Effects of Irrelevant Attributes in Fuzzy Clustering

Christian Döring, Christian Borgelt, and Rudolf Kruse
Dept. of Knowledge Processing and Language Engineering
Otto-von-Guericke-University of Magdeburg
Universitätsplatz 2, D-39106 Magdeburg, Germany,
{doering,borgelt,kruse}@iws.cs.uni-magdeburg.de

***Abstract*— In fuzzy clustering soft cluster partitions are formed based on the similarity of data points to the respective cluster prototypes. Similarity is defined in terms of simultaneous closeness regarding all attributes. In some applications the values of many attributes have been measured, but a natural clustering, if it exists, occurs within a (small) subset of attributes. The remaining dimensions can be considered irrelevant. They can obscure an existing grouping and make it harder to discover the cluster structure. In probabilistic fuzzy clustering irrelevant attributes can lead to coincidental cluster centers in the worst case. We study this effect in detail as well as the robustness of different similarity functions and their possible parameterizations against irrelevant input dimensions. Empirical evidence is given for the different properties of the membership functions.**

I. Fuzzy Clustering

Most fuzzy clustering algorithms are objective function based: they determine an optimal (fuzzy) partition of a given data set $\mathbf{X} = \{\vec{x}_j \mid j = 1, \ldots, n\}$ into clusters by minimizing an objective function

$$J(\mathbf{X}, \mathbf{U}, \mathbf{C}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \qquad (1)$$

subject to the constraints

$$\sum_{j=1}^{n} u_{ij} > 0, \qquad \text{for all } i \in \{1, \ldots, c\}, \qquad \text{and} \qquad (2)$$

$$\sum_{i=1}^{c} u_{ij} = 1, \qquad \text{for all } j \in \{1, \ldots, n\}. \qquad (3)$$

Here $u_{ij} \in [0,1]$ is the membership degree of datum $\vec{x}_j$ to cluster i and d_{ij} is the distance between datum $\vec{x}_j$ and cluster i. The $c \times n$ matrix $\mathbf{U} = (u_{ij})$ is called the fuzzy partition matrix and $\mathbf{C}$ describes the set of clusters by stating location parameters (i.e. the cluster center) and maybe size and shape parameters for each cluster. The parameter m, $m > 1$, is called the *fuzzifier* or *weighting exponent*. It determines the "fuzziness" of the classification: with higher values for m the boundaries between the clusters become softer, with lower values they get harder. Usually $m = 2$ is chosen.

Constraint (2) guarantees that no cluster is empty. Constraint (3) ensures that the membership degrees of a datum to the clusters sum up to 1 and thus that each datum has the same total influence. Because of the second constraint this approach is usually called *probabilistic fuzzy clustering*, since with it the membership degrees for a datum formally resemble the probabilities of its being a member of the corresponding clusters. The partitioning property of a probabilistic clustering algorithm, which "distributes" the weight of a datum to the different clusters, is due to this constraint.

Unfortunately, the objective function J cannot be minimized directly. Therefore an iterative algorithm is used, which alternately optimizes the membership degrees and the cluster parameters. That is, first the membership degrees are optimized for fixed cluster parameters, then the cluster parameters are optimized for fixed membership degrees. The main advantage of this scheme is that in each of the two steps the optimum can be computed directly. By iterating the two steps the joint optimum is approached (although it cannot be guaranteed that the global optimum will be reached—the algorithm may get stuck in a local minimum of the objective function J).

The update formulae are derived by simply setting the derivative of the objective function J w.r.t. the parameters to optimize equal to zero (necessary condition for a minimum). Independent of the chosen distance measure we thus obtain the following update formula for the membership degrees [1]:

$$u_{ij} = \frac{d_{ij}^{-\frac{2}{m-1}}}{\sum_{t=1}^{c} d_{tj}^{-\frac{2}{m-1}}}. \qquad (4)$$

The update formulae for the cluster parameters depend, of course, on what parameters are used to describe a cluster (location, shape, size) and on the chosen distance measure. In this paper we take the fuzzy c-means (FCM) algorithm [2] as an example. In the FCM the Euclidean distance measures the dissimilarity of data points to the clusters, which are described by their centers $\vec{c}_i$ only. Therefore the update formula for the clusters in the alternating optimization scheme is given by

$$\vec{c}_i = \frac{\sum_{j=1}^{n} u_{ij}^m \vec{x}_j}{\sum_{j=1}^{n} u_{ij}^m}. \qquad (5)$$

This paper is organized as follows. In the next section we give an illustrative example of the effects that occur when datasets with irrelevant attributes are clustered. In Section III we analyze the observed effects and show which properties of the similarity functions cause these observations. We further describe an alternative similarity function and its deviating characteristics which contrast the observed properties in standard fuzzy clustering. Empirical evidence is gathered for different choices of similarity functions and for some parameterizations in Section IV.

II. Illustrative Example

To demonstrate the effects of irrelevant attributes we generated an artificial data set. We used two classes with 150 data points each and a high number of normally distributed attributes. However, only one of the input dimensions is grouping the generated example data. The generating model as well as the data set is shown in Figure 1, with the relevant attribute on the horizontal axis and one of the irrelevant attributes on the vertical. The clustering attribute is distributed with mean $\mu = 3.5$ and variance $\sigma^2 = 1$ in the left cluster and with $\mu = 6.5$, $\sigma^2 = 1$ in the right cluster. All other attributes have $\mu = 5$ and $\sigma^2 = 1$ in both clusters and thus do not provide any information for grouping the data points.

The interesting effects become visible when this artificial data set is endowed with an increasing number of irrelevant input dimensions. The result of FCM clustering with the relevant attribute and only one noisy attribute is shown in Figure 2. The two cluster centers are very slightly repelling each other due to the partitioning property of probabilistic FCM, i.e., their distance is slightly larger than in the generating model. This situation changes in the cluster partitions when more and more irrelevant dimensions are added. Then the cluster centers seem to get attracted to each other and move closer. With eight irrelevant attributes the clusters are about to collapse (see Figure 3). The cluster centers finally coincide in the center of the data cloud when 9 out of the 10 input attributes are irrelevant (see Figure 4). All data points are assigned to both clusters with equal membership degree of 0.5 in the resulting probabilistic cluster partition.

III. Analysis

To understand the above effects better it is worthwhile to look deeper into the computation of the membership degrees. Since the cluster centers are centers of gravity in the clouds of weighted data points, their attraction and coincidence must occur because of changes of the weights u_{ij}^m of the data points. The weight calculations as given in Equation 4 can be divided into two individual steps: the computation of un-normalized membership degrees first, followed by the normalization step:

$$u_{ij}^* = f(d_{ij}) = d_{ij}^{-\frac{2}{m-1}}, \tag{6}$$

$$u_{ij} = \frac{u_{ij}^*}{\sum_{t=1}^{c} u_{tj}^*}. \tag{7}$$

In Equation 6 the un-normalized membership degrees are a function of the distance d_{ij} of the data point $\vec{x}_j$ to the cluster center $\vec{c}_i$. We call this function f *similarity function*, because it assigns different membership degrees for data points to clusters depending how similar they are to each other. These un-normalized membership degrees are normalized in Equation 7 in order to satisfy the constraints stated in Section I.

The illustrative example shows that in the case of collapsing clusters all membership degrees of the data points are equal. In the general case, when such a cluster center coincidence occurs, we have $u_{ij} = 1/c$, $i \in \{1, \ldots, c\}$, $j \in \{1, \ldots, n\}$.

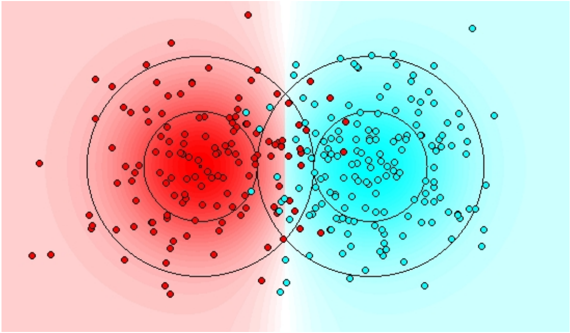

Fig. 1. The generating model.

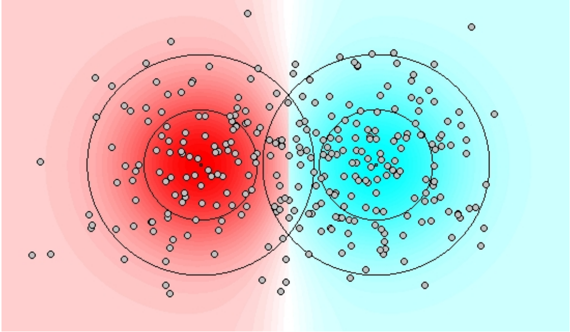

Fig. 2. FCM ($m = 2$) and one irrelevant dimension.

It follows that with an increasing number of noisy inputs the ratios of the membership degrees,

$$\frac{u_{ij}}{u_{tj}} = \frac{u_{ij}^*}{u_{tj}^*} = \frac{f(d_{ij})}{f(d_{tj})}, \qquad i, t \in \{1, \ldots, c\}, \tag{8}$$

of each data point $\vec{x}_j$ get closer to 1. Therefore, in order to explain the observed effects we examine the magnitudes of the distances d_{ij} and the influence of the membership function when the number of irrelevant features increases.

A. Increased Distances

In clustering algorithms distance is usually measured w.r.t. all attributes. For instance, the Euclidean distance used in the FCM algorithm is an aggregate of attribute specific distances $d_{ij,k}^2 = (x_{j,k} - c_{i,k})^2$, where k specifies the k-th attribute or the k-th feature vector component. With the set of relevant attributes A_{rel} and a set of irrelevant and noisy attributes A_{irr} the squared Euclidean distance can be written as

$$d_{ij}^2 = \sum_{k \in A_{\text{rel}}} d_{ij,k}^2 + \sum_{p \in A_{\text{irr}}} d_{ij,p}^2. \tag{9}$$

That is, distances w.r.t. the irrelevant dimensions are added to the distances w.r.t. those dimensions which group the data. Therefore the resulting distance is larger than a distance measured w.r.t. the relevant attributes only. Consequently, the more irrelevant attributes are present, the more the true dissimilarity w.r.t. A_{rel} diminishes in value—and thus in importance—compared to the dissimilarity measured w.r.t. A_{irr}.

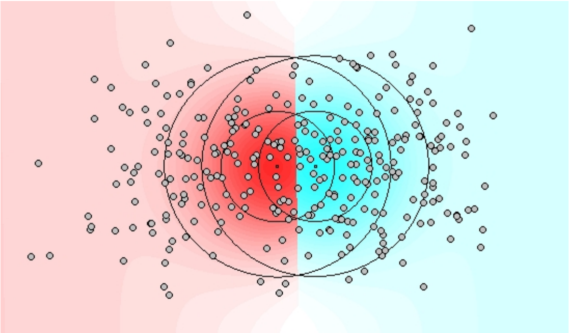

Fig. 3. FCM ($m = 2$) with 8 irrelevant dimensions.

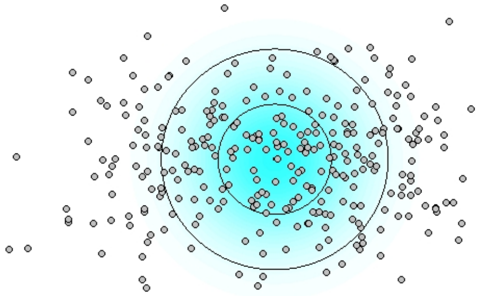

Fig. 4. FCM ($m = 2$) and 9 irrelevant attributes.

In addition, similarity information is lost due to noise that is invariably present along the irrelevant input dimensions. W.r.t. the relevant attributes the within-cluster distances are small whereas the inter-cluster cluster distances are significantly larger. There is some variation in both, intra-cluster as well as inter-cluster distance, but they can easily be distinguished. However, if the distances w.r.t. the irrelevant attributes, which are randomly dispersed, are added to these distance values, the variation in both the intra-cluster as well as the inter-cluster distances increases. It may increase so much that these two types of distances cannot be told apart any more. As a consequence, clusters are the harder to determine the more irrelevant attributes are present, because a higher number of noisy features leads to higher variance of the measured distances. Note that this effect would be present even if the total distance would not be increased on average.

Summarizing, we see that with a higher number of irrelevant attributes the distances d_{ij} of a data point $\vec{x}_j$ to the different clusters tend to get (1) larger and (2) more dispersed. Here we focus on the first effect. For a clustering algorithm it is equivalent to moving a data point $\vec{x}_j$ further away from the cluster centers, as illustrated in Figure 5 for two clusters. If the data point $\vec{x}_j$ in the figure appears to be (moved) further away from the centers due to irrelevant features, $d_{1j} \approx d_{2j}$ since $d_{ij} \gg \delta$ for $i = \{1, 2\}$. In the following we study the influence of membership functions on the relative differences of membership degrees (Equation 8) when the magnitudes of measured distances to clusters have changed as described.

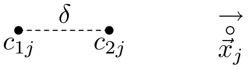

Fig. 5. A simple two cluster setting.

B. Influence of the Similarity Function

A similarity function f is a strictly monotonic decreasing function: a high degree of similarity (membership) is assigned to the cluster close to data point, whereas a lower similarity (membership) is assigned to clusters that are further away. The similarity function f in the FCM (see Equation 6) can be seen as a special case of a generalized Cauchy function with two parameters:

$$f_{cauchy}(d_{ij}; a, b) = \frac{1}{d_{ij}^a + b}, \tag{10}$$

where the exponent $a = \frac{2}{m-1}$ and the reference radius $b = 0$. In the presence of many irrelevant inputs the influence of a membership function depends on how it judges smaller differences in distance, especially when the overall distances of the data points to clusters are high. In the two cluster example we have for the relative difference in degrees of membership with increasing distance to the clusters (in horizontal direction):

$$\begin{aligned} \lim_{d_{2j}\to\infty} \frac{u_{2j}}{u_{1j}} &= \lim_{d_{2j}\to\infty} \frac{f(d_{2j})}{f(d_{2j}+\delta)} \\ &= \lim_{d_{2j}\to\infty} \frac{d_{2j}^2 + 2d_{2j}\delta + \delta^2}{d_{2j}^2} \\ &= \lim_{d_{2j}\to\infty} \left(1 + \frac{2d_{2j}\delta + \delta^2}{d_{2j}^2}\right) \\ &= 1 + 0 = 1, \end{aligned} \tag{11}$$

with $m = 2$ such that $f = f_{\text{cauchy}}(d_{ij}; 2, 0)$. That is, the ratio of the degrees of membership (here: u_{2j}/u_{1j}) to both clusters approaches 1 the more irrelevant input dimensions are in the data. The same situation can also be shown to exist in the general case of more than two clusters: when the distances of a data point to the cluster centers are high without being significantly different, the Cauchy functions tend to assign almost equal degrees of membership. Due to this property we observe the attraction of clusters at high numbers of irrelevant features as well as coincidental clusters in the worst case.

This behavior of the Cauchy similarity function is known to the extent that counter-intuitive membership degrees are assigned to data points which are outliers or just further away from the bulk of data. Such data points are assigned with membership degrees of about $1/c$ to all clusters. Noise clustering approaches have been proposed for gathering such data points in a noise cluster [3], [4]. When clustering data with mainly relevant input dimensions the noise cluster approaches can prevent the counter-intuitive membership degrees. In presence of an increasing number of irrelevant attributes, however, the assignment of data points to the noise cluster actually has a deteriorating effect: the farther away from all clusters a data point is, the higher will be its degree of membership to the noise cluster. Rather than solving the problem, this tendency promotes the attraction of clusters, because the part of the data point weight that is assigned to the noise cluster reduces the "attraction" exerted by the data points on the outskirts of the data set. Hence the clusters are even more likely to move to the center of gravity of the data set. As a consequence, with noise clustering approaches, we observed cluster coincidence already at even lower numbers of irrelevant attributes.

With irrelevant features it would certainly be better to use a membership function that counteracts the situation that relative differences in distance to the clusters diminish with higher overall distance to the clusters. A better suited membership function would still give a significantly higher membership degree to the cluster that is closest and a much lower membership to clusters further away. The Gaussian function

$$f_{\text{gauss}}(d_{ij}) = e^{-\frac{1}{2}d_{ij}^2} \tag{12}$$

has this property. For the simple two cluster example we get:

$$\begin{aligned}
\lim_{d_{2j}\to\infty} \frac{u_{2j}}{u_{1j}} &= \lim_{d_{2j}\to\infty} \frac{f(d_{2j})}{f(d_{2j}+\delta)} \\
&= \lim_{d_{2j}\to\infty} \frac{e^{-\frac{d_{2j}^2}{2}}}{e^{-\frac{(d_{2j}+\delta)^2}{2}}} \\
&= \lim_{d_{2j}\to\infty} \frac{e^{\frac{d_{2j}^2}{2}} e^{d_{2j}\delta} e^{\frac{\delta^2}{2}}}{e^{\frac{d_{2j}^2}{2}}} \\
&= \lim_{d_{2j}\to\infty} e^{d_{2j}\delta} e^{\frac{\delta^2}{2}} = \infty, \qquad (13)
\end{aligned}$$

with $f = f_{\text{gauss}}$. That is, when a data point is moved further away (i.e., the overall distance to clusters increases), the relative difference in degree of membership gets even more expressed. Due to this relation this function implements a winner-takes-all-principle in the limit: the data point will be assigned to the closest cluster, with higher distances this assignment becomes even stronger, and gets exclusive for the distance approaching infinity. Thus we expect the Gaussian function to be more robust than the Cauchy function. High numbers of irrelevant attributes should not cause clusters to get attracted or to coincide when f_{gauss} is used.

The abovementioned behavior of the Gaussian function can also be observed in soft learning vector quantization (SLVQ) when a Gaussian mixture approach is used [5]. In SLVQ the (un-normalized) assignment probabilities of a data point x_j to prototype i are given by $\exp(d_{ij}^2/2\sigma^2)$, where σ is the width of the Gaussian component densities. If the width σ goes to zero, the assignment probabilities become hard assignments (winner-takes-all case) [5]. Smaller widths ($\sigma \to 0$) and increasing distance to the cluster ($d_{ij} \to \infty$) are equivalent for this behavior of the Gaussian function: when σ approaches 0 the entire exponent $d_{ij}^2/2\sigma^2$ approaches ∞ just as the exponents in the equation above.

Apparently the way in which membership degrees are assigned by f_{cauchy} and f_{gauss} differs because of their different asymptotic behavior for increasing distance. For larger distances ($d_{ij} \to \infty$) both functions approach membership degrees of 0. The difference between the two functions, however, lies in the strength of their monotonic descent: the Gaussian function is falling much more steeply than the Cauchy membership function, thus avoiding the deteriorating effects that results from the enlarged distances.

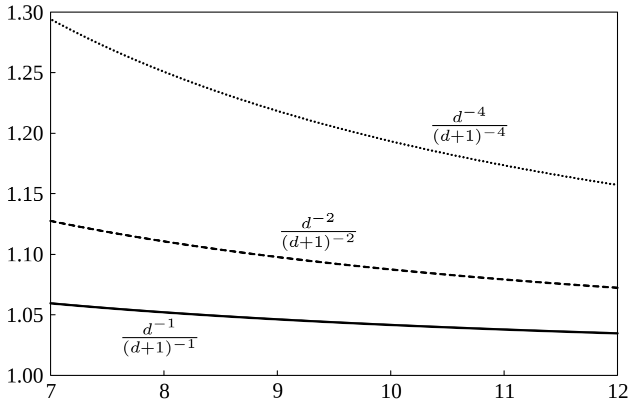

Fig. 6. Ratios of membership degrees for increasing distance.

C. The Role of the Fuzzifier

The generalized Cauchy function contains the fuzzifier as a parameter of the membership computation (see Equation 10). Since $a = \frac{2}{m-1}$, fuzzifiers $m > 2$ lower the value of the exponent in the Cauchy function. Hence cluster boundaries get softer and the Cauchy function $f_{\text{cauchy}}(d_{ij}, a, 0)$ falls less steeply. On the other hand, fuzzifiers $m < 2$ result in harder cluster boundaries and the Cauchy function falls more steeply. The ratios of membership degrees for different choices of the fuzzifier in the Cauchy function can be seen in Figure 6 for the simple two cluster example ($\delta = 1$). With increasing overall distance to the clusters and lower values of the exponent, the tendency of the membership ratio to approach 1 is stronger than for high exponents in f_{cauchy}. From the example it can be seen that the steeper Cauchy functions (with higher exponents a) assign almost equal degrees of membership only at higher distances. Therefore it can be expected that the observed effects of cluster attraction and coincidence will be more pronounced for lower exponents a and may occur even with a low number of irrelevant features.

The considerations for $m < 2$ above are complemented by a reformulation of the fuzzy c-means in [6], [7]. In the limiting case when the fuzzifier m approaches 1 from above ($m \to 1_+$), the membership degrees correspond to the nearest prototype condition in classical learning vector quantization. That is, data points are assigned to the closest (cluster) prototype with full weight and have no association to other clusters (hard memberships). Due to such hardened cluster assignments we expect that smaller values for the weighting exponent should make the clustering algorithm less sensitive to higher numbers of noisy attributes. On the other hand, a stronger tendency towards attracting or coinciding clusters can be expected if softer cluster partitions are desired.

TABLE I

NUMBERS OF IRRELEVANT ATTRIBUTES AT WHICH CLUSTERS COINCIDE

m	1.1	1.5	2	3	6	15	100	1000
no. irr.	-	15	9	6	3	3	3	3

TABLE II

FOUND CLUSTER CENTERS WITH f_{gauss}

no. of irr. attributes	cluster 1	cluster 2
1	3.24341	6.61652
10	3.26322	6.62256
20	3.26294	6.60603
40	3.31177	6.62485

IV. EXPERIMENTS

In our experiments we investigated the tendency for cluster attraction and coincidence with an increasing numbers of noisy attributes using the example dataset described in Section II. We clustered the dataset using the fuzzy c-means algorithm with different values for the fuzzifier m to validate the influence of the Cauchy function. For each parameter value we executed FCM several times, adding one more irrelevant inputs in each step. For an increasing number of irrelevant inputs we observed the attraction of cluster centers as expected for the Cauchy function. The values $m = 1.1$ and $m = 1.5$ resulted in the weakest attraction observed. The attraction of the two clusters got stronger even at lower numbers of noisy attributes for $m \geq 2$. For $m = 1.1$ we did not observe coinciding clusters even for 25 irrelevant features. In all other cases clusters coincided after they had been attracting each other. The numbers of irrelevant inputs at which clusters coincided are summarized in Table I for all tested values of m. For higher weighting exponents (softer clustering) clusters coincided already at lower counts of irrelevant input features. We can conclude an increased tendency for the studied effects at higher degrees of fuzziness, which complies with the expectations formed in the preceding section.

In the second part of our experiments we used the Gaussian membership function and again increased the number of noisy features step by step. However, we did not observe attracting or coinciding clusters. Table II shows that even for a fairly high number of irrelevant features the found cluster coordinates stayed close to the corresponding values in the generating model. The experiments also showed that data points were strongly assigned either to the left or the right cluster. These results comply with the considerations of the properties of f_{gauss} above and the expectation formed.

V. CONCLUSIONS

In standard fuzzy clustering clusters can attract each other or even collapse in presence of irrelevant features. This tendency is stronger when fuzzier cluster partitions are desired ($m > 2$). The effects are caused due to properties of the Cauchy membership function. We found the Gaussian function to be more robust against irrelevant input dimensions.

Irrelevant features lead to noisier and in average increasing distance values while relative differences in dissimilarity of points to the clusters diminish. In such settings the tendency for attracting or coincidental clusters is weaker the more expressed differences in the assigned memberships are. Using the Gaussian function higher relative differences in degrees of membership are obtained due to the increased overall distances. Using the Cauchy functions, however, significantly different membership are assigned only for harder values of the fuzzifier ($m \rightarrow 1_+$).

REFERENCES

[1] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Clustering*. J. Wiley & SOns, Chichester, United Kingdom 1999

[2] J.C. Bezdek. *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York, NY, USA 1981

[3] R.N. Davé. Characterization and Detection of Noise in Clustering. *Pattern Recognition Letters* 12:657–664. Elsevier Science, Amsterdam, Netherlands 1991

[4] R.N. Davé and R. Krishnapuram. Robust Clustering Methods: A Unified View. *IEEE Trans. on Fuzzy Systems* 5:270–293. IEEE Press, Piscataway, NJ, USA 1997

[5] S. Seo and K. Obermayer. Soft Learning Vector Quantization. *Neural Computation* 15(7):1589–1604. MIT Press, Cambridge, MA, USA 2003

[6] N.B. Karayiannis and J.C. Bezdek. An Integrated Approach to Fuzzy Learning Vector Quantization and Fuzzy c-means Clustering. *IEEE Trans. on Fuzzy Systems* 5(4):622–628. IEEE Press, Piscataway, NJ, USA 1997

[7] N.B. Karayiannis and P.-I. Pai. Fuzzy Algorithms for Learning Vector Quantization. *IEEE Trans. on Neural Networks* 7:1196–1211. IEEE Press, Piscataway, NJ, USA 1996