Adaptation of Cluster Sizes in Objective Function Based Fuzzy Clustering

Annette Keller
Institute for Flight Guidance
German Aerospace Center
Lilienthalplatz 7
D-38108 Braunschweig
Germany

Frank Klawonn
Dept. of Electrical Engineering
and Computer Science
Ostfriesland University of
Applied Sciences
Constantiaplatz 4
D-26723 Emden, Germany

Abstract

This paper discusses new approaches in objective function based fuzzy clustering. Some well-known approaches are extended by a supplementary component. The resulting new clustering techniques are able to adapt single clusters to the expansion of the corresponding group of data in an iterative optimization procedure. Another new approach based on volume centers as cluster representatives with varying radii for individual groups is also described. The corresponding objective functions are presented and alternating optimization schemes are derived. Experimental results demonstrate the significance of the presented techniques.

Keywords

Objective function based fuzzy clustering, size adaptable clustering, alternating optimization

1 Introduction

Standard fuzzy clustering methods like the fuzzy c-means algorithm are based on the idea of optimizing an objective function. This objective function depends on the distances of the data to the cluster centers weighted by the membership degrees. By taking the first derivative of the objective function with respect to the cluster parameters, one obtains necessary conditions for the objective function to receive an optimum. These conditions are then applied in an iteration procedure and define a clustering algorithm. Numerous approaches have been developed to detect different forms of cluster shapes in data sets. The more flexible the clustering algorithms are in general, the more they depend on a suitable cluster initialization. Also with the flexibility of cluster structures the complexity of the proposed algorithms highly increases. In this article we present an extension that can be applied to well-known simple and fast clustering techniques enabling these to adapt to the cluster sizes without highly increasing the computational effort.

In section 2 we review the necessary background on objective function based fuzzy clustering techniques. Well-known and often applied clustering techniques as probabilistic, possibilistic, and noise clustering are discussed as well as several possible distance measures. The restrictions imposed by the algorithms lead us to the idea of our new approaches. In section 3 we show that only slight modifications of the basic ideas are necessary to enable better adaptations by clustering. Nevertheless our new techniques are not based on special cluster shapes that could be restrictive for the underlying models. The results of the proposed methods compared to the basic algorithms presented in section 2 are illustrated in section 4. There some artificial data sets are used to demonstrate the possibilities of the presented algorithms. The last example consists of real world data and demonstrates the applicability of the new approaches. Section 5 finally summarizes our experiences.

2 Objective Function Based Fuzzy Clustering

In this section we present a short introduction to objective function based fuzzy clustering and describe some well known and often applied techniques. For a detailed overview on fuzzy clustering see for example [8]. Most objective function based fuzzy clustering algorithms aim at minimizing an objective function that evaluates the partition of data into a given number of clusters.

2.1 Basic Objective Functions

Before discussing several special clustering techniques, general forms of objective functions for fuzzy clustering are introduced that still depend on the choice of a suitable distance measure. Two very common basic clustering techniques are probabilistic and possibilistic clustering. Both depend on a distance or dissimilarity measure weighted by the membership degrees. Probabilistic clustering [1] uses a constraint ensuring that all data points totally belong to the partition, whereas possibilistic clustering [15] considers outliers with small membership degrees to all groups of data. A third approach related to possibilistic clustering is called *noise clustering* [3]. The idea of this approach is to assign outliers to a special group of data called noise cluster and reduce the influence of this group on the whole partition. Selim and Ismail [18] introduced other approaches to avoid the drawback of probabilistic clustering. They suggest to let a datum belong to a maximum number of clusters, to set the membership degrees to zero if a predefined maximal distance is exceeded, or to define a minimum threshold for the membership degrees.

2.1.1 Probabilistic Clustering

In case of probabilistic fuzzy clustering the objective function has the form

$$J^{prob}(X, U, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^{m} \cdot d^2(v_i, x_k) \tag{1}$$

$X = \{x_1, \ldots, x_n\} \in \mathbb{R}^p$ is the data set, n the number of data points, c denotes the number of fuzzy clusters, $u_{ik} \in [0, 1]$ is the membership degree

of datum x_k to cluster i, v_i is the prototype or the vector of parameters for cluster i, and $d(v_i, x_k)$ is a distance between prototype v_i and datum x_k. The parameter $m > 1$ is called fuzziness index. For $m \to 1$ the clusters tend to be crisp, i.e. either $u_{ik} \to 1$ or $u_{ik} \to 0$, for $m \to \infty$ we have $u_{ik} \to 1/c$. Usually $m = 2$ is chosen.

To avoid the trivial solution that all membership degrees u_{ik} are 0, constraints have to be taken into account. In this case the constraints are

$$\sum_{k=1}^{n} u_{ik} \; > \; 0 \qquad \text{for all } i \in \{1, \ldots, c\} \tag{2}$$

and

$$\sum_{i=1}^{c} u_{ik} \; = \; 1 \qquad \text{for all } k \in \{1, \ldots, n\}. \tag{3}$$

Constraint (2) guarantees that only non-empty clusters are admitted in the partition. Constraint (3) ensures that the sum of all membership degrees for one datum equals 1.0. This can be interpreted as "each datum is fully divided among the clusters and belongs totally to the partition of the data set".

Differentiating (1) and taking the constraints into account by a Lagrange function leads to the necessary condition

$$u_{ik} \; = \; \frac{1}{\sum_{j=1}^{c} \left(\frac{d^2(v_i, x_k)}{d^2(v_j, x_k)} \right)^{\frac{1}{m-1}}} \tag{4}$$

for (1) to have a (local) minimum. Therefore, equation (4) is used in an iteration procedure for updating the membership degrees u_{ik}. If a suitable distance function and parameter form is chosen, equations for the prototypes can be derived analogously, assuming the membership degrees are fixed. The alternating optimization scheme starts with a random initialization and applies the equations for the u_{ik} and the prototypes until the difference between the membership matrices $\left(u_{ik}^{old}\right)$ and (u_{ik}^{new}) in two succeeding iterations is smaller than a given bound ε.

2.1.2 Possibilistic Clustering

In probabilistic clustering the strong constraint (3) possibly leads to undesirable membership degrees of some data. Assume a data point in great distance to all clusters exists in the data set. This noise point would be assigned the same high membership degree $\frac{1}{c}$ to all c clusters and therefore would have a greater influence on the partition than desired. To avoid such a drawback the approach of possibilistic clustering was introduced with remaining constraint (2) but modified constraint (3):

$$\sum_{i=1}^{c} u_{ik} \; > \; 0 \qquad \text{for all } k \in \{1, \ldots, n\}. \tag{5}$$

With these constraints the membership degree u_{ik} could be interpreted as a degree of representativeness of datum x_k for cluster i. To avoid the trivial solution all $u_{ik} \to 0$ by minimizing equation (1) considering constraint (5) the objective function has to be modified as well (6).

$$J^{poss}(X, U, v) \; = \; \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^{m} \cdot d^2(v_i, x_k) \; + \; \sum_{i=1}^{c} \eta_i \; \sum_{k=1}^{n} (1 - u_{ik}^{m}) \tag{6}$$

The additional parameter η_i determines the permissible extension of cluster i. Differentiating (6) considering the constraints (5) and (2) leads to

$$u_{ik} \; = \; \frac{1}{1 + \left(\frac{d^2(v_i, x_k)}{\eta_i} \right)^{\frac{1}{m-1}}} \tag{7}$$

To illustrate the influence of η_i, assume $\eta_i = d^2(v_i, x_k)$. The resulting membership degrees are $u_{ik} = \left(1 + 1^{\frac{1}{m-1}}\right)^{-1} = \frac{1}{2}$. Defining a membership degree of $\frac{1}{2}$ as lower bound for assigning a data point x_k to cluster i gives parameter η_i the mentioned meaning. The permissible extension of cluster i is in some way defined by η_i. If the cluster shapes are known in advance, η_i could be estimated for all $i = 1, \ldots, c$ easily. Otherwise additional assumptions have to be made. One possible approach is to assume clusters containing about the same number of data points and estimate

$$\eta_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} \cdot d^2(v_i, x_k)}{\sum_{k=1}^{n} u_{ik}^{m}}. \tag{8}$$

Krishnapuram and Keller [15] have also proposed other methods to estimate the parameters η_i.

2.1.3 Noise Clustering

Possibilistic clustering is one approach to deal with noisy data. Another related technique is called *noise clustering*, see e.g. [3] or [4] and the references therein. The principle idea is to add one noise cluster to the set of clusters. Since the objective function considers only the distance function and the membership degrees, the noise cluster could be represented by the weighted membership degrees of the data to this cluster. The second term in equation (9) expresses the noise cluster.

$$J^{noise}(X,U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} \cdot d^2(v_i, x_k) + \sum_{k=1}^{n} \delta^2 \left(1 - \sum_{i=1}^{c} u_{ik}\right)^m \tag{9}$$

Parameter δ has to be chosen in advance and is supposed to be the (large) constant distance of each datum to the noise cluster. In this case the constraints (2) and (5) have to be considered as in possibilistic clustering in order to derive equations for the membership degrees

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d^2(v_i,x_k)}{d^2(v_j,x_k)}\right)^{\frac{1}{m-1}} + \left(\frac{d^2(v_i,x_k)}{\delta^2}\right)^{\frac{1}{m-1}}} \tag{10}$$

as necessary conditions for (9) to have a minimum. An interesting result is that

$$\sum_{i=1}^{c} u_{ik} < 1 \qquad \text{for all } k \in \{1,\ldots,n\}. \tag{11}$$

This illustrates that each datum belongs at least with a small membership degree to the noise cluster. In 1984 Ohashi [16] already made an attempt to consider noise in data. Davé and Krishnapuram [4] showed that the minimization of Ohashi's objective function is equivalent to the presented approach introduced by Davé [3].

2.2 Distance measures

In the previous section several general clustering concepts have been described. All techniques rely on the definition of suitable distance measures. Choosing a certain dissimilarity measure defines the structure which is searched for in the sample data. Different distance measures that are able to describe varying forms or shapes of clusters are possible.

2.2.1 The Fuzzy c-Means Algorithm

One simple fuzzy clustering technique is the fuzzy c-means algorithm (FCM), see e.g. [1], where the distance $d(v_i, x_k)$ is simply the Euclidean distance

$$\mathfrak{D}_{FCM} \;=\; d^2(v_i, x_k) \;= \parallel x_k - v_i \parallel^2 = \sum_{\nu=1}^{p} \left(x_k^{(\nu)} - v_i^{(\nu)} \right)^2 \tag{12}$$

and the prototypes are vectors $v_i \in \mathbb{R}^p$, where p is the dimensionality of the data. $x_k^{(\nu)}$ $\left(v_i^{(\nu)}\right)$ denote the ν'th coordinate of the data vector (cluster center representative). Due to the Euclidean distance measure, this technique searches for spherical clusters of approximately the same size. By differentiating (1), (6) or (9) we obtain the necessary condition

$$v_i \;=\; \frac{\sum_{k=1}^{n} u_{ik}^m \cdot x_k}{\sum_{k=1}^{n} u_{ik}^m} \tag{13}$$

as prototype calculation instruction for the objective functions to have a (local) minimum using $\mathfrak{D}_{FCM}$. Since the first summand is identical in the three mentioned objective functions (see section 2.1) and the second term in equations (6) and (9) does not depend on a certain distance measure, the derived prototype equation holds in all three cases. These prototypes could be used alternately with (4) in the iteration procedure. The update equation for the membership degrees depends on the chosen basic objective function as described in the previous section.

2.2.2 The Algorithm by Gustafson and Kessel

Gustafson and Kessel [7] designed a fuzzy clustering method that is able to adapt to hyper-ellipsoidal forms. The prototypes consist of the cluster

centers v_i as in FCM and (positive definite) covariance matrices C_i. The Gustafson and Kessel algorithm (GK) replaces the Euclidean distance by the transformed Euclidean distance

$$\mathfrak{D}_{GK} = d^2(v_i, x_k) \;=\; (\rho_i \det C_i)^{1/p} \cdot (x_k - v_i)^\top C_i^{-1}(x_k - v_i). \qquad (14)$$

The factor $(\rho_i \det C_i)^{1/p}$ in $\mathfrak{D}_{GK}$ guarantees the volume for all cluster to be constant. Factor ρ_i could be used to determine the size of cluster i and is not changed during the alternating optimization. If the sizes cannot be estimated in advance, the parameters ρ_i might be set to one. The covariance matrices C_i are computed using equation (15).

$$C_i = \sum_{k=1}^{n} u_{ik}^m \cdot (x_k - v_i)(x_k - v_i)^\top. \qquad (15)$$

The prototype calculation instruction of equation (13) does not depend on the norm used in the distance measure, so again we obtain

$$v_i \;=\; \frac{\sum_{k=1}^{n} u_{ik}^m \cdot x_k}{\sum_{k=1}^{n} u_{ik}^m}$$

as a necessary condition for the objective functions (1), (6), or (9) to have a (local) minimum. With these equations the alternating iteration procedure for the Gustafson-Kessel algorithm is defined. In the update equation for the membership degrees $\mathfrak{D}_{GK}$, see equation (14), has to be inserted as distance measure. This general form searches for hyper-ellipsoidal forms in the domain of interest.

Considering e.g. the task of rule learning, where fuzzy clusters are projected to single dimensions, non-axes parallel ellipsoids would lead in general to a serious loss of information caused by the construction of the fuzzy sets for the single domains. One approach to avoid this drawback is to restrict the covariance matrices C_i to diagonal matrices resulting in axes-parallel hyper-ellipsoids [11, 14]. The distance measure in this case can be rewritten as

$$\mathfrak{D}_{AGK} \;=\; d^2(v_i, x_k) \;=\; \left(\rho_i \prod_{\nu=1}^{p} c_i^{(\nu)}\right)^{1/p} \cdot \left(\sum_{\nu=1}^{p} (x_k^{(\nu)} - v_i^{(\nu)})^2 \cdot \frac{1}{c_i^{(\nu)}}\right). \qquad (16)$$

Here, p is the dimensionality of the data vectors and $x_k^{(\nu)}, v_i^{(\nu)}$ denote the ν'th component of the k'th data point, i'th cluster center, respectively. For the alternating optimization the calculation instruction for the covariance matrices can be simplified in the following way

$$c_i^{(\gamma)} = \sum_{k=1}^{n} u_{ik}^m \cdot (x_k^{(\gamma)} - v_i^{(\gamma)})^2, \tag{17}$$

where $c_i^{(\gamma)}$ denotes the γ'th diagonal element of the covariance matrix. $\mathfrak{D}_{AGK}$ could be used as distance measure in the membership update equations from section 2.1.

2.2.3 The Algorithm by Gath and Geva

Another clustering technique (GG) was designed by Gath and Geva [5]. This extension of the Gustafson-Kessel Algorithm is in some way able to adapt the cluster size and like the GK adapts to hyper-ellipsoidal forms. Actually this approach is not based on an objective function optimizer. Instead the GG is a heuristic method derived from the fuzzification of a maximum likelihood estimator. Here, the distance measure is of the following form:

$$\mathfrak{D}_{GG} \;=\; d^2(v_i, x_k) \;=\; \frac{1}{\pi_i} \;\cdot\; \sqrt{\det(A_i)} \;\cdot\; \exp^{\left(\frac{1}{2}\cdot(x_k - v_i)^\top A_i^{-1}(x_k - v_i)\right)}. \tag{18}$$

The parameter π_i denotes the a-priori probability for a datum to belong to the i-th normal distribution. π_i is estimated as described by equation (19), i.e. *"number of data belonging to cluster i in relation to total number of data"*.

$$\pi_i \;=\; \frac{\sum_{k=1}^{n} u_{ik}^m}{\sum_{j=1}^{c}\sum_{k=1}^{n} u_{jk}^m} \tag{19}$$

The covariance matrix of the i-th normal distribution is denoted by A_i, where (20) is the calculation instruction for estimating matrix A_i.

$$A_i \;=\; \frac{\sum_{k=1}^{n} u_{ik}^m \cdot (x_k - v_i)(x_k - v_i)^\top}{\sum_{k=1}^{n} u_{ik}^m} \tag{20}$$

The prototype coordinates are the estimated expected values of the assumed normal distribution for cluster i. Again the calculation of the prototypes can

be done using equation (13) as in the FCM or GK, respectively:

$$v_i \;=\; \frac{\sum_{k=1}^{n} u_{ik}^m \cdot x_k}{\sum_{k=1}^{n} u_{ik}^m}.$$

Now the equations for the alternating iteration procedure of the Gath-Geva algorithm are complete. To obtain equations for the prototypes as a necessary condition for the optimization function having a local minimum, the objective functions described in section 2.1 would have to be differentiated. Using $\mathfrak{D}_{GG}$ as distance measure would lead to equations for which no analytical solution exists. Therefore, the estimation analogous to probability theory provides a good heuristic method.

Since GG is able to adapt to hyper-ellipsoidal forms as well as to different cluster sizes, the same problems as with the Gustafson-Kessel algorithm arise in the task of rule learning. As with GK it is possible to restrict this approach to detect axes-parallel hyper-ellipsoids [11, 14]. Nevertheless, the axes-parallel version of the algorithm introduced by Gath and Geva is able to adapt to different cluster sizes. In this case the distance measure can be rewritten as

$$\mathfrak{D}_{AGG} \;=\; d^2(v_i, x_k) \;=\; \frac{1}{\pi_i} \sqrt{\prod_{\nu=1}^{p} a_i^{(\nu)}} \;\cdot exp^{\frac{1}{2}\cdot\left(\sum_{\nu=1}^{p}(x_k^{(\nu)}-v_i^{(\nu)})^2\cdot\frac{1}{a_i^{(\nu)}}\right)}. \qquad (21)$$

Again, p denotes the dimensionality of the data, $x_k^{(\nu)}$ and $v_i^{(\nu)}$ designate the ν'th component of the k'th data point, i'th cluster center, respectively. In the alternating optimization the covariance matrices could be simplified in the following way:

$$a_i^{(\gamma)} = \frac{\sum_{k=1}^{n} u_{ik}^m \cdot (x_k^{(\gamma)} - v_i^{(\gamma)})^2}{\sum_{k=1}^{n} u_{ik}^m}, \qquad (22)$$

where $a_i^{(\gamma)}$ denotes the γ'th diagonal element of the covariance matrix. Parameter π_i is estimated again as denoted in (19).

In [21] other clustering methods based on the maximum likelihood principle are described.

2.3 Alternating Optimization Approaches

In order to increase the influence of the user in extracting functional models from data, Runkler and Bezdek, see e.g. [17] and the references therein, have developed alternative approaches based on the presented basic ideas. They call the general clustering form with interchanging update equations for prototypes and membership degrees as presented above *alternating optimization*. In one approach the expert has to specify the input space components in form of prototype parameters. In case of the fuzzy c-means algorithm (see section 2.2.1) the expert would have to state prototype coordinates for each input domain. For other clustering algorithms also a suitable distance measure would have to be chosen. The components for the output space are alternately updated during the optimization phase of that algorithm. Runkler and Bezdek call this form of alternating optimization *regular alternating optimization, rAO*. Since some parameters are defined by the user and do not have to be updated during the alternating optimization the computational effort is reduced.

By *alternating cluster estimation, ACE*, Runkler and Bezdek denote a clustering method where the expert has to select suitable membership function shapes and thereby defines the update equations for the cluster parameters.

The combination of both approaches where the expert has to specify suitable prototype parameters for the input domain as well as to choose suitable membership function shapes is called *regular alternating cluster estimation, rACE*. The algorithm generates a partition of the data and then evaluates the projections of the cluster centers into the output space.

Although the resulting functional models are easy to understand and reflect the experts interpretation of the modeled system, they are not necessarily based on objective functions. Problems may arise if the expert associates a system behavior with the data and assigns suitable parameters for the clustering algorithm but the so defined model has a different basis. The greater the influence of the expert the greater are the restrictions of the associated functional model. Knowledge about unknown dependencies in the data is dif-

ficult to extract with these models, but under assumptions about the system behavior these are computationally fast methods resulting in interpretable and easily understandable functional models.

The algorithms from section 2 can be applied to learn fuzzy rules from data for classification problems [6, 13] or function approximation [12, 14, 20]. Fuzzy rules are usually obtained from fuzzy clusters by projecting the clusters to the coordinate spaces leading to a certain loss of information. The more flexible the cluster algorithms are in finding different shape forms, the greater is the resulting loss of information in rule generation. One method to avoid a major part of this information loss is described in [10]. There we start with a partition of the single domains in fuzzy sets and try to find a suitable partition for the data under consideration. Here we propose another approach. We modify (1) in a way that enables simple fuzzy clustering algorithms like FCM to adapt to the cluster size – meaning to make the algorithm more flexible with respect to the cluster shape – without increasing the existing loss of information in case of rule learning.

3 Adaptation of Cluster Volumes

In this section we present some approaches to adapt to clusters with different volumes. The first approach was previously presented in [9] to reduce the loss of information in rule learning. Therefore, only small modifications of some in section 2 displayed algorithms have to be made. The principle of objective function based fuzzy clustering with cluster representatives in form of real-valued prototypes remains unchanged. The second presented method does no longer use multidimensional center-points as representatives but center-volumes. As we will see, this approach has a non-negligible drawback. A combination of the presented methods seems to eliminate this lack and is presented in section 3.3. Another approach using volume prototypes

to enable the fuzzy c-means to detect clusters with different densities was presented by M. Setnes and U. Kaymak [19].

3.1 Center-Based Clustering Algorithm

For each cluster we introduce an additional parameter τ_i to the objective function in order to enable the clustering algorithm to adapt the cluster volumes. τ_i can be considered as the (relative) radius of the corresponding cluster. The resulting probabilistic objective function is shown in (23), with constant real-valued parameter $l > 0$.

$$J_{SACB}^{prob}(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \cdot \frac{1}{\tau_i^l} \cdot d^2(x_k, v_i) \tag{23}$$

To avoid the trivial solution that all $\tau_i \to \infty$, the constraint

$$\sum_{i=1}^{c} \tau_i = \tau \tag{24}$$

has to be taken into account, where τ is a predefined constant parameter, e.g. $\tau = c$ or $\tau = 1$.

Since the objective function (23) does not require special properties of the distance measure $\mathfrak{D}$, most of the described distance measures need only small modifications to use the advantages of the proposed objective function. Let us define

$$\mathfrak{D}_{SACB,*} = d_\tau^2(x_k, v_i) = \frac{1}{\tau_i^l} \cdot d^2(x_k, v_i) \tag{25}$$

as a new group of distance measures. Then the objective function (23) can be rewritten as

$$J_{SACB,*}^{prob}(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \cdot d_\tau^2(x_k, v_i). \tag{26}$$

Considering constraints (2) and (3) from section 2.1 we obtain the same equations for the membership degrees as in (2.1), except that we have to

replace the old distance $d^2(v_i, x_k)$ by $d_\tau^2(v_i, x_k)$, i.e. in the probabilistic case

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_\tau^2(x_k, v_i)}{d_\tau^2(x_k, v_j)} \right)^{\frac{1}{m-1}}}.$$

Analogously the distance measure for the membership equations in case of possibilistic and noise clustering could be replaced. The modified distance measure for the FCM is shown in (27).

$$\mathfrak{D}_{SACB,FCM} = d_\tau^2(x_k, v_i) = \frac{1}{\tau_i^l} \cdot (x_k - v_i)^T (x_k - v_i) \tag{27}$$

Analogously to the objective function from section 2.1 minimizing (23) leads to the necessary condition (13)

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m \cdot x_k}{\sum_{k=1}^{n} u_{ik}^m}$$

for the evaluation of the prototype coordinates.

Assuming that the parameters $l > 0$ and $\tau > 0$ are fixed, we have to take constraint (24) into account, to determine the values τ_i with predefined and during the iteration procedure unchanged $l > 0$ and $\tau > 0$. So we obtain the Lagrange function

$$J_{SACB,\lambda}^{prob}(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \cdot \frac{1}{\tau_i^l} \cdot d^2(x_k, v_i) + \lambda \left(\sum_{i=1}^{c} \tau_i - \tau \right). \tag{28}$$

Note that the last term of (28) does neither depend on u_{ik} nor on v_i so that the formulae for the optimal choices of the u_{ik} and the v_i remain valid.

Since the distance measure is independent of τ_i, differentiating (28) gives us

$$\frac{\partial J_{SACB,\lambda}^{prob}(X, U, v)}{\partial \tau_i} = -\frac{l}{\tau_i^{l+1}} \cdot \sum_{k=1}^{n} u_{ik}^m \cdot d^2(x_k, v_i) + \lambda \stackrel{!}{=} 0 \tag{29}$$

and therefore

$$\tau_i = \left(\frac{l \cdot \sum_{k=1}^{n} u_{ik}^m \cdot d^2(x_k, v_i)}{\lambda} \right)^{\frac{1}{l+1}}. \tag{30}$$

With (24) λ evaluates to

$$\lambda \;=\; \frac{\left(\sum_{j=1}^{c}\left(l\cdot\sum_{k=1}^{n}u_{jk}^{m}\cdot d^{2}\left(x_{k},v_{j}\right)\right)^{\frac{1}{l+1}}\right)^{l+1}}{\tau^{l+1}}. \tag{31}$$

After inserting (31) in (30), equation (32) represents the resulting calculation instruction for the τ_i.

$$\tau_i \;=\; \frac{\left(\sum_{k=1}^{n}u_{ik}^{m}\cdot d^{2}\left(x_{k},v_{i}\right)\right)^{\frac{1}{l+1}}}{\sum_{j=1}^{c}\left(\sum_{k=1}^{n}u_{jk}^{m}\cdot d^{2}\left(x_{k},v_{j}\right)\right)^{\frac{1}{l+1}}}\cdot\tau \tag{32}$$

The parameter $l > 0$ plays a similar role as the fuzzifier m. When we choose a small value for l, a strong emphasis is put on adapting to the cluster size. Too small values for l can have a negative effect on algorithms as the GK, since the priority is put on the cluster size instead of the cluster shape. For $l \to \infty$, no adaptation of cluster sizes is carried out any more, and we obtain the original algorithms.

Equation (32) can be used alternately with equations (4) and (13) and a suitable distance measure for fuzzy clustering algorithms. We call this group of clustering techniques Size-Adaptable Center-Based clustering algorithms (SACB). Applying our results on the described FCM or GK enables these algorithms to detect clusters of different sizes. In case of the FCM rule generation only results in a small loss of information. Adapting the sizes of the detected spherical structures has no influence on the precision of the resulting fuzzy rules. Also the axes-parallel version of the GK, i.e. AGK, does not lead to a significant loss of information in rule-learning. Not only considering the task of rule learning this approach can be in combination with the GK as well as the AGK an objective function based alternative to the GG or AGG, respectively. It is possible to combine our approach with the objective function approaches of possibilistic (section 2.1.2) or noise clustering (section 2.1.3). The difference of these methods compared to the probabilistic objective function of section 2.1.1 does not depend on a special distance measure. In case of possibilistic clustering, equations (7) for the membership degrees u_{ik}, (32) for the size parameters τ_i, and the necessary conditions derived from the chosen distance measure ($\mathfrak{D}_{FCM}$, $\mathfrak{D}_{GK}$ or $\mathfrak{D}_{AGK}$), e.g. the equations for

cluster centers or covariance matrices could be used in an alternating optimization procedure. Applying noise clustering to the size-adapting clustering approach, equation (10) has to be used to calculate the accessory membership degrees. The other parameters are equivalent to probabilistic and possibilistic clustering.

We call the corresponding alternating optimization incorporating cluster size adaptation the sized algorithm (FCM-sized, GK-sized etc.).

3.2 Volume-Based Clustering Algorithm

In the former described fuzzy clustering techniques the clusters are characterized by a vector, consisting of real-valued attributes, and a distance measure. Only the data points that coincide with a prototype may be assigned to the corresponding cluster with a membership degree of 1.0. Let us imagine dense spherical clusters. Instead of having just one ideal prototype for each cluster to which we calculate the distances of the data points, we now assume that we have a complete circle or (hyper-)ball as the cluster center. This means that data within this area have distance zero to the cluster. This idea was proposed in [19]. However, there it was not based on an objective function, but on pure heuristic considerations. Here we want derive an alternating optimization scheme for this approach as well.

Taking these considerations into account, we obtain a probabilistic objective function (33) reflecting this idea of volume prototypes using (34) as the distance function.

$$J_{SAVB}^{prob}(X,U,v) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \cdot \max\{0,\ (x_k - v_i)^T (x_k - v_i) - \tau_i\} \qquad (33)$$

$$\mathfrak{D}_{SAVB} = d^2(v_i,\ x_k) = max\{0,\ (x_k - v_i)^T (x_k - v_i) - \tau_i\} \qquad (34)$$

If the clusters' radii τ_i are known in advance, these values should be used directly. Otherwise the τ_i have to be adapted during the alternating optimization taking constraint (24) into account, to avoid the trivial solution

$\tau_i \to \infty$ for all $i \in \{1, \cdots, c\}$ in minimizing the objective function (33). Here τ is a predefined constant parameter. Assigning 0 to τ (all τ_i are 0) leads to the previously described fuzzy c-means (FCM) clustering technique, see section 2.2.1. To derive equations for prototype coordinates (37) and radii values (40) respectively, the partial derivatives (36), (38) and (39) of (35) have to be computed.

$$J^{prob}_{SAVB,\lambda}(X,U,v) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \cdot \; d^2\left(v_i,\; x_k\right) \;+\; \lambda \cdot \left(\sum_{i=1}^{c} \tau_i^2 \;-\; \tau\right), \qquad (35)$$

with $d^2\left(v_i,\; x_k\right)$ the distance measure from equation (34).

$$\frac{\partial J^{prob}_{SAVB,\lambda}}{\partial v_i} \;=\; 2 \cdot \sum_{k:(x_k-v_i)^T(x_k-v_i)>\tau_i} u_{ik}^m \cdot \; (v_i - x_k) \stackrel{!}{=} 0 \qquad (36)$$

$$\stackrel{(36)}{\Rightarrow} \; v_i \;=\; \frac{\sum_{k:(x_k-v_i)^T(x_k-v_i)>\tau_i} u_{ik}^m \cdot \; x_k}{\sum_{k:(x_k-v_i)^T(x_k-v_i)>\tau_i} u_{ik}^m} \qquad (37)$$

$$\frac{\partial J^{prob}_{SAVB,\lambda}}{\partial \tau_i} = \sum_{k:(x_k-v_i)^T(x_k-v_i)>\tau_i} u_{ik}^m \;+\; 2 \cdot \lambda \cdot \tau_i \stackrel{!}{=} 0 \qquad (38)$$

$$\frac{\partial J^{prob}_{SAVB,\lambda}}{\partial \lambda} = \sum_{i=1}^{c} \tau_i^2 \;-\; \tau \stackrel{!}{=} 0 \qquad (39)$$

$$\stackrel{(38),(39)}{\Rightarrow} \; \tau_i = \frac{\sum_{k:(x_k-v_i)^T(x_k-v_i)>\tau_i} u_{ik}^m}{\sqrt{\sum_{i=1}^{c}\left(\sum_{k:(x_k-v_i)^T(x_k-v_i)>\tau_i} u_{ik}^m\right)^2}} \cdot \; \sqrt{\tau} \qquad (40)$$

In the alternating optimization scheme the distance measure has to be replaced by (34). Depending on the used clustering technique (probabilistic, possibilistic or noise) the corresponding update equations of the membership degrees u_{ik} (4, 7, or 10) have to be used.

Note that the objective function is not differentiable everywhere. The necessary conditions (36) and (40) lead to a local minimum, if no data points leave a volume center or wander into a volume center. This is why in equations (37) and (40) only data points with Euclidean distance greater τ_i to the prototypes v_i have influence on the next alternating parameters τ_i^{new} and v_i^{new} for cluster i. Imagine two well separated spherical clusters are given. In the first alternating optimization steps the structures are identified correctly. The τ_i are assigned the correct radius values of the circles containing the data points. In the next step each prototype and each radius is only calculated on the basis of the data points assigned to the opposite cluster. So the cluster parameters are alternately interchanged. Even if the τ_i are smaller than the correct radius values, convergence is neither guaranteed nor plausible.

3.3 Combination of Volume-Based and Center-Based Clustering Algorithm

To avoid restrictions as in case of the volume-based clustering technique (SAVB) from section 3.2 we have modified objective function (33) so as to combine distance measure (34) with the Euclidean distance used for the fuzzy c-means algorithm. The resulting objective function is shown in (41). Parameter $0 < q < 1$ determines the influence of each summand in the distance function (42).

$$\begin{aligned} J_{SAVCB}^{prob}(X,U,v) = \\ \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \cdot \Big(q \cdot \max\{0,\ (x_k - v_i)^T (x_k - v_i) - \tau_i\} \\ + (1-q) \cdot (x_k - v_i)^T (x_k - v_i) \Big) \end{aligned} \tag{41}$$

$$\begin{aligned} \mathfrak{D}_{SAVCB} = d^2(v_i,\ x_k) = q \cdot \max\{0,\ (x_k - v_i)^T (x_k - v_i) - \tau_i\} \\ + (1-q) \cdot (x_k - v_i)^T (x_k - v_i) \end{aligned} \tag{42}$$

To adapt the cluster radii during the alternating optimization, again constraint (24) has to be considered, leading to equation (43).

$$
\begin{aligned}
&J^{prob}_{SAVCB,\,\lambda}(X,U,v) = \\
&\quad \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \cdot \Big(q \cdot \ \max\{0,\ (x_k - v_i)^T (x_k - v_i) - \tau_i\} \\
&\quad + (1-q) \cdot \ (x_k - v_i)^T (x_k - v_i) \Big) + \lambda \cdot \ \left(\sum_{i=1}^{c} \tau_i^2 \ - \ \tau \right)
\end{aligned}
\tag{43}
$$

The alternating optimization scheme needs calculation instructions for the cluster centers v_i, the radii τ_i and the membership degrees u_{ik}. Equations (44) and (45) are the corresponding partial derivatives of our objective function. The partial derivative for λ remains as denoted in section 3.2 (39)

$$
\frac{\partial J^{prob}_{SAVCB,\,\lambda}}{\partial \lambda} = \sum_{i=1}^{c} \tau_i^2 \ - \ \tau \stackrel{!}{=} 0
$$

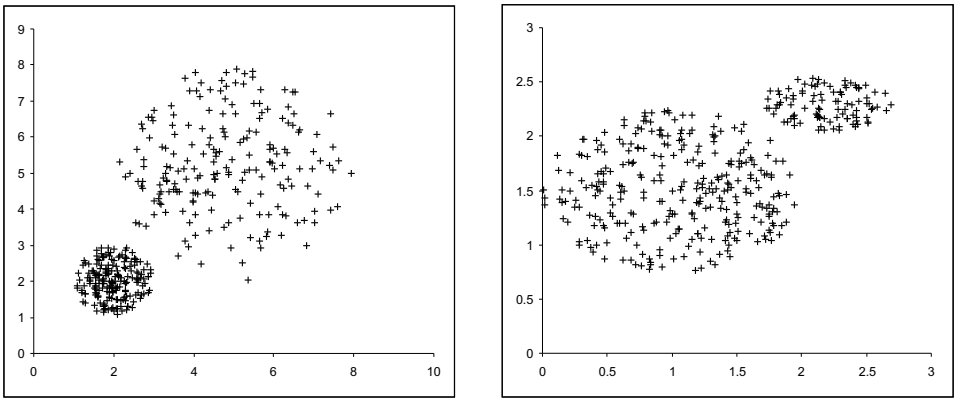
$$
\begin{aligned}
\frac{\partial J^{prob}_{SAVCB,\,\lambda}}{\partial v_i} &= 2 \cdot \ (1-q) \cdot \sum_{k=1}^{n} u_{ik}^{m} \cdot \ (v_i - x_k) \\
&+ \ 2 \cdot \ q \cdot \sum_{k:(x_k - v_i)^T (x_k - v_i) > \tau_i} u_{ik}^{m} \cdot \ (v_i - x_k) \stackrel{!}{=} 0
\end{aligned}
\tag{44}
$$

$$
\frac{\partial J^{prob}_{SAVCB,\,\lambda}}{\partial \tau_i} = - \ q \cdot \sum_{k:(x_k - v_i)^T (x_k - v_i) > \tau_i} u_{ik}^{m} + 2 \cdot \ \lambda \cdot \ \tau_i \stackrel{!}{=} 0 \tag{45}
$$

The resulting calculation instructions are shown in (46) and (47), respectively. The greater the influence of the Euclidean distance ($q \to 0$), the smaller are the calculated center radii τ_i.

$$
\stackrel{(44)}{\Rightarrow} \ v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} \cdot \ x_k \ - \ q \cdot \ \sum_{k:(x_k - v_i)^T \cdot (x_k - v_i) \le \tau_i} u_{ik}^{m} \cdot \ x_k}{\sum_{k=1}^{n} u_{ik}^{m} \ - \ q \cdot \ \sum_{k:(x_k - v_i)^T \cdot (x_k - v_i) \le \tau_i} u_{ik}^{m}} \tag{46}
$$

$$
\stackrel{(45)}{\Rightarrow} \ \tau_i = \frac{\sum_{k:(x_k - v_i)^T (x_k - v_i) > \tau_i} u_{ik}^{m}}{\sqrt{\sum_{i=1}^{c} \left(\sum_{k:(x_k - v_i)^T (x_k - v_i) > \tau_i} u_{ik}^{m} \right)^2}} \cdot \ \sqrt{\tau} \cdot \ q \tag{47}
$$



(a) Two spherical clusters (b) Two ellipsoidal clusters

Figure 1: Artificial test data sets

Depending on the chosen basic clustering technique (probabilistic, possibilistic, or noise) the adequate calculation instruction for the membership degrees u_{ik} has to be chosen. Therein the distance measure has to be replaced by $\mathfrak{D}_{SAVCB}$ (42). Even if the influence of the Euclidean distance is rather small ($q \approx 0.99$), the alternating optimization converges reliable in our experiments.

We refer to the corresponding alternating optimization scheme as the FCM-volume algorithm.

4 Examples

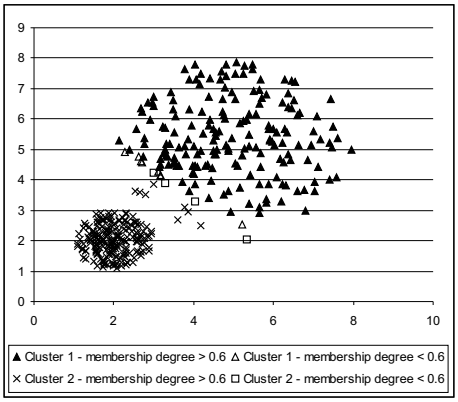
To demonstrate the properties of our new approaches we have designed two artificial test data sets as can be seen in figure 1. Part (a) of figure 1 shows two spherical clusters with uniformly distributed data points for both clusters but different radii. Here the number of data points for each cluster is the same. In part (b) of figure 1 two ellipsoidal clusters with equally distributed data points but different extents are displayed. The larger cluster has about twice as much data points as the smaller one.

In figure 2 the results for the data set from figure 1(a) with the algorithms

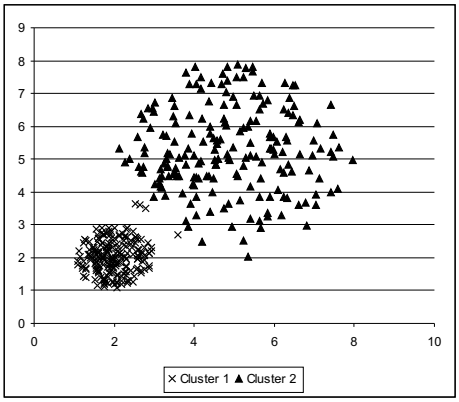
using the Euclidean distance measure are compared. The fuzzifier m was in all cases set to 2.0. The constraint parameter τ was set to 1.0 in both cases, FCM-sized and FCM-volume. For the size adaptable version of the fuzzy c-means algorithm the exponent l was set to 0.5. For the influence of the radii part in case of the FCM-volume approach, $q = 0.99$ was chosen. The original fuzzy c-means algorithm has difficulties in assigning the data to the correct clusters (see figure 2 (a)). A datum is assigned to the cluster with the highest membership degree. The approach using the Euclidean distance combined with volume centers is in a position to adapt the volume centers and therefore yields slightly better results than the original FCM (see part (c) of figure 2). Only the size adaptable approach (part (b)) has the ability to assign most data points correctly.

In figure 3 the results for the ellipsoidal test data set from figure 1(b) are shown. As clustering algorithms the axes-parallel versions of the Gustafson-Kessel algorithm and our new size-adaptable version of that algorithm have been chosen. The fuzzifier m was assigned the value 2.0 in both cases. For the size-adaptable approach constraint parameter τ has been set to 1 and the exponent l was assigned 0.4. Our new approach is able to adapt to the ellipses' content (figure 3 part (b)) whereas the result in part (a) of figure 3 shows that the Gustafson-Kessel algorithm searches for groups of about the same size. Our approach can be further improved if a smaller value for parameter l, e.g. $l = 0.3$, is chosen.

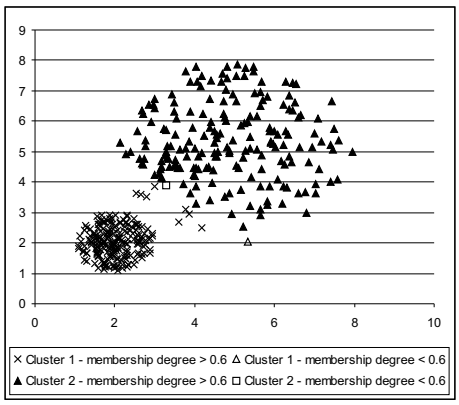
As another example, we used the Wisconsin Breast Cancer Database [22, 2] to test our new approaches with the probabilistic objective function. This classified data set originally contains 699 data points with 9 attributes and a classification attribute. 16 data points with missing values have been deleted from the data set for our tests. From the remaining 683 data points 444 were classified as benign and 239 as malignant. In Figure 4 the results for the original fuzzy c-means algorithm (FCM) are compared to our size adaptable version of this algorithm (FCM-sized) and the combination of the FCM with the volume-center-based approach (FCM-volume). In figure 4 the percentage of wrong classified data for two to ten clusters is displayed. The fuzzifier m was in all cases set to 2.0. The values for the other parameters



(a) FCM

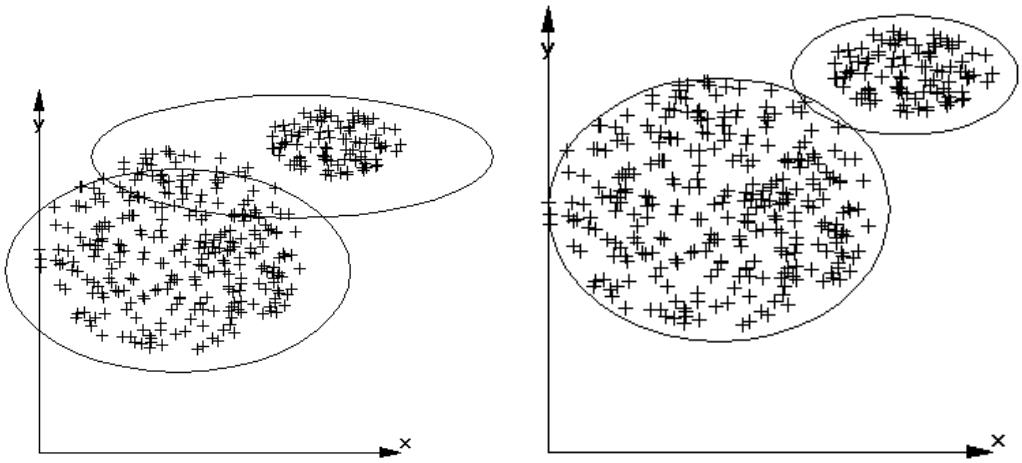


(b) FCM-sized



(c) FCM-volume

Figure 2: Classification results for a circular test data set



(a) GK-parallel (b) GK-parallel-sized

Figure 3: Classification results for an ellipsoidal test data set

are $\tau = 1.0$ for FCM-sized as well as for FCM-volume, $l = 0.8$ for FCM-sized and $q = 0.9$ for FCM-volume. In this case the priority of the FCM-volume lies upon the volume-based component. The best results are obtained with our new algorithms. The FCM-sized algorithm yields the best classification where 2.6% of the data entries are misclassified with four clusters. A similar good result (2.8% misclassified data points) is reached in case of the FCM-volume at 5 clusters. The best result for the FCM (3.1% wrong classified data) is obtained with 4 clusters. Our approaches seem both to improve the results for the Wisconsin breast cancer database.

As can be seen in Figure 5 where the results for the axes-parallel version of the method designed by Gustafson and Kessel (GK parallel, see section 2.2.2) are compared to our size adaptable version of this algorithm (GK parallel sized, see section 3.1). As mentioned, the percentage of wrong classified data for two to twelve clusters is shown. The fuzzifier m was again set to 2.0. The values for the other parameters are $\tau = 1$ and $l = 0.5$ for the GK-parallel-sized. That both algorithms show similar worse results indicates that the model of ellipsoidal clusters is not suited for this data set and that the GK and it's variantsalthough they are extensions of the FCM variants are not

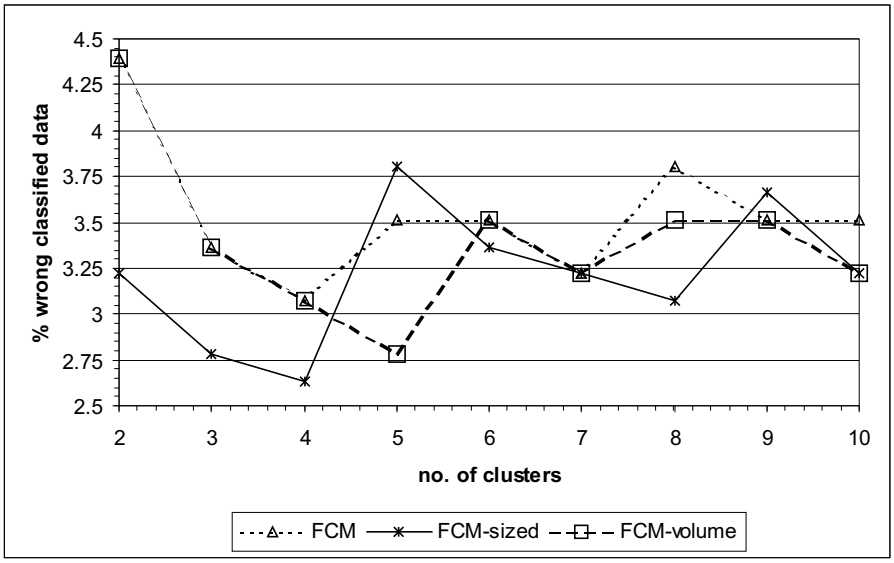


Figure 4: Classification results for Wisconsin Breast Cancer - Part A

able to find the spherical clustersof the good results found by the FCM and it's variants. Not shown are the results obtained by the non-restricted version of the GK and our new size adaptable approach for this algorithm. They are worse than those of the restricted versions GK parallel and GK parallel sized. Nevertheless, if the principal chosen model is well suited for the considered data set, size adaptation has the ability to improve the classification results.

In table 1 clustering results in case of the size-adaptable fuzzy c-means for different values of parameter l are shown. The values denote the percentage of misclassified data. To calculate this value first for all clusters c the class which is represented by one particular cluster is determined. Then the data points corresponding to that cluster but originally belonging to a different class than the cluster's are counted. The sum of misclassified data over all clusters in ratio to the total number of data gives the percentage of misclassified data, also called error rate. It can be seen, that the result depends on the choice of the exponent l. For figure 4 the value for $l = 0.8$ obtaining best results has been chosen.

For table 2 we have chosen those c-partitions for each l-value from table 1, where the least error values occurred, and calculated the maximal mem-

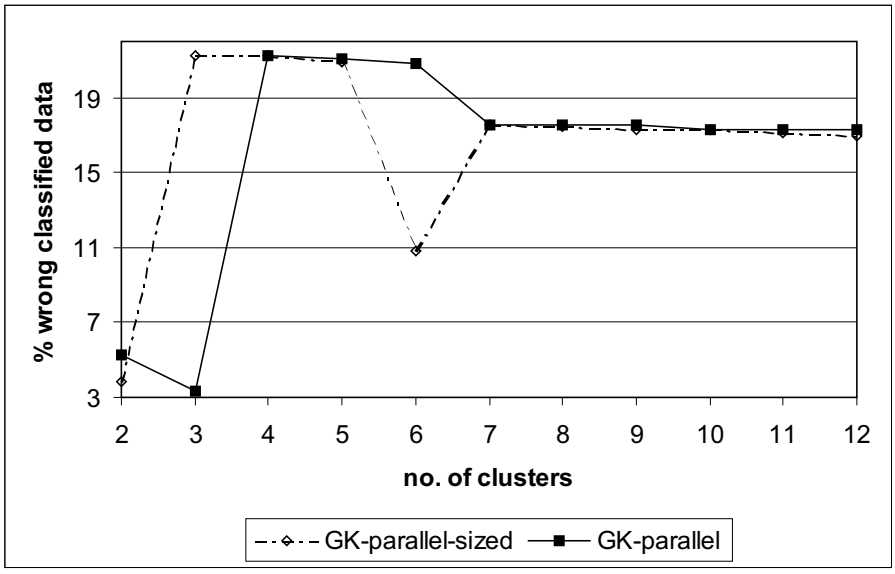


Figure 5: Classification results for Wisconsin Breast Cancer - Part B

Table 1: Influence of parameter l in FCM-sized - Part A

c	**l = 0.2**	**l = 0.5**	**l = 0.8**	**l = 1.0**	**l = 3.0**	**l = 5.0**
2	25.5	*2.8*	3.2	3.1	4.0	4.2
3	30.9	*2.8*	2.8	2.9	3.2	3.2
4	36.6	2.9	*2.6*	*2.6*	*2.9*	*3.1*
5	47.4	4.4	3.8	3.8	3.7	3.5
6	34.8	2.9	3.4	3.2	3.5	3.5
7	13.6	2.9	3.2	3.4	3.2	3.2
8	17.4	*2.8*	3.1	3.2	3.8	3.8
9	22.3	4.2	3.7	3.5	3.5	3.5
10	34.6	3.4	3.2	3.2	3.2	3.4
11	*12.2*	2.9	3.8	3.8	3.8	3.8
12	13.3	3.1	3.5	3.5	3.5	3.5

Table 2: Influence of parameter l in FCM-sized - Part B

$\mathbf{l = 0.2}$	0.5964
$\mathbf{l = 0.5}$	0.9078
$\mathbf{l = 0.8}$	0.7045
$\mathbf{l = 1.0}$	0.7105
$\mathbf{l = 3.0}$	0.7112
$\mathbf{l = 5.0}$	0.7119

bership degree for each datum separately. In table 2 the average of these maximal membership degrees is shown for each partition. It is obvious that this value in probabilistic clustering depends on the total number of clusters for the partition, e.g. is the value for $l = 0.2$ and $c = 11$ clusters less than the result for $l = 0.5$ and $c = 2$ clusters. The last four entries illustrate the influence of parameter l on the membership degrees, see section 3.1. Here, for the l-values $0.8, 1.0, 3.0$ and 5.0 the number of clusters was in all cases $c = 4$. The calculated values are slightly increasing for increasing l-values.

5 Conclusions

Our approach seems to be well suited to adapt to different sizes of clusters. One remaining problem that also exists concerning the original versions of the algorithms presented in sections 2 and 4 is that all these approaches presuppose uniformly distributed data over all clusters, i.e. the number of data points per cluster are assumed to be equal for all clusters. To cluster data with varying sizes and numerical differences regarding the data points per structure correctly, adaptation to the density has to be taken into account.

Especially in applying fuzzy clustering techniques to the task of rule learning it is not necessary to implement highly form-adaptable algorithms since more flexible cluster algorithms referring to the cluster shape generally result in a higher loss of information by rule generation. Several approaches to apply fuzzy clustering algorithms to the task of rule learning have been

developed in recent years, see for instance [11, 13, 14, 20, 23]. However, a loss of information by the process of rule generation is unavoidable.

As long as we stay with such simple clustering algorithms as FCM or the parallel version of GK, loss of information in case of rule learning can be mostly avoided. This is also valid for the size adaptable versions of these algorithms since the form describing distance measure is not changed. In case of rule learning the proposed modified versions of the described algorithms are a good alternative to more complex algorithms like the method introduced by Gath and Geva, see section 2.2.3.

References

[1] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[2] C. Blake, E. Keogh, and C. Merz. Uci repository of machine learning databases., 1998.

[3] R. Davé. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12:657–664, 1991.

[4] R. Davé and R. Krishnapuram. Robust clustering methods: A unified view. *IEEE Transactions on Fuzzy Sets and Systems*, 5(2):270–293, 1997.

[5] I. Gath and A. Geva. Unsupervised optimal fuzzy clustering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11:773–781, 1989.

[6] H. Genther and M. Glesner. Automatic generation of a fuzzy classification system using fuzzy clustering mehtods. In *ACM Symposium on Applied Computing (SAC'94)*, pages 180–183, Phoenix, 1994.

[7] D. Gustafson and W. Kessel. Fuzzy clustering with a fuzzy covariance matrix. In *IEEE CDC*, pages 761–766, San Diego, 1979.

[8] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis*. Wiley, Chichester, 1999.

[9] A. Keller and F. Klawonn. Clustering with volume adaptation for rule learning. In *EUFIT'99*, Aachen, 1999.

[10] F. Klawonn and A. Keller. Fuzzy clustering and fuzzy rules. In *7th Intern. Fuzzy Systems Association World Congress (IFSA'97)*, volume I, pages 193–198, Prague, 1997. Academia.

[11] F. Klawonn and R. Kruse. Automatic generation of fuzzy controllers by fuzzy clustering. In *1995 IEEE Intern. Conference on Systems, Man, and Cybernetics*, pages 2040–2045, Vancouver, 1995.

[12] F. Klawonn and R. Kruse. Clustering methods in fuzzy control. In W. Gaul and D. Pfeifer, editors, *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organization.*, pages 195–202. Springer-Verlag, Berlin, 1995.

[13] F. Klawonn and R. Kruse. Derivation of fuzzy classification rules from multidimensional data. In G. Lasker and X. Liu, editors, *Advances in Intelligent Data Analxsis*, pages 90–94. The International Institute for Advanced Studies in Systems Research and Cybernetics, Windsor, Ontario, 1995.

[14] F. Klawonn and R. Kruse. Constructing a fuzzy controller from data. *Fuzzy Sets and Systems*, 85:177–193, 1997.

[15] R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE Trans. on Fuzzy Systems*, 1:98–110, 1993.

[16] Y. Ohashi. Fuzzy clustering and robust estimation. In *9th Meeting SAS Users Group Int.*, Hollywood Beach, Fl., 1984.

[17] T. Runkler and J. C. Bezdek. Alternating cluster estimation: A new tool for clustering and function approximation. *IEEE Transactions on Fuzzy Systems*, 7(4):377–393, 1999.

[18] S. Z. Selim and M. A. Ismail. Soft clustering of multidimensional data: A semi-fuzzy approach. *Pattern Recognition*, 17(5):559–568, 1984.

[19] M. Setnes and U. Kaymak. Extended fuzzy c-means with volume prototypes and cluster merging. In *EUFIT'98*, pages 1360–1364, Aachen, 1998.

[20] M. Sugeno and T. Yasukawa. A fuzzy-logic-based approach to qualitative modelling. *IEEE Trans. on Fuzzy Systems*, 1:7–31, 1993.

[21] E. Trauwaert, L. Kaufmann, and P. Rousseeuw. Fuzzy clustering algorithms based on the maximum likelihood principle. *Fuzzy Sets and Systems*, 42:213–227, 1991.

[22] W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *National Academy of Sciences*, volume 87, pages 9193–9196, USA, 1990.

[23] Y. Yoshinari, W. Pedrycz, and K. Hirota. Construction of fuzzy models through fuzzy clustering techniques. *Fuzzy Sets and Systems*, 54:157–165, 1993.