Clustering with Volume Adaptation for Rule Learning

Annette Keller[1] Frank Klawonn[2]

[1]Institute for Flight Guidance
German Aerospace Center
Lilienthalplatz 7
D-38108 Braunschweig, Germany
e-mail annette.keller@dlr.de

[2]Department of Electrical Engineering and Computer Science
Ostfriesland University of Applied Sciences
Constantiaplatz 4
D-26723 Emden, Germany
e-mail klawonn@et-inf.fho-emden.de

Abstract

The well-known fuzzy c-means algorithm – an objective function based fuzzy clustering technique – is often used for rule learning. This method is suitable to detect equally sized (hyper-)spheres in the domain of interest. In replacing the Euclidean distance measure in the objective function, other cluster shapes, for instance ellipsoids or lines, can be found. The results of known techniques to adapt to the cluster sizes highly depend on the initialisation of the clustering procedure and often need increased computation time in comparison to techniques as the fuzzy c-means. We propose a modified objective function that enables us to retain the possibilities of well-known clustering algorithms and enabling these algorithms to adapt to the cluster sizes.

1 Introduction

Standard fuzzy clustering methods like the fuzzy c-means algorithm are based on the idea to optimise an objective function. This objective function depends on the distances of the data to the cluster centres 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

have 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 adaptable the clustering algorithms are in general, the more they depend on a suitable cluster initialisation. Also with the flexibility of cluster structures the complexity of the proposed algorithms highly increases. Several approaches to apply fuzzy clustering algorithms to the task of rule learning have been developed in recent years, see for instance [7, 9, 10, 11, 13]. However, a loss of information by the process of rule generation is unavoidable. More flexible cluster algorithms referring to the cluster shape generally result in a higher loss of information by rule generation. In this paper we briefly review the necessary background on objective function based fuzzy clustering and rule learning in 2. Based on the clustering approaches described in 2 we propose a modification of the objective function in 3. This enables well-known clustering algorithms to adapt to the shape sizes without increasing the loss of information by rule generation. Finally some examples for the proposed objective function are shown in 4.

2 Objective Function-Based Fuzzy Clustering and Rule Learning

We cannot give a complete overview on objective function based fuzzy clustering and mention here only the background necessary to understand our new approach. For an overview on fuzzy clustering see for example [5]. Most fuzzy clustering algorithms aim at minimising the objective function

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

under the constraints

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

and

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

$X = \{x_1, \ldots, x_n\} \in \mathbb{R}^p$ is the data set, c is 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 the 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. (2) guarantees that no cluster is

empty, (3) enforces that for each datum its classification can be distributed over different clusters, but the sum of the membership degrees to all clusters has to be one for each datum. Differentiating (1) considering the constraints making use of Lagrange multipliers 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 optimisation scheme starts with a random initialisation and applies the equations for the u_{ik} and the prototypes until the difference between the matrices $\left(u_{ik}^{old}\right)$ and (u_{ik}^{new}) in two succeeding iterations is smaller than a given bound ε. The most simple fuzzy clustering algorithm is the fuzzy c-means (FCM) (see e.g. [1]) where the distance $d(v_i, x_k)$ is simply the Euclidean distance and the prototypes are vectors $v_i \in \mathbb{R}^p$. It searches for spherical clusters of approximately the same size and by differentiating (1) we obtain the necessary conditions

$$v_i \;=\; \frac{\sum_{k=1}^{n} u_{ik}^m \cdot x_k}{\sum_{k=1}^{n} u_{ik}^m} \tag{5}$$

for the prototypes that are used alternatingly with (4) in the iteration procedure.

Gustafson and Kessel [4] designed a fuzzy clustering method that can adapt to hyper-ellipsoidal forms. The prototypes consist of the cluster centres 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

$$d^2(v_i, x_k) \;=\; (\rho_i \det C_i)^{1/p} \cdot (x_k - v_i)^\top C^{-1} (x_k - v_i).$$

Another clustering technique (GG) that was designed by Gath and Geva [2] is in some way able to adapt the cluster size but is no longer based on an objective function approach. Instead the GG is a heuristic method derived from the fuzzification of a maximum likelihood estimator. In [12] other clustering methods based on the maximum likelihood principle are described.

These or similar algorithms can be applied to learn fuzzy rules from data for classification problems [3, 9] or function approximation [8, 10, 11]. 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 [6]. 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 the 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.

3 Clustering with Adaptation of Cluster Sizes

For each cluster we add an additional parameter r_i to the objective function in order to enable the clustering algorithm to adapt the cluster sizes. r_i can be considered as the (relative) radius of the corresponding cluster. The resulting objective function is shown in (6), with constant real-valued parameter $l > 0$.

$$J(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \cdot \frac{1}{r_i^l} \cdot d^2(x_k, v_i) \tag{6}$$

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

$$\sum_{i=1}^{c} r_i = r \tag{7}$$

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

Since the objective function (6) does not require special properties of the distance measure d, most of the described clustering procedures need only little modification to use the advantages of the proposed objective function. Let us define

$$d_r^2(x_k, v_i, r_i) = \frac{1}{r_i^l} \cdot d^2(x_k, v_i) \tag{8}$$

as a new group of distance measures. Then the objective function (6) can be rewritten as

$$J(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \cdot d_r^2(x_k, v_i, r_i). \tag{9}$$

For this modified objective function (8) with the constraints (2) and (3) we obtain the same equations for the membership degrees as in (4), except that we have to replace the old distance $d^2(v_i, x_k)$ by $d_r^2(v_i, x_k)$, i.e.

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_r^2(x_k, v_i, r_i)}{d_r^2(x_k, v_j, r_j)} \right)^{\frac{1}{m-1}}}. \tag{10}$$

The modified objective function for the FCM is shown in (11).

$$J(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \cdot \frac{1}{r_i^l} \cdot (x_k - v_i)^T (x_k - v_i) \tag{11}$$

Analogous to the objective function from section (2) minimising (6) leads to the necessary condition

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m \cdot x_k}{\sum_{k=1}^{n} u_{ik}^m} \tag{12}$$

for the evaluation of the prototype coordinates.

Assuming that the parameters $l > 0$ and $r > 0$ are fixed, we have to take the constraint (7) into account, when we have to determine the values r_i by predefined and during the iteration procedure unchanged $l > 0$ and $r > 0$, the constraint (7) has to be taken into account. With condition (7) we obtain the Lagrange function

$$J_\lambda(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \cdot \frac{1}{r_i^l} \cdot d^2(x_k, v_i) + \lambda \left(\sum_{i=1}^{c} r_i - r \right). \tag{13}$$

Since the distance measure is independent of r_i, differentiating (13) gives us

$$\frac{\partial J_\lambda(X, U, v)}{\partial r_i} = -\frac{l}{r_i^{l+1}} \cdot \sum_{k=1}^{n} u_{ik}^m \cdot d^2(x_k, v_i) + \lambda \stackrel{!}{=} 0 \tag{14}$$

and therefore

$$r_i = \left(\frac{l \cdot \sum_{k=1}^{n} u_{ik}^m \cdot d^2(x_k, v_i)}{\lambda} \right)^{\frac{1}{l+1}}. \tag{15}$$

With (7) λ evaluates to

$$\lambda = \frac{\left(\sum_{j=1}^{c} \left(l \cdot \sum_{k=1}^{n} u_{jk}^m \cdot d^2(x_k, v_j) \right)^{\frac{1}{l+1}} \right)^{l+1}}{r^{l+1}}. \tag{16}$$

Equation (17) represents the resulting calculation instruction for the r_i.

$$r_i = \frac{\left(\sum_{k=1}^{n} u_{ik}^m \cdot d^2(x_k, v_i) \right)^{\frac{1}{l+1}}}{\sum_{j=1}^{c} \left(\sum_{k=1}^{n} u_{jk}^m \cdot d^2(x_k, v_j) \right)^{\frac{1}{l+1}}} \cdot r \tag{17}$$

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 bad effect on algorithms like 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 (17) can be used alternatingly with equations (10) and (12) and a suitable distance measure for fuzzy clustering algorithms. 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. In the next section we present some examples for the modified versions of the FCM and GK respectively.

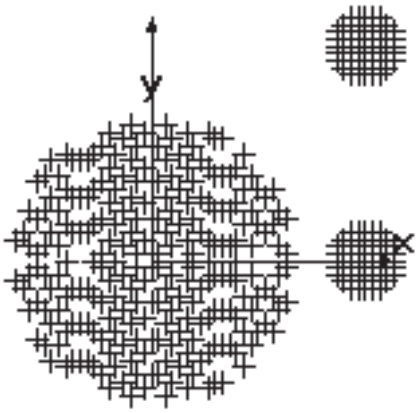

Figure 1: Three circular groups of different size

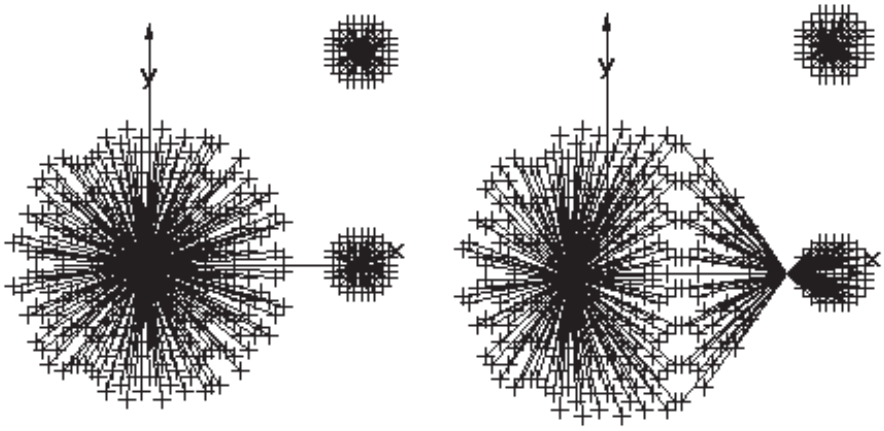

Figure 2: Result with size adapting FCM (left) and normal FCM (right)

4 Examples

Figure 1 shows one of our test data sets for the modified versions of the FCM and GK with three clusters. For all of the applied cluster algorithms the number of clusters was set to 3. For the fuzzifier we have chosen $m = 2$ in all cases.

Since the three clusters differ extremely in their size, the original FCM (right side of figure 2) has no chance to detect the small lower right cluster correctly. Our modified version (left hand side) has no difficulties to find the correct classification. Here the parameter r from the necessary condition (7) was set to 1 and the exponent of the parameter r_i, l, was assigned the value 0.3.

Figure 3 shows the clustering results for the original GK (at right) and the modified version (at left). These algorithms have more difficulties in finding the correct classification. They are assuming ellipsoidal cluster structures of the same size and depend more on a suitable initialisation of the prototype coordinates.

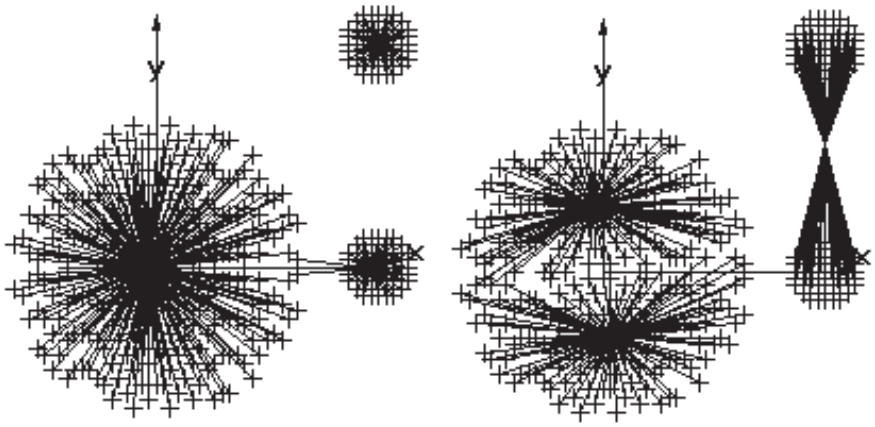

Figure 3: Result with size adapting GK (left) and normal GK (right

Nevertheless the modified version of the GK determines the three clusters nearly correct.

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 equally distributed data over all clusters, i.e. the number of datapoints 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. As long as we stay with such simple clustering algorithms like the FCM or the parallel version of the GK, presented e.g. in [5], loss of information in case of rule learning can be 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 [2].

References

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

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

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

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

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