x_i - \mu_t)'/n_t$$
$$n_t = \sum_i u_{it}. \tag{21}$$

This method allows for clusters of different shapes, but tries to obtain clusters with similar volumes. This restriction is largely avoided in the following two methods.

4. The *minimum total volume* method of fuzzy clustering [7] proceeds by minimizing

$$F = \sum_t |S_t|^{1/2} \tag{22}$$

where S_t again depends on the memberships u_{it} through (21).

5. The *sum of all normalized determinants* (SAND) method is defined by minimizing the objective function

$$F = \sum_t |S_t|^{1/p} \tag{23}$$

(see [7]) with p the dimension.

All these objective functions depend only on the memberships u_{it}. They can all be minimized by the same algorithm, described in [7], which puts the derivatives with respect to the u_{it} equal to zero and uses the Lagrange method of constrained optimization.

In case the memberships in the objective function occur in the form u_{it}^2, the results are

$$\hat{\mu}_t = \frac{\sum_i u_{it}^2 x_i}{\sum_i u_{it}^2} \tag{24}$$

and

$$\hat{u}_{it} = \frac{1/B_{it}}{\sum_{r \notin T_i} 1/B_{ir}} - \frac{1}{B_{it}}\left[\frac{\sum_{r \notin T_i} A_r/B_{ir}}{\sum_{r \notin T_i} 1/B_{ir}} - A_t\right] \geqslant 0 \tag{25}$$

for $t \notin T_i$, and

$$\hat{u}_{it} = 0 \text{ for } t \in T_i.$$

In this solution, T_i (which depends on i) represents the set of indices t for which (25) would become strictly negative. The values of A_t and B_{it} are typical for each method and can be found in the corresponding references.

Replacing u_{it}^2 in the objective function by (11) yields the same results but with

$$\hat{\mu}_t = \frac{\sum_i f(u_{it}) x_i}{\sum_i f(u_{it})}, \tag{26}$$

$$S = \sum_t \sum_i f(u_{it})(x_i - \mu_t)(x_i - \mu_t)'/n, \tag{27}$$

$$S_t = \sum_i f(u_{it})(x_i - \mu_t)(x_i - \mu_t)'/n_t \tag{28}$$

and where A_t and B_{it} are replaced by A'_{it} and B'_{it} defined as follows:

$$A'_{it} = A_t - cB_{it}/\alpha,$$

$$B'_{it} = (1 - c)B_{it}.$$

Based on these relations it is possible to develop a unified algorithm, generalizing the algorithm proposed in [1] for the fuzzy k-means method. The main steps are as follows:

1. Initialize membership values of all objects with respect to each cluster, in accordance with the conditions (2) and (3);
2. Calculate the center of each cluster according to (26) and – if relevant – its scatter matrix according to (27) or (28);
3. For each object i, initalize the set $T_i = \emptyset$ and evaluate the membership functions according to (25). If some of the memberships are negative, put them equal to zero, put their index in the set T_i and recalculate the other memberships according to (25). Iterate as long as some memberships are strictly negative. Repeat this for all objects i;
4. Compare the memberships with those of the previous passage. If no values differ by more than a quantity ε, stop; otherwise go back to step 2.

By choosing adequate values for the contrast factor c, it is now possible to choose to which extent the memberships will be crisp or fuzzy, as we will see in the following examples.

6. Examples

In a first example it will be shown that the proposed approach effectively enhances the contrast between the elements that clearly belong to a cluster and those that are doubtful.

The data, from Kaufman and Rousseeuw [5], consist of three well-separated clusters with two outliers or bridges, one of which is almost at the same distance from all three clusters, whereas the other is equally distant from two clusters but further away from the third cluster (Fig. 3). When applying the fuzzy k-means clustering method (with zero contrast factor) to these data, the three clusters are easily found, with memberships at least 0.95 for all objects but the two outliers. We see that the membership to a cluster decreases with the distance to its center (Table 1). When using fuzzy k-means with a contrast factor of (say) 0.3 the result is even sharper, as only the two outliers are left with fuzzy memberships, all other objects being crisp. Changing the contrast factor further does not influence the results much, apart from slightly varying the memberships of the outliers.

The second example shows that enhancing the contrast may be necessary to find the natural clusters.

The example uses a data set of 64 four-dimensional objects, obtained by generating 8 objects from one population and 56 from another [8]. Each analysis is repeated for 10 different starting

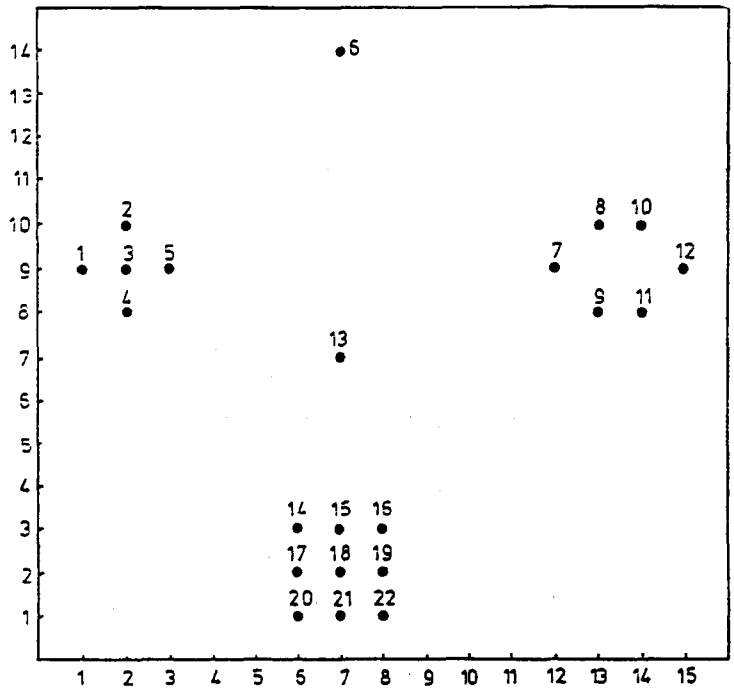

Fig. 3. Plot of 22 objects.

partitions, which are needed by the fuzzy algorithm of Section 5. Applying the uncontrasted ($c = 0$) minimum volume method to this data set yields a reasonable result, but it does not provide the expected natural clusters. This is because the typical fuzziness bias occurs: due to the positive memberships of the objects of the large cluster to the small cluster, the center of the small cluster is shifted towards the large cluster.

To avoid this fuzziness bias we can use a positive contrast factor. Table 2 shows, for different values of the contrast factor c, the frequency of obtaining the natural and alternative clusterings. For $c = 0$ we see the fuzziness bias. For values of c between 0.33 and 0.50 we obtain the correct partition in a stable way (at least 9 times out of 10). Increasing c to 0.80 and 0.95 only seldom yields the partition corresponding to the global optimum, because then the method starts to behave as a crisp algorithm that converges to a local minimum. It appears that an intermediate value of c (such as $c = \frac{1}{3}$ in this example) works best.

7. Conclusions

The proposed method of enhanced contrast combines the best of both worlds by using fuzziness features for performing the calculations and for characterizing truly fuzzy objects, while leaving the others crisp. The contrasted minimum volume method becomes capable of finding clusters of

Table 1
Comparison between uncontrasted and contrasted clusterings

	Data		Fuzzy k-means					
			$c = 0$			$c = 0.3$		
i	x	y	u_{i1}	u_{i2}	u_{i3}	u_{i1}	u_{i2}	u_{i3}
1	1	9	0.97	0.01	0.02	1.00	0.00	0.00
2	2	10	0.98	0.01	0.01	1.00	0.00	0.00
3	2	9	1.00	0.00	0.00	1.00	0.00	0.00
4	2	8	0.96	0.01	0.03	1.00	0.00	0.00
5	3	9	0.99	0.00	0.01	1.00	0.00	0.00
6	7	14	0.50	0.34	0.16	0.69	0.31	0.00
7	12	9	0.02	0.96	0.02	0.00	1.00	0.00
8	13	10	0.01	0.98	0.01	0.00	1.00	0.00
9	13	8	0.01	0.97	0.02	0.00	1.00	0.00
10	14	10	0.01	0.98	0.01	0.00	1.00	0.00
11	14	8	0.01	0.97	0.02	0.00	1.00	0.00
12	15	9	0.02	0.96	0.02	0.00	1.00	0.00
13	7	7	0.37	0.22	0.41	0.45	0.09	0.46
14	6	3	0.03	0.02	0.95	0.00	0.00	1.00
15	7	3	0.01	0.01	0.98	0.00	0.00	1.00
16	8	3	0.02	0.03	0.95	0.00	0.00	1.00
17	6	2	0.02	0.01	0.97	0.00	0.00	1.00
18	7	2	0.00	0.00	1.00	0.00	0.00	1.00
19	8	2	0.01	0.01	0.98	0.00	0.00	1.00
20	6	1	0.03	0.02	0.95	0.00	0.00	1.00
21	7	1	0.02	0.01	0.97	0.00	0.00	1.00
22	8	1	0.02	0.02	0.96	0.00	0.00	1.00

Table 2
Results of the contrasted minimum volume method applied to data with two natural clusters, with 8 and 56 objects

Outcomes in 10 trials			Contrast factor c					
			0.00	0.33	0.50	0.67	0.80	0.95
Typical	clusters	8/56	—	10	9	7	4	—
		13/51	10	—	—	1	—	—
		30/34	—	—	—	2	2	—
		others	—	—	1	—	4	10
		Total	10	10	10	10	10	10

unequal shapes (provided they are somewhat ellipsoidal) and unequal numbers of objects. Provided its contrast factor is chosen adequately, the corresponding algorithm typically finds the global minimum.

References

[1] J.C. Bezdek, Numerical taxonomy with fuzzy sets, *J. Math. Biology* **1** (1974) 57–71.
[2] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum Press, New York, 1981).
[3] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. Cybernet.* **3** (1974) 32–57.
[4] D.E. Gustafson and W. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: K.S. Fu, Ed., *Proc. IEEE-CDC, Vol 2* (IEEE Press, Piscataway, 1979) 761–766.
[5] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York, 1990).
[6] M. Roubens, Pattern classification problems and fuzzy sets, *Fuzzy Sets and Systems* **1** (1978) 239–253.
[7] P.J. Rousseeuw, L. Kaufman and E. Trauwaert, Fuzzy clustering using scatter matrices, *Comput. Statist. Data Anal.* (1995) to appear.
[8] E. Trauwaert, Grouping of objects using objective functions with applications to the monitoring and control of industrial production processes, Ph.D. Thesis, Faculty of Applied Sciences, Vrije Universiteit Brussel, 1991.
[9] E. Trauwaert, L. Kaufman and P. Rousseeuw, Fuzzy clustering algorithms based on the maximum likelihood principle, *Fuzzy Sets and Systems* **42** (1991) 213–227.
[10] L.A. Zadeh, Fuzzy sets and their application to pattern classification and cluster analysis, in: J. Van Ryzin, Ed., *Classification and Clustering* (Academic Press, New York, 1977) 251–299.