Constructing a Fuzzy Controller from Data

Frank Klawonn, Rudolf Kruse
Department of Computer Science
University of Braunschweig
D–38106 Braunschweig, Germany

Abstract

Fuzzy control can be interpreted as an approximation technique for a control function based on typical, imprecisely specified input–output tuples that are represented by fuzzy sets. The imprecision is characterized by similarity relations that are induced by transformations of the canonical distance function between real numbers. Taking this interpretation of fuzzy controllers into account, in order to derive a fuzzy controller from observed data typical input–output tuples have to be identified. In addition, a concept of similarity based on a transformations of the canonical distance is needed in order to characterize the typical input–output tuples by suitable fuzzy sets.

A variety of fuzzy clustering algorithms exists that are exactly working in this spirit: They identify prototypes and assign fuzzy sets to the prototypes on the basis of a suitable transformed distance. In this paper we discuss how such fuzzy clustering techniques can be applied to construct a fuzzy controller from data and introduce special clustering algorithms that are tailored for this problem.

Keywords: similarity relation; fuzzy cluster analysis; fuzzy control

1 Introduction

The principal idea behind fuzzy control is to define a control function on the basis of linguistic control rules that describe an adequate control strategy. These linguistic rules can for example be formulated by an operator who is able to control the plant. In this case one encounters well known problems of knowledge acquisition like the difficulties that experts have with specifying their complete, but not always consciously applied knowledge.

Instead of asking the operator directly, one can observe him instead and try to derive rules from these observation data and use these rules as a support for the knowledge acquisition process. More generally, the problem can be stated as follows. Given a set of data for which we presume some functional dependency, is there a suitable methodology to derive (linguistic) rules from these data that characterize the unknown function at least vaguely?

In this paper we briefly discuss fuzzy control as an approximation technique based on typical input–output tuples and a concept of indistinguishability that is induced by transformations of the canonical distance between real numbers. This discussion shows that certain fuzzy clustering algorithms seem to be very suitable for deriving (linguistic) rules from data since they are also based on the idea of typical elements and indistinguishability induced by transformed distances. We introduce modifications of fuzzy clustering techniques that are well suited for deriving rules from given data

The aim of our proposed methods for constructing a fuzzy controller from data is not to find a kind of best approximation of the data. For such a task other mathematical approximation techniques depending on what we understand by a best approximation are better suited. Our intention is to extract knowledge from the data that can be easily understood and interpreted, which is often not possible when the functional dependency inherent in the data is approximated by some complex function.

2 Fuzzy Control and Similarity

In this section we briefly review the interpretation of fuzzy controllers in the framework of similarity relations. The principal idea is that each rule in the rule base of a fuzzy controller describes vaguely a crisp point of the control surface and the fuzzy sets appearing in the fuzzy partitions of the input and output spaces stand for vaguely defined crisp values.

We will not dive into a discussion on the variety of possibilities for representing a concept like similarity or indistinguishability. For our purpose it is sufficient to restrict ourselves to the following definition of a similarity relation.

Definition 2.1 *A similarity relation on a set X is a mapping $S : X \times X \to [0,1]$ fulfilling the axioms*

(S1) $S(x,x) = 1$

(S2) $S(x,y) = S(y,x)$

(S3) $S(x,y) + S(y,z) - 1 \ \leq\ S(x,z)$.

The value $S(x,y)$ should be interpreted as the degree of similarity or indistinguishability of the elements x and y. The reflexivity condition (S1) requires that each element is similar to itself to the degree 1. (S2) guarantees for symmetry, i.e. that x is similar to y to the same degree as y is similar to x. (S3) can be interpreted as a transitivity condition since it is equivalent to the axiom

$$S(x,y) * S(y.z) \ \leq\ S(x,z) \tag{1}$$

where $*$ stands for the Lukasiewicz t–norm which is defined by $\alpha * \beta = \max\{\alpha + \beta - 1, 0\}$. Reading the Lukasiewicz t–norm as the valuation function of a conjunction, (1) stands for the statement that if x and y as well as y and z are similar, then also x and z have to be similar. In principal $*$ could be replaced an arbitrary t–norm effecting of course the axiom (S3) and leading to other definitions of similarity relations [14, 12, 5].

We have chosen the transitivity condition (S3) with respect to the Lukasiewicz t–norm for the reason that in this case we have a close connection between similarity relations and metrics. This means precisely that similarity relations and pseudo–metrics bounded by one are dual concepts, since a similarity relation S induces a pseudo–metric $\varrho_S(x, y) = 1 - S(x, y)$ and vice versa a pseudo–metric δ bounded by one defines a similarity relation $R_\delta(x, y) = 1 - \delta(x, y)$ and we have the one–to–one correspondence $S = R_{\varrho_S}$ and $\delta = \varrho_{R_\delta}$. Thus on the real numbers we obtain the canonical similarity relation $S(x, y) \;=\; 1 - \min\{|x - y|, 1\}$ induced by the standard metric $\delta(x, y) \;=\; |x - y|$.

Although this canonical similarity relation on the real numbers looks very natural, it is too restrictive to be suitable for most of the applications. At least, we have to take into account the unit of measurement. Otherwise it would be possible that two measurements might have different degrees of similarity depending on the unit in which they were measured. For instance, when using Fahrenheit for the temperature the absolute value of the difference between two temperatures is greater than their difference taken in Celsius. In order to avoid such anomalies it is necessary to introduce scaling factors. A scaling factor $c \geq 0$ defines a transformation $t_c : \mathbb{R} \to \mathbb{R}$ and instead of using the distance between two values for the definition of their degree of similarity the distance of their transformed values is used, i.e.

$$S(x, y) \;=\; |t_c(x) - t_c(y)| \;=\; |c \cdot x - c \cdot y|. \tag{2}$$

In [6] the concept of scaling factors is generalized first to the case of defining different scaling factors for different regions or intervals and then more generally to assigning to each real value x an individual scaling factor $c(x)$, which specifies how strong one has to distinguish between values in the neighbourhood of x. In this way one obtains a more general form of (2)

$$S(x, y) \;=\; |t_c(x) - t_c(y)| \;=\; \left| \int_x^y c(s)ds \right|. \tag{3}$$

There is a close relation between fuzzy sets and similarity relations in the sense that fuzzy sets are induced by crisp sets when taking a similarity relation into account and vice versa, from given fuzzy sets a similarity relation and crisp sets can be derived under certain assumptions so that the fuzzy sets are induced by the crisp sets.

Let us first recall how a fuzzy set can be interpreted as the representation of a crisp set in the presence of a similarity relation. Let S be a similarity relation on the set X and let $x_0 \in X$. The element x_0 induces the (fuzzy) set μ_{x_0} of all elements that are similar to x_0. The membership degree of an element x to this fuzzy set is simply the grade to which x and x_0 are similar, i.e.

$$\mu_{x_0}(x) \;=\; S(x, x_0).$$

Definition 2.2 *The fuzzy set* $\mu_{x_0}(x) = S(x, x_0)$ *is called the extensional hull of* x_0 *(with respect to the similarity relation* S*).*

An interesting observation is that when $X = \mathbb{R}$ and the similarity relation S is induced by the standard metric, i.e. $S(x, y) = 1 - \min\{|x - y|, 1\}$, then the extensional hull μ_{x_0} of the value $x_0 \in \mathbb{R}$ is a symmetrical triangular membership function, having its maximum at x_0. When a scaling function c is admitted so that the similarity relation is of the form (3), not only triangular membership functions may appear as extensional hulls, but any fuzzy set μ satisfying the following axioms. There exists $x_0 \in \mathbb{R}$ such that

(C1) $\mu(x_0) = 1$,

(C2) μ is a non–decreasing function on $(-\infty, x_0]$,

(C3) μ is a non–increasing function on $[x_0, \infty)$,

(C4) μ is continuous,

(C5) μ is almost everywhere differentiable.

A natural question that arises is whether a given family of fuzzy sets, for instance a fuzzy partition like it is used in fuzzy control, can be interpreted as the extensional hulls of crisp values with respect to a suitable similarity relation. In [6] the following theorem was proved, providing the answer to this question when it is required that the corresponding similarity relation is defined on the basis of a scaling function.

Theorem 2.3 *Let* $(\mu_i)_{i \in I}$ *be an at most countable family of fuzzy sets on* $\mathbb{R}$ *and let* $(x_0^{(i)})_{i \in I}$ *be a family of real numbers such that for each* $i \in I$ *the fuzzy set* μ_i *and the value* $x_0^{(i)}$ *satisfy the axioms (C1)–(C5). There exists a scaling function* $c : \mathbb{R} \to [0, \infty)$ *such that* μ_i *coincides with the extensional hull of* $x_0^{(i)}$ *(for each* $i \in I$*) with respect to the similarity relation induced by* c*, if and only if*

$$\min\{\mu_i(x), \mu_j(x)\} > 0 \quad \Rightarrow \quad \left|\frac{d\mu_i(x)}{dx}\right| = \left|\frac{d\mu_j(x)}{dx}\right| \tag{4}$$

holds almost everywhere for all $i, j \in I$.

In case condition (4) holds, the corresponding similarity relation is given by (2) with the scaling function

$$c : \mathbb{R} \to [0, \infty), \qquad x \mapsto \begin{cases} \left|\frac{d\mu_i(x)}{dx}\right| & \text{if there exists } i \in I \text{ s.t. } \mu_i(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that the typical requirement for fuzzy partitions used in fuzzy control that the membership degrees of two neighbouring fuzzy sets add up to one implies (4). Thus in many examples of fuzzy control applications corresponding scaling functions inducing suitable similarity relations can be derived from the fuzzy partitions.

The aim of this section is to understand fuzzy control from the viewpoint of similarity relations. As it is already shown in [7, 10], fuzzy control can be interpreted as interpolation in the presence of indistinguishability expressed by similarity relations.

Since fuzzy control is concerned with the construction of a control function that is defined for a number of inputs we have to consider similarity relations in multidimensional spaces. Let $S_1, \ldots, S_n$ be similarity relations on $X_1, \ldots, X_n$, respectively. We are looking for a suitable similarity relation on the product space $X_1 \times \ldots \times X_n$ as a combination of the similarity relations $S_1, \ldots, S_n$. There is of course more than one possibility to define such a similarity relation. However, the similarity relation

$$S_{\min\{S_1,\ldots,S_n\}}\big((x_1, \ldots, x_n), (\tilde{x}_1, \ldots, \tilde{x}_n)\big) \;=\; \min\big\{S_i(x_i, \tilde{x}_i) \mid i \in \{1, \ldots, n\}\big\}$$

is distinguished from others since it is the greatest similarity relation S such that

$$S\big((x_1, \ldots, x_n), (x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_n)\big) \;\geq\; S_i(x_i, \tilde{x}_i)$$

as well as

$$S\big((x_1, \ldots, x_n), (\tilde{x}_1, \ldots, \tilde{x}_n)\big) \;\leq S_i(x_i, \tilde{x}_i)$$

hold for all $i \in \{1, \ldots, n\}$. Therefore, $S_{\min\{S_1,\ldots,S_n\}}$ is the coarsest similarity relation on the product space that does distinguish at least as well as each S_i.

Let us consider a single fuzzy control rule of the form

If ξ_1 is A_1 and $\ldots$ ξ_n is A_n then η is B

where ξ_i denotes the i^{th} input variable, η is the output variable, and A_i and B represent linguistic terms to which suitable fuzzy sets μ_i and μ are associated. Let us furthermore assume that there are similarity relations S_i and S on X_i and Y, the domains of the variables ξ_i and η, such that the fuzzy sets μ_i and μ correspond to the extensional hulls of single points $x_0^{(i)}$ and y_0, respectively, i.e. $\mu_i(x_i) = S_i(x_i, x_0^{(i)})$ and $\mu(y) = S(y, y_0)$. In this way we can interpret the fuzzy control rule as the a (vague) specification of the point $(x_0^{(1)}, \ldots, x_0^{(n)}, y_0)$ in the multi–dimensional space $X_1 \times \ldots \times X_n \times Y$ so that the rule simply states that this point is an element of the graph of the control function, i.e. it belongs to the control surface.

When we are dealing with r fuzzy control rules

If ξ_1 is $A_1^{(j)}$ and $\ldots$ ξ_n is $A_n^{(j)}$ then η is $B^{(j)}$ $\qquad (j = 1, \ldots, r)$

where we associate the fuzzy sets $\mu_i^{(j)}$ and $\mu^{(j)}$ with the linguistic terms $A_i^{(j)}$ and $B^{(j)}$, we obtain in the same way for each rule one point $(x_0^{(1,j)}, \ldots, x_0^{(n,j)}, y_0^{(j)})$ of the control function so that we know that

$$\varphi_0 \;=\; \big\{(x_0^{(1,j)}, \ldots, x_0^{(n,j)}, y_0^{(j)}) \mid j \in \{1, \ldots, r\}\big\} \tag{5}$$

is contained in the graph of the control function.

The similarity relations S_i and S can be joint together to the similarity relation $S_{\min\{S_1,\ldots,S_n,S\}}$ on the product space $X_1 \times \ldots \times X_n \times Y$. In the same way as in Definition 2.2 a single point x_0 induces the fuzzy set $\mu_{x_0}(x) = S(x, x_0)$ by taking its extensional hull with respect to the similarity relation S, the extensional hull of a set M can be defined as the fuzzy set

$$\mu_M(x) \;=\; \sup\{S(x, x_0) \mid x_0 \in M\}. \tag{6}$$

μ_M can be understood as the (fuzzy) set of elements that are similar to at least one of the elements of M.

The extensional hull of the set (5) corresponding to the fuzzy set of elements that are similar to at least one of the points of the graph of the control function given by the control rules, with respect to the similarity relation is therefore

$$\begin{aligned}
\mu_{\varphi_0}(x_1, \ldots, x_n, y) &= \sup_{j\in\{1,\ldots,r\}} \left\{S_{\min\{S_1,\ldots,S_n,S\}}\big((x_1, \ldots, x_n, y), (x_0^{(1,j)}, \ldots, x_0^{(n,j)}, y_0^{(j)})\big)\right\} \\
&= \sup_{j\in\{1,\ldots,r\}} \Big\{ \min\big\{ \min_{i\in\{1,\ldots,n\}} \{S_i(x_i, x_0^{(i,j)})\}, S(y, y_0^{(j)})\big\}\Big\} \qquad (7) \\
&= \sup_{j\in\{1,\ldots,r\}} \Big\{ \min\big\{ \min_{i\in\{1,\ldots,n\}} \{\mu_{x_0^{(i,j)}}(x_i)\}, \mu_{y_0^{(j)}}(y)\big\}\Big\} \\
&= \sup_{j\in\{1,\ldots,r\}} \Big\{ \min\big\{ \min_{i\in\{1,\ldots,n\}} \{\mu_i^{(j)}(x_i)\}, \mu^{(j)}(y)\big\}\Big\}.
\end{aligned}$$

If an input tuple $(x_1, \ldots, x_n)$ is given, by using (7) we can compute for each possible output value y the degree to which the tuple $(x_1, \ldots, x_n, y)$ is similar to at least one of the points of the control function specified by the control rules and use this as the grade to which y is accepted as a suitable output. Thus we obtain for the given input $(x_1, \ldots, x_n)$ the fuzzy set

$$\mu_{(x_1,\ldots,x_n)}(y) \;=\; \max_{j\in\{1,\ldots,r\}} \Big\{ \min\big\{ \min_{i\in\{1,\ldots,n\}} \{\mu_i^{(j)}(x_i)\}, \mu^{(j)}(y)\big\}\Big\}$$

which is exactly the same result that the max–min rule yields (before defuzzification).

In the view of interpreting fuzzy sets on the basis of similarity relations, the principal concept of fuzzy control is the following. A control function is specified by some 'typical' points $(x_0^{(1,j)}, \ldots, x_0^{(n,j)}, y_0^{(j)})$ on the control surface plus some notion of similarity that allows us to determine to which degree a given point can be considered to be similar to one of the typical points of the control function. As we have seen, the similarity does not have to be specified explicitly. It can be derived from given fuzzy sets whenever the fuzzy sets satisfy reasonable restrictions like they are for example described in Theorem 2.3.

The similarity relation $S_{\min\{S_1,\ldots,S_n,S\}}$ on the product space $X_1 \times \ldots \times X_n \times Y$ is an aggregation of the similarity relation S_i and S on the spaces X_i and Y, respectively.

These simple similarity relations on one–dimensional spaces are induced by a scaling factor based transformation of the standard distance between real numbers. In the above considerations we chose as the aggregation operation for the similarity relations the minimum motivated by the fact that it is the coarsest similarity relation that distinguishes at least as well as the given similarity relations. In some sense this corresponds to the assumption that the similarity relations are non–interacting. If we discard this assumption other operations than the minimum have to be considered.

3 Fuzzy Cluster Analysis and Similarity

In the previous section we have discussed fuzzy control from the viewpoint of similarity relations. It turned out that in this light the control function is specified by some 'typical' input–output tuples and a similarity relation which determines the degree to which an arbitrary tuple is similar to one of the typical ones. Let us consider the typical input–output tuple $(x_0^{(1,j)}, \ldots, x_0^{(n,j)}, y_0^{(j)})$. Taking the similarity relation $S_{\min\{S_1,\ldots,S_n,S\}}$ that was explained in the previous section into account $(x_0^{(1,j)}, \ldots, x_0^{(n,j)}, y_0^{(j)})$ induces the fuzzy set

$$\mu_{(x_0^{(1,j)},\ldots,x_0^{(n,j)},y_0^{(j)})}(x_1, \ldots, x_n, y) \;=\; S_{\min\{S_1,\ldots,S_n,S\}}\big((x_1, \ldots, x_n, y), (x_0^{(1,j)}, \ldots, x_0^{(n,j)}, y_0^{(j)})\big)$$

of all tuples that are similar to $(x_0^{(1,j)}, \ldots, x_0^{(n,j)}, y_0^{(j)})$. Thus the space $X_1 \times \ldots \times X_n \times Y$ is 'partitioned' by the fuzzy sets induced by the the typical input–output tuples. A very similar kind of fuzzy partition is the aim of many fuzzy clustering algorithms. Each (fuzzy) cluster is represented by a prototype and the membership degrees of the data to the cluster that are depending on the distances between the prototypes and the data, and in the ideal are decreasing with increasing distance.

Let us briefly review some objective function based fuzzy clustering methods. The cluster algorithm aims at minimizing the objective function

$$J(X, U, v) \;=\; \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d^2(v_i, x_k) \tag{8}$$

under the constraints

$$\sum_{k=1}^{n} u_{ik} \;>\; 0 \qquad \text{for all } i \in \{1, \ldots c\} \tag{9}$$

and

$$\sum_{i=1}^{c} u_{ik} \;=\; 1 \qquad \text{for all } k \in \{1, \ldots n\}. \tag{10}$$

$X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^p$ 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, $v_i \in \mathbb{R}^p$ is the prototype 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.

The objective function (8) to be minimized uses the sum over the quadratic distances of the data to the prototypes weighted with their membership degrees. (9) guarantees that no cluster is completely empty, (10) ensures 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 1 for each datum.

Differentiating (8) one obtains

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d^2(v_i, x_k)}{d^2(v_j, x_k)} \right)^{\frac{1}{m-1}}} \tag{11}$$

and

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

as a necessary condition for (8) to have a (local) minimum. The equations (11) and (12) are therefore used for updating the membership degrees u_{ik} and the prototypes v_i in an iteration procedure until the difference between the matrix (u_{ik}^{new}) and the matrix (u_{ik}^{old}) in the previous iteration step is less than a given tolerance bound ε.

The most simple fuzzy clustering algorithm is the fuzzy c–means (FCM) (see f.e. [1]) where the distance d is simply the Euclidean distance. It searches for spherical clusters of approximately the same size.

Gustafson and Kessel [4] and Gath and Geva [2] designed fuzzy clustering methods that are looking for hyper–ellipsoidal clusters of varying size. We refer to the corresponding algorithms by the abbreviations GK and GG, respectively. In both cases, in addition to the prototypes v_i and the membership degrees u_{ik} for each cluster i a (positive definite) covariance matrix C_i is calculated. The GK replaces the Euclidean distance by the transformed Euclidean distance

$$d^2(v_i, x_k) = (\rho_i \det C_i)^{1/p} \cdot (x_k - v_i)^\top C^{-1} (x_k - v_i), \tag{13}$$

whereas the GG is based on normal distributions and uses the distance

$$d^2(v_i, x_k) = \frac{(\det(C_i))^{1/2}}{p_i} \cdot \exp\left(\frac{(x_k - v_i)^\top C^{-1} (x_k - v_i)}{2} \right) \tag{14}$$

where

$$p_i = \frac{\sum_{k=1}^{n} (u_{ik})^m}{\sum_{j=1}^{c} \sum_{k=1}^{n} (u_{jk})^m} \tag{15}$$

so that in both cases for each cluster a matrix inversion and a determinant has to be computed in every iteration step. For the GK the size of each cluster has to be specified implicitly in advance by the value ρ_i, whereas in the GG the sizes of the clusters need not be known in advance. For the GG it is assumed that each cluster is associated with a normal distribution and that the distance of a datum to a prototype is inversely proportional to the a posteriori probability (likelihood) that the datum was generated by the normal distribution associated with the cluster belonging to the prototype. (15) estimates the a priori probability that a datum is generated by cluster i.

The GK and GG algorithm are almost in the spirit of what was motivated by our considerations of fuzzy control. The prototypes are exactly what we need as 'typical' input–output tuples. The membership degrees are computed on the basis of scaled distances encoded in the matrices C_i. However, the matrices also incorporate rotations that do not fit into our framework. Speaking in terms of fuzzy control this would mean that instead of directly using the input and output we would apply an additional transformation which is reasonable in data analysis but not in fuzzy control since control rules on transformed variables are not easily understood and cannot be well interpreted. Thus, we would prefer only a scaling of the distance in the direction of the coordinate axes without an additional rotation. This means that we have to restrict the matrices C_i appearing in the GK and the GG to diagonal matrices. Since in the derivation of the GK and the GG (fuzzy) covariance matrices are needed, one does in general not obtain a diagonal matrix. Therefore, it is necessary to derive new formulae for updating the matrices C_i that induce the transformation of the Euclidean distance via (13) and (14).

Let us first consider the modification of the GK. We rewrite the distance (13) in the form

$$d^2(v_i, x_k) \;=\; (x_k - v_i)^\top A_i (x_k - v_i)$$

where we require that $\det A_i = \rho_i$, i.e. A_i equals $(\rho_i \det C_i)^{1/p} C^{-1}$. Now we assume that A_i is a diagonal matrix. The objective function (8) depends in this case in addition to X, U, and v on the c diagonal matrices $A_1, \ldots, A_c$ for which we require $\det A_i = \rho_i$. Let $a_\alpha^{(i)}$ denote the α^{th} diagonal element of the matrix A_i. Thus we assume

$$\rho_i \;=\; \prod_{\alpha=1}^{p} a_\alpha^{(i)}. \tag{16}$$

We optimize the matrices A_i individually. Taking care of the condition (16) by a Lagrange multiplier we have to minimize the function

$$F(a^{(i)}, \lambda) \;=\; \sum_{k=1}^{n} (u_{ik})^m \sum_{\alpha=1}^{p} a_\alpha^{(i)} (x_{k,\alpha} - v_{i,\alpha})^2 - \;\lambda \left(\left(\prod_{\alpha=1}^{p} a_\alpha^{(i)} \right) - \rho_i \right). \tag{17}$$

Thus, requiring that the partial derivative of (17) with respect to λ is zero is equivalent to (16).

Let w be an arbitrary unit vector. Then we obtain

$$\frac{\partial F}{\partial w}(a^{(i)}) \;=\; \lim_{t \to 0} \frac{F(a^{(i)} + t \cdot w, \lambda) - F(a^{(i)}, \lambda)}{t}$$

$$
\begin{aligned}
&= \lim_{t\to 0}\frac{1}{t}\left(\sum_{\alpha=1}^{p} tw\sum_{k=1}^{n}(u_{ik})^m(x_{k,\alpha}-v_{i,\alpha})^2 \;-\; \lambda\sum_{\alpha=1}^{p} tw_\alpha \prod_{\beta=1,\,\beta\neq\alpha}^{p} a_\beta^{(i)} \;+\; o(t^2)\right)\\
&= \sum_{\alpha=1}^{p} w_\alpha \sum_{k=1}^{n}(u_{ik})^m(x_{k,\alpha}-v_{i,\alpha})^2 - \;\lambda\sum_{\alpha=1}^{p} w_\alpha \prod_{\beta=1,\,\beta\neq\alpha}^{p} a_\beta^{(i)}\\
&= \sum_{\alpha=1}^{p} w_\alpha\left(\sum_{k=1}^{n}(u_{ik})^m(x_{k,\alpha}-v_{i,\alpha})^2 \;-\; \lambda\prod_{\beta=1,\,\beta\neq\alpha}^{p} a_\beta^{(i)}\right)\\
&= 0
\end{aligned}
$$

independent of w. This implies that

$$
\lambda \prod_{\beta=1,\,\beta\neq\gamma}^{p} a_\beta^{(i)} \;=\; \sum_{k=1}^{n}(u_{ik})^m(x_{k,\gamma}-v_{i,\gamma})^2 \tag{18}
$$

holds for all $\gamma \in \{1,\ldots,p\}$. Taking (16) into account, we may replace the left–hand side of equation (18) by $\rho_i/a_\gamma^{(i)}$ and obtain

$$
a_\gamma^{(i)} \;=\; \frac{\lambda\rho_i}{\sum_{k=1}^{n}(u_{ik})^m(x_{k,\gamma}-v_{i,\gamma})^2}.
$$

Using this result in equation (16) yields

$$
\lambda \;=\; \left(\rho_i^{1-p}\prod_{\alpha=1}^{p}\sum_{k=1}^{n}(u_{ik})^m(x_{k,\gamma}-v_{i,\gamma})^2\right)^{1/p}
$$

so that we finally obtain

$$
a_\gamma^{(i)} \;=\; \frac{\left(\rho_i\prod_{\alpha=1}^{p}\sum_{k=1}^{n}(u_{ik})^m(x_{k,\alpha}-v_{i,\alpha})^2\right)^{1/p}}{\sum_{k=1}^{n}(u_{ik})^m(x_{k,\gamma}-v_{i,\gamma})^2} \tag{19}
$$

as the updating scheme for the modified version of the GK. (The other parameters are again updated in the same way as in the original GK by equations (11) and (12).)

Let us now turn to the modification of the GG. As for the original GG we assume that each cluster is associated with a normal distribution and that the distance of a datum to a prototype is inversely proportional to the a posteriori probability (likelihood) that the datum was generated by the normal distribution associated with the cluster belonging to the prototype. Therefore, we also use (14) with the additional assumption that the covariance matrix C_i is a diagonal matrix with $1/c_\alpha^{(i)}$ as its α^{th} diagonal element.

Let

$$f_i(x) = \frac{1}{(2\pi)^{p/2}} \cdot \frac{1}{\sqrt{\prod_{\alpha=1}^{p} \frac{1}{c_\alpha^{(i)}}}} \cdot \exp\left(-\frac{1}{2} \sum_{\alpha=1}^{p} c_\alpha^{(i)} (x_{k,\alpha} - v_{i,\alpha})^2 \right)$$

denote the density function of the normal distribution associated with cluster i. Thus the a posteriori probability (likelihood) that datum x_k was generated by the normal distribution associated with cluster i is $p_i \cdot f_i(x_k)$ where p_i is the a priori probability that a datum is generated by cluster i.

We now have to choose for each cluster i the (diagonal) covariance matrix in such a way that the a posteriori probability is as high as possible for those data belonging to the cluster. This corresponds to computing the maximum likelihood estimator. The a posteriori probability (likelihood) that *all* data are generated by cluster i is

$$\prod_{k=1}^{n} p_i f_i(x_k). \tag{20}$$

However, we only have to take those data into account that actually belong to the cluster. Therefore, we modify (20) by

$$\prod_{k=1}^{n} \left(p_i f_i(x_k)\right)^{(u_{ik})^m}. \tag{21}$$

Note that in the crisp case, i.e. $u_{ik} \in \{0, 1\}$, (21) becomes

$$\prod_{k:\, x_k \text{ belongs to cluster } i} p_i f_i(x_k).$$

Instead of maximizing (21) directly we maximize the logarithm of (21), i.e.

$$\begin{aligned} F(c_1^{(i)}, \ldots, c_p^{(i)}) = \sum_{k=1}^{n} (u_{ik})^m \Bigg(& \ln(p_i) - \frac{p}{2} \ln(2\pi) + \frac{1}{2} \sum_{\alpha=1}^{p} \ln(c_\alpha^{(i)}) \\ & - \frac{1}{2} \sum_{\alpha=1}^{p} c_\alpha^{(i)} (x_{k,\alpha} - v_{i,\alpha})^2 \Bigg). \end{aligned} \tag{22}$$

Taking the partial derivatives of (22) we obtain

$$\frac{\partial F(c_1^{(i)}, \ldots, c_p^{(i)})}{\partial c_\alpha^{(i)}} = \frac{1}{c_\alpha^{(i)}} \cdot \frac{1}{2} \sum_{k=1}^{n} (u_{ik})^m - \frac{1}{2} \sum_{k=1}^{n} (u_{ik})^m (x_{k,\alpha} - v_{i,\alpha})^2 \tag{23}$$

To maximize (22) it is necessary that (23) equals zero so that we finally derive

$$c_\alpha^{(i)} = \frac{\sum_{k=1}^{n} (u_{ik})^m}{\sum_{k=1}^{n} (u_{ik})^m (x_{k,\alpha} - v_{i,\alpha})^2} \tag{24}$$

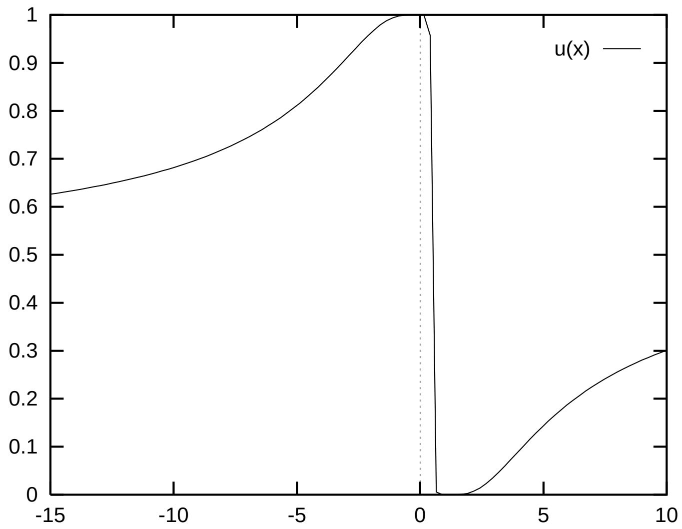

Figure 1: The membership function for cluster 0.

as the updating scheme of the modified version of the GG. (The other parameters are again updated in the same way as in the original GG using equations (11), (12), and (15).)

Note that neither the inverse nor the determinant of a matrix has to be computed in the modified versions of the GK and the GG so that the corresponding algorithms are much simpler and faster than the original GK and GG.

In the ideal case when there is very few noise in the data, fuzzy clusters obtained by an algorithm based on the objective function (8) usually have the desired property that the membership degrees of the data to a cluster decrease with increasing distance. However, since the membership degree u_{ik} of datum x_k to cluster i depends only on the relative distance between its distance to prototype v_i and the other prototypes (compare equation (11)), a datum that is far away from all clusters will be assigned the membership degree of approximately $1/c$ to each cluster. To illustrate this problem let us for reasons of simplicity consider one–dimensional data that were divided into two fuzzy clusters by the FCM with 0 and 1 as prototypes. According to equation (11) the membership degree of datum x to the cluster with prototype 0 is

$$u(x) \;=\; \frac{1}{1+\left(\frac{x^2}{(1-x)^2}\right)^2}$$

where we have chosen $m = 2$. The graph of this function is shown in Figure 1. Note that the membership degree is increasing for negative values x reaching the maximum value 1 at $x = 0$ at the corresponding prototype 0. For positive x it is decreasing to the membership degree 0 at the other prototype 1 and then the function is increasing again to 0.5 for $x \to \infty$.

Krishnapuram and Keller [9] introduced the notion of possibilistic clustering to avoid this anomaly. The name possibilistic clustering is motivated by the fact that they no longer require the probabilistic condition (10) that the membership degrees to the fuzzy clusters sum up to one for each datum. For possibilistic clustering the objective function (8) is replaced by

$$J(X,U,v) \;=\; \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d^2(v_i,x_k) + \sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-u_{ik})^m. \tag{25}$$

The second sum guarantees that the trivial optimum $u_{ik}=0$ for all i,k is excluded. The parameters η_i are estimated in advance by

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

and are not updated during the iteration procedure. Requiring that the first derivative of the modified objective function (25) equals zero leads to the necessary condition that the prototypes satisfy the same equation (12) as for probabilistic clustering, whereas the membership degrees have to fulfill

$$u_{ik} \;=\; \frac{1}{1+\left(\frac{d^2(v_i,x_k)}{\eta_i}\right)^{\frac{1}{m-1}}}. \tag{26}$$

In principal any of the mentioned fuzzy clustering algorithm can be modified to possibilistic clustering simply by replacing (12) by (26) in the updating scheme. However, the GG algorithm and our proposed modification are based on a probabilistic interpretation so that a possibilistic version counteracts the underlying semantics.

In order to estimate the parameters η_i and to have a good initialization one should first apply the corresponding probabilistic fuzzy clustering algorithm and use it as an initialization for the possibilistic version.

Even for the probabilistic case it is also recommended to use the result of an FCM run as initialization for more complicated algorithms like the GK and the GG. Due to the high number of parameters that have to be optimized in the GK and the GG a bad (random) initialization can lead to undesirable results.

4 Construction of a Fuzzy Controller

The idea of deriving fuzzy if–then rules from fuzzy clusters is the following. We apply a fuzzy clustering algorithm to the given data and then obtain for each cluster a rule by projecting the fuzzy cluster to the one–dimensional coordinate spaces. Thus, if the fuzzy set $\mu_\alpha^{(i)}$ is the projection of the fuzzy cluster i to the α^{th} coordinate space the corresponding rule is

If ξ_1 is $\mu_1^{(i)}$ and ... and ξ_{p-1} is $\mu_{p-1}^{(i)}$ then ξ_p is $\mu_p^{(i)}$

where $\xi_1, \ldots, \xi_{p-1}$ are the input variables and ξ_p is the output variable. Of course, appropriate linguistic terms have to be associated with the fuzzy sets $\mu_\alpha^{(i)}$.

In principal all fuzzy cluster algorithms mentioned in the previous section could be applied. The disadvantage of the FCM is however that it is not very flexible, since it only admits spherical clusters of approximately the same size. Although this problem can be overcome by the GK and the GG, in these cases the derived fuzzy rules are in general not very coherent with the fuzzy clusters. One cannot avoid a certain loss of information by projecting the fuzzy clusters, since the same effect as for crisp projections appears. The Cartesian product (computed by the derived rules) of the projections of a sphere or an ellipsoid yields the smallest rectangle containing the sphere/ellipsoid. The difference between the original ellipsoid and the rectangle is small when the axes of the ellipsoid are parallel to the coordinate axes. Therefore, our proposed modified versions of the GK and the GG are a compromise between flexibility and loss of information.

We have not explicitly explained how the projections of a fuzzy cluster are computed. From a mathematical point of view the membership degree of the value y to the α^{th} projection $\mu_\alpha^{(i)}$ of the fuzzy cluster i is the supremum over the membership degrees of all vectors with y as α^{th} component to the fuzzy cluster i, i.e.

$$\mu_\alpha^{(i)}(y) \;=\; \sup\left\{ \frac{1}{1+\left(\frac{d^2(v_i,x)}{\eta_i}\right)^{\frac{1}{m-1}}} \;\middle|\; x=(x_1,\ldots,x_{\nu-1},y,x_{\nu+1},\ldots,x_p)\in \mathbb{R}^p \right\}$$

or

$$\mu_\alpha^{(i)}(y) \;=\; \sup\left\{ \frac{1}{\sum_{j=1}^{c}\left(\frac{d^2(v_i,x)}{d^2(v_j,x)}\right)^{\frac{1}{m-1}}} \;\middle|\; x=(x_1,\ldots,x_{\nu-1},y,x_{\nu+1},\ldots,x_p)\in \mathbb{R}^p \right\}$$

in the possibilistic case.

Since this is difficult to compute and leads also to quite complicated fuzzy sets, we use the following method. We project the data (with their membership degrees to the considered fuzzy cluster) and obtain a discrete one–dimensional fuzzy set. A typical result of this procedure is shown in Figure 2. In order to extend this discrete fuzzy set to all real numbers we compute the convex hull or when simple membership functions are required, we approximate the convex hull for instance by a trapezoidal function by a heuristic algorithm that aims at minimizing the sum of quadratic errors. A detailed description of such a method can be found in [11].

In order to illustrate our method we consider the function $f(x,y)=\sin(x)\cdot\cos(y)$ for $0 \le x \le 3.1$ and $0 \le y \le 6.2$ shown in Figure 3.

We used an artificial data set with 91 data that are approximately uniform distributed over the input set $[0,3.1]\times[0,6.2]$ (compare Figure 4).

We first applied the FCM to this data set with a fixed number of five clusters and used the result as an initialization for our modified versions of the GK and GG.

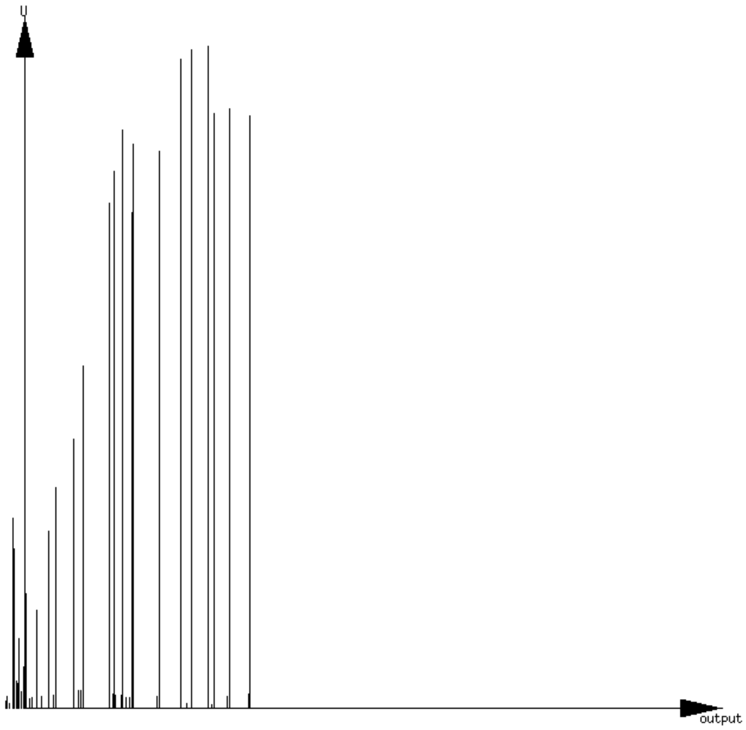

Figure 2: A typical projection of a fuzzy cluster.

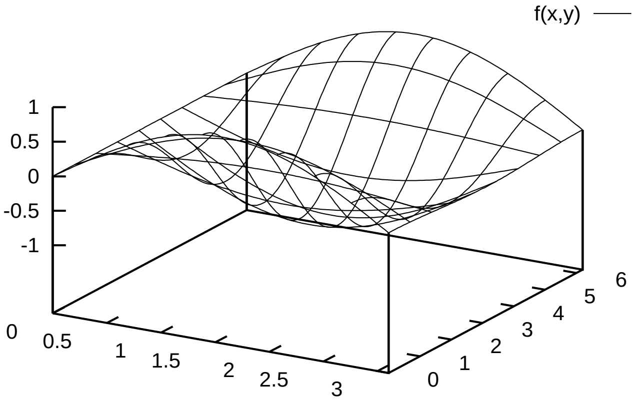

Figure 3: The function $f(x,y) = \sin(x) \cdot \cos(y)$.

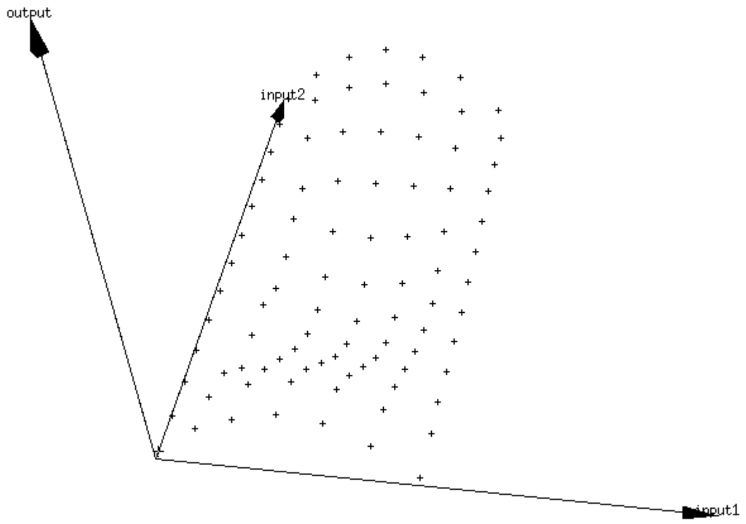

Figure 4: The data set.

The results of the cluster analysis are shown in Figures 5–7. In these pictures each datum is connected to the prototype which represents the fuzzy cluster yielding the highest membership degree for the datum.

From each of these clustering results we obtain a rule base with five rules by projecting the fuzzy clusters and approximating the projections by trapezoidal fuzzy sets. The rule base induced by the modified GK clustering result can be seen in Figure 8.

The three rule bases derived from the clustering results of the FCM and the modified versions of the GK and the GG can be applied to the data (with max–min inference and simplified defuzzification where the output fuzzy set of each single rule is defuzzified before aggregation). Figures 9–11 compare the original data set with the results that are obtained when the output is computed on the basis of the corresponding rule bases. Each datum is connected with the output computed on the basis of the rule base so that longer lines indicate greater errors. It is obvious that the FCM shows the worst performance of the three algorithms.

Five rules are of course a very small number. In order to obtain a better approximation one has to increase the number of rules (clusters). Figures 12, 13, and 14 provide the results of the modified GG clustering with nine clusters, the induced rule base, and the comparison with the original data.

Figure 14 shows an improvement in comparison with the results obtained using only five rules, but still not a solution that is fully satisfactory. The approximation in the inner area is quite good. The errors are mainly lying on the boundary of the input set. The reason for this is that the data in the inner area are covered by more than one (fuzzy) clusters whereas the data on the boundary are usually assigned to only one cluster. Therefore, one should try to extrapolate some data going further

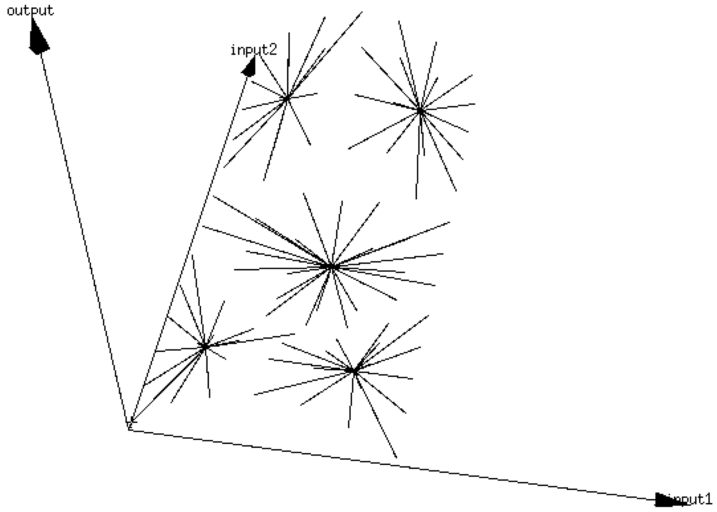

Figure 5: The result of the FCM with five clusters.

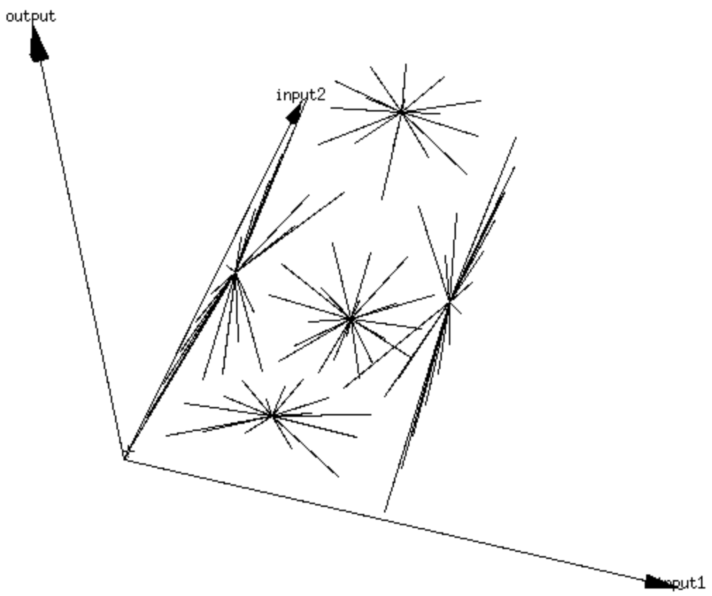

Figure 6: The result of the modified GK with five clusters.

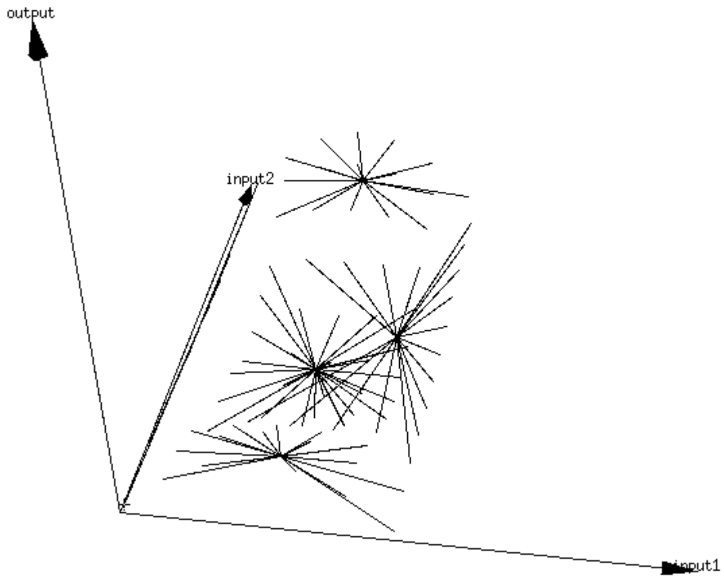

Figure 7: The result of the modified GG with five clusters.

than the boundary in order to have the relevant data within the inner area where the approximation is good enough.

5 Discussion

In this paper we have elucidated the close relation between fuzzy cluster analysis and fuzzy control. Fuzzy control can be interpreted as an approximation technique based on

- some typical points of the graph of the control function to be approximated (each point corresponding to one linguistic control rule),
- some indistinguishability concept which can be formalized by similarity relations that are induced by transformations of the canonical distance between real numbers,
- and on the representation of the typical points of the function by fuzzy sets that take the indistinguishability induced by the similarity relations into account.

A similar concept is used in fuzzy cluster analysis where each fuzzy cluster is represented by a prototype (typical element) and a fuzzy set whose membership degrees are computed on the basis of a transformed distance. Thus it seems to be natural to apply fuzzy clustering in order to derive fuzzy rules from given data.

It should be emphasized that the goal of this method is not to find the best approximation of the data with respect to sum error measure. For this purpose

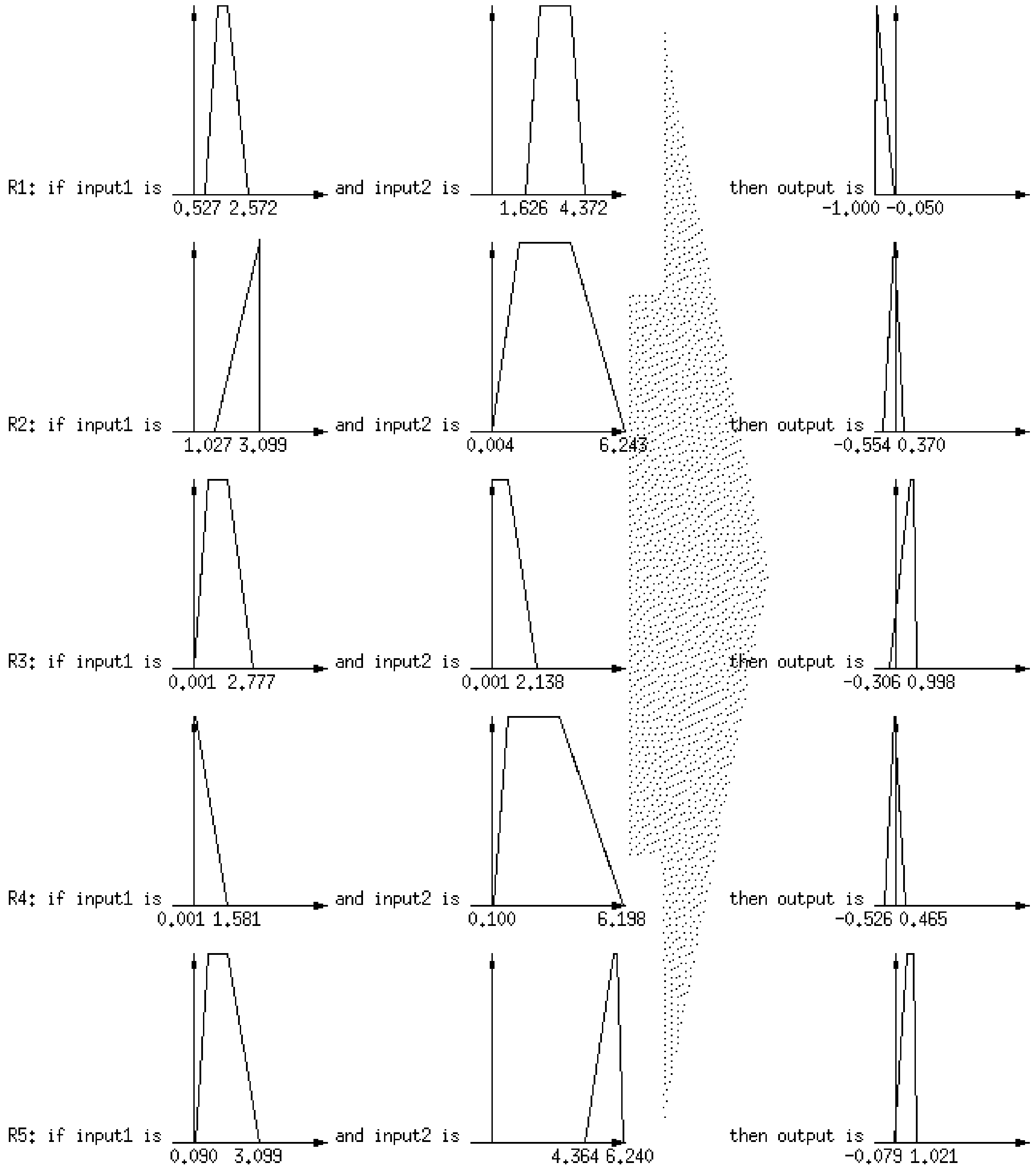

Figure 8: The rule base induced by the clustering result of Figure 6.

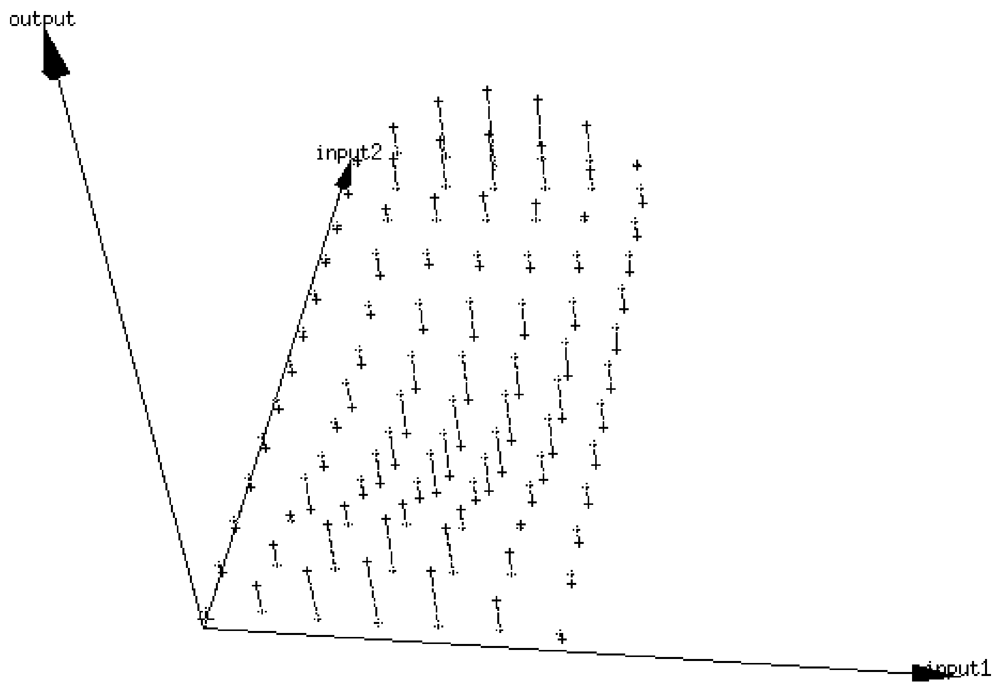

Figure 9: Comparison of the original data with what is obtained from the rule base induced by the FCM result.

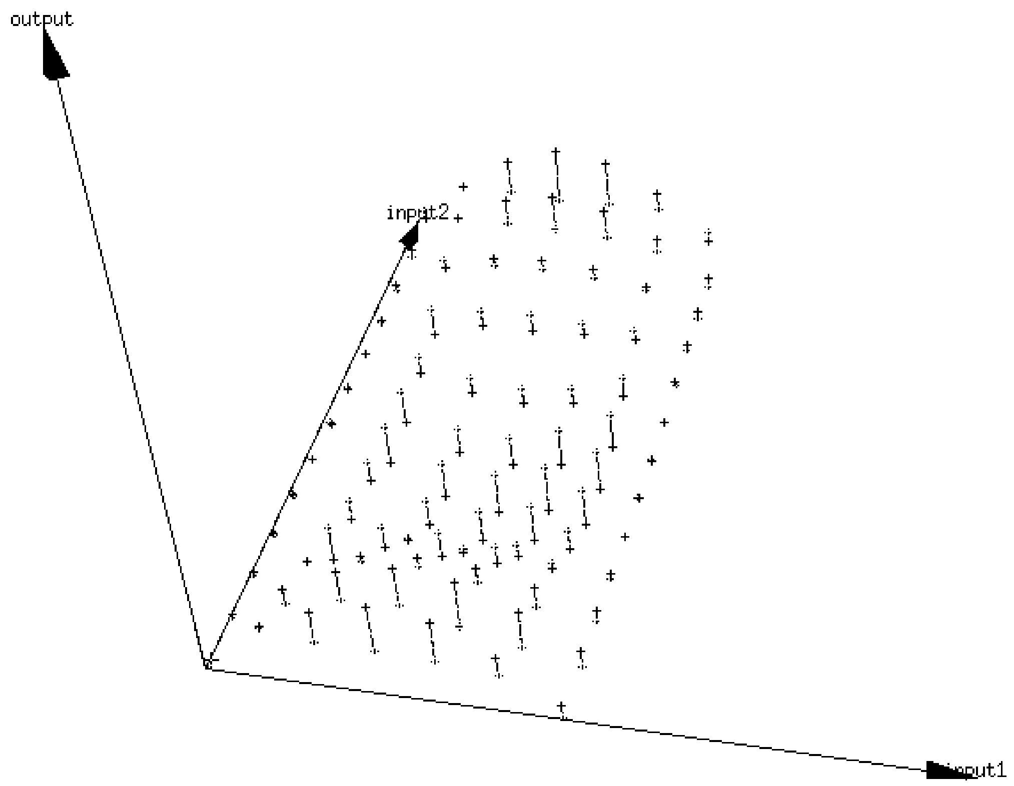

Figure 10: Comparison of the original data with what is obtained from the rule base induced by the modified GK result.

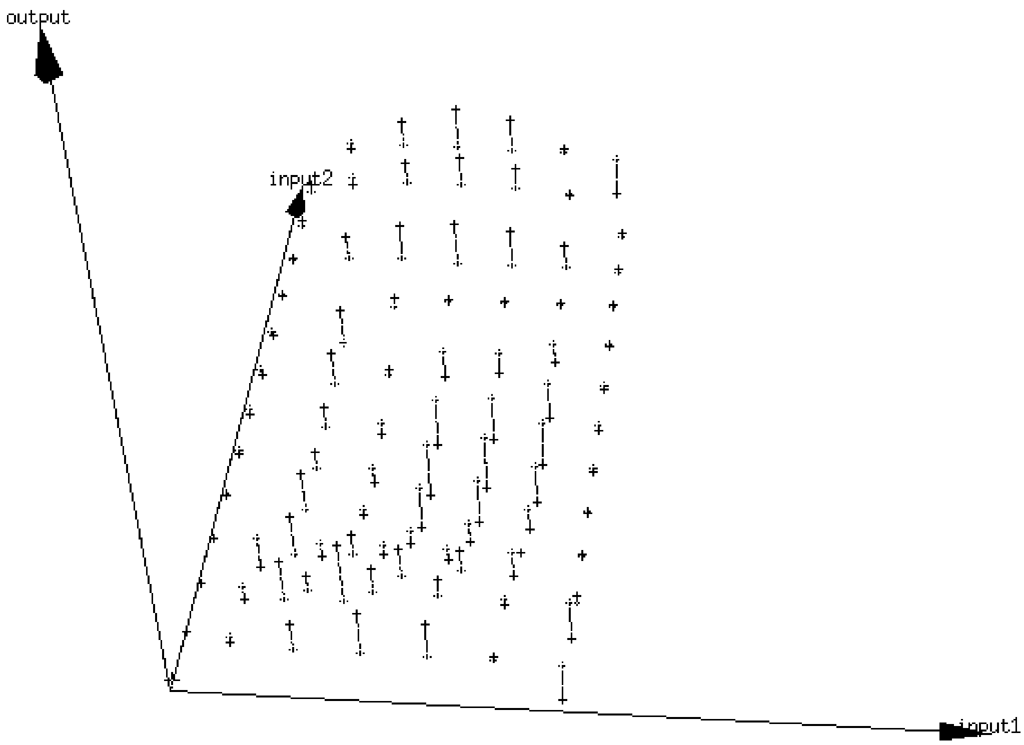

Figure 11: Comparison of the original data with what is obtained from the rule base induced by the modified GG result.

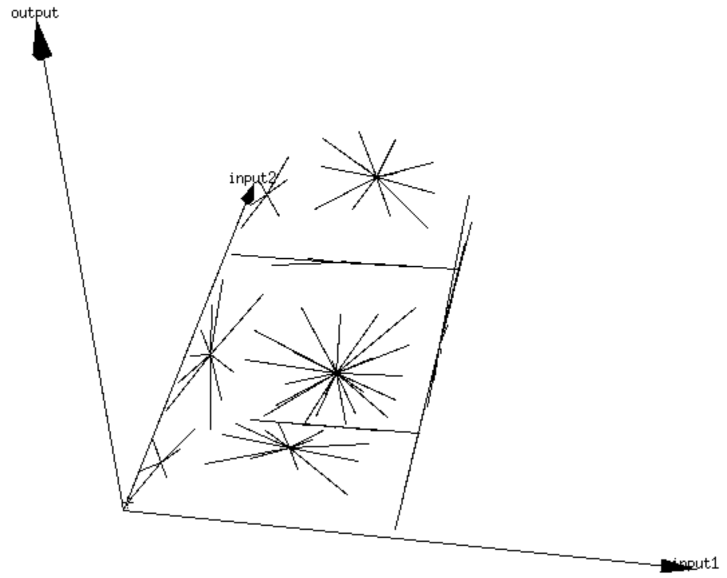

Figure 12: The result of the modified GG with nine clusters.

other mathematical approximation techniques are more suitable. The aim is to derive (linguistic) rules from the data that can be understood by humans instead of finding an abstract best fitting function.

One should be aware of the fact that the method for the construction of the rules incorporates a certain loss of information in comparison with the original fuzzy clusters. First of all, the approximation of the projections of the fuzzy clusters by trapezoidal functions causes a loss of information. But even if one would use the correct projections themselves, the original clusters cannot be reconstructed from the rules. The reason for this is that the aggregation of the fuzzy sets appearing in one rule by the minimum corresponds to building their Cartesian product, i.e. what we obtain from the rules are the Cartesian products of the projections of the original clusters, which gives in general larger fuzzy sets (in the product space) than the original ones. The amount of enlargement depends on the form of the original fuzzy set. The original fuzzy set can only be reconstructed from the projections when all α–cuts are axes parallel rectangles. For the considered fuzzy clustering techniques we obtain as α–cuts spheres for the FCM, arbitrary ellipsoids for the GK and GG, and axes parallel ellipsoids for our modified versions of the GK and GG. The loss of information caused by the projection corresponds to the difference between the smallest rectangle containing the corresponding geometric form (sphere, ellipsoid, axes parallel ellipsoid) and the geometric form itself. It is obvious that non–axes parallel ellipsoids cause the biggest difference. Therefore, our modified versions of the GK and GG can be seen as a compromise between flexibility and loss of information, since they are less restrictive than the FCM, which only allows

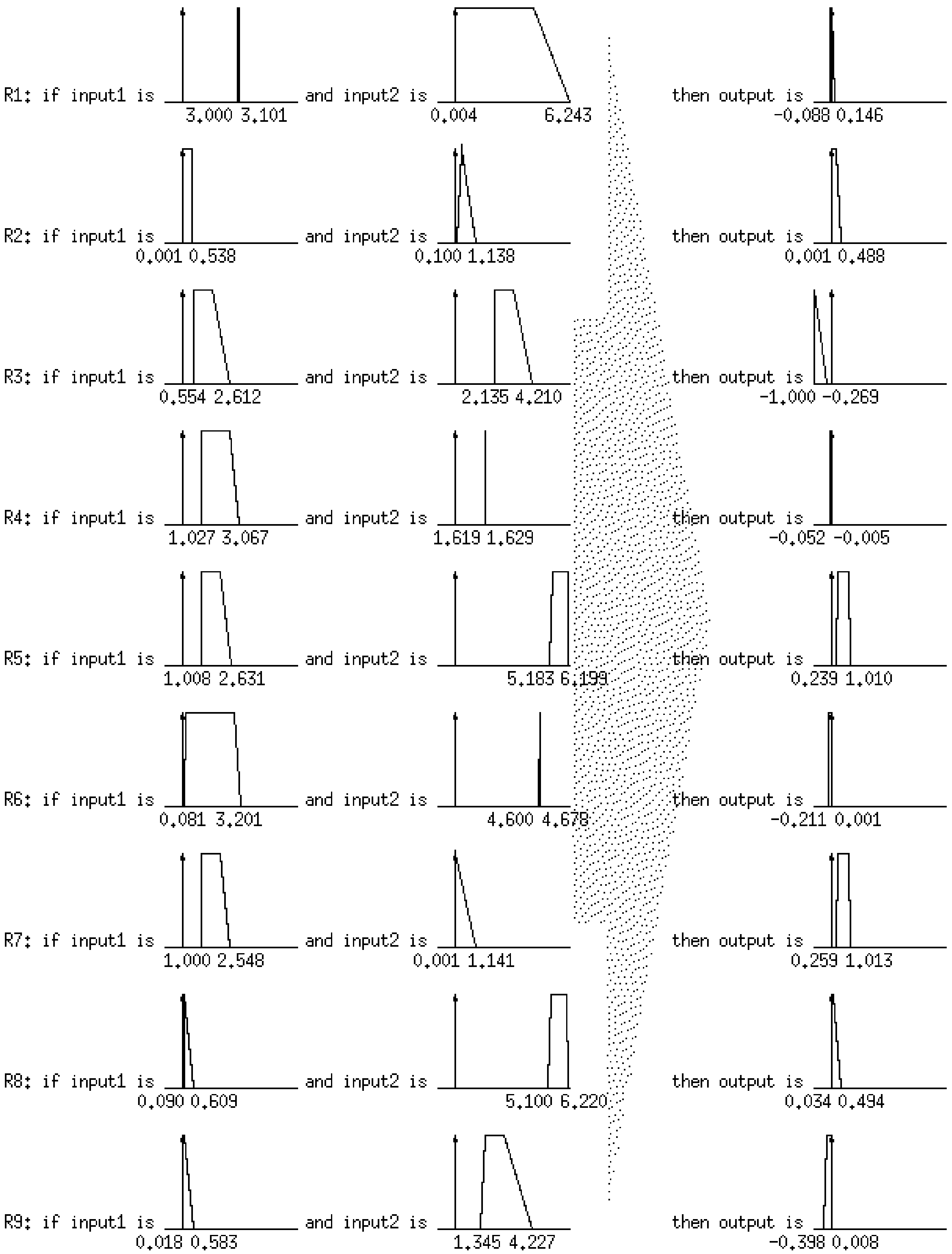

Figure 13: The rule base induced by the clustering result of Figure 12.

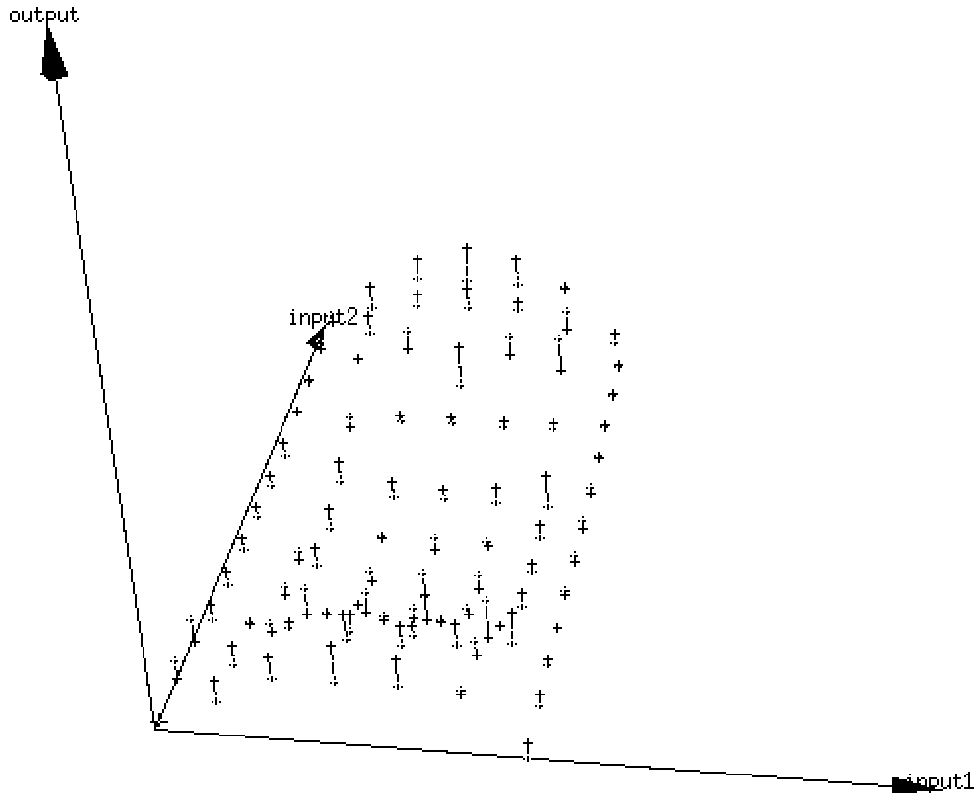

Figure 14: Comparison of the original data with what is obtained from the rule base induced by the modified GG result with nine clusters.

spheres, but avoid the great loss of information caused by non–axes parallel ellipsoids that may result from the original GK and the GG.

Constructing a rule base from fuzzy clusters gives a first approximation for the data which can be used as a basis for further improvements either by inspecting the rules and fuzzy sets 'by hand' or by applying for instance fuzzy–neuro techniques for tuning the fuzzy sets.

Another problem is to determine the number of rules which is identical to the number of clusters. If one does not want to fix the number of rules in advance it is possible to use a cluster validity measure to compute the number of rules/clusters. For proposals for various cluster validity measures see [1, 2, 13, 11].

As already mentioned one should be aware of the fact that the projection of a fuzzy cluster might not be interpretable as a convex fuzzy number (meaning that the α–cuts are not necessarily convex). Instead of applying possibilistic clustering to avoid this undesired effect one can also define a threshold for the membership degrees to be considered. In order to obtain a convex fuzzy set one starts at the point of the projection of a fuzzy cluster where the membership degree is maximal and then goes to the right and to the left taking all other elements of the projection into account as long as their membership degree exceeds the threshold. Once a datum has membership degree lower than the threshold, itself and all the data beyond it (even if the membership degrees are increasing again) are neglected in the projection of the cluster.

In [11] Sugeno and Yasukawa proposed to apply fuzzy clustering only to the output data, compute their cylindrical extensions (in the product space of the input and the output data) and then use the projections of these fuzzy sets for deriving fuzzy rules from the data. [11] considers only the FCM. Of course, one can also apply our modified versions of the GK and the GG only to the output data and then continue with the procedure described in [11]. Note that the modified versions of the GK and GG coincide with the original algorithms for one–dimensional data. Nevertheless, there are enough problems with more than one output variable.

Finally, let us remark that it also reasonable to use the modified version of the GK and the GG for deriving classification rules [8]. A similar method for learning fuzzy classification rules based on the original GG was described in [3]. There it is proposed to transform the data (according to the transformations encoded in the matrices computed by the GG) before projection. In this way the loss of information for non–axes parallel ellipsoids is also reduced for the price that the rules are formulated for transformed variables and are therefore often difficult to interpret. The restriction to our modified versions of the GK and GG makes such transformations superfluous (for the price of a little less flexible clustering algorithm).

References

[1] J.C. Bezdek, Pattern recognition with fuzzy objective function algorithms, (Plenum Press, New York, 1981).

[2] I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (1989), 773–781.

[3] H. Genther, M. Glesner, Automatic generation of a fuzzy classification system using fuzzy clustering methods. Proc. ACM Symposium on Applied Computing (SAC'94), Phoenix (1994), 180–183.

[4] D.E. Gustafson, W.C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, Proc. IEEE CDC, San Diego (1979), 761–766.

[5] U. Höhle, M–valued sets and sheaves over integral commutative CL–monoids, in: S.E. Rodabaugh, E.P. Klement, and U. Höhle, Eds., Applications of category theory to fuzzy subsets, (Kluwer, Dordrecht, 1992) 33–72.

[6] F. Klawonn, Fuzzy sets and vague environments, Fuzzy Sets and Systems 66 (1994), 207–221.

[7] F. Klawonn, R. Kruse, Equality relations as a basis for fuzzy control. Fuzzy Sets and Systems 54 (1993), 147–156.

[8] F. Klawonn, R. Kruse, Derivation of fuzzy classification rules from multidimensional data, Proc. International Symposium on Intelligent Data Analysis (IDA–95), Baden–Baden (to appear).

[9] R. Krishnapuram, J. Keller, A possibilistic approach to clustering. IEEE Trans. on Fuzzy Systems 1 (1993), 98–110.

[10] R. Kruse, J. Gebhardt, F. Klawonn, Foundations of fuzzy systems. Wiley, Chichester (1994).

[11] M. Sugeno, T. Yasukawa, A Fuzzy–Logic–Based Approach to Qualitative Modeling, IEEE Transactions on Fuzzy Systems 1. 1993. 7–31.

[12] E. Trillas, L. Valverde, An Inquiry into Indistinguishability Operators. In: H.J. Skala, S. Termini, E. Trillas (eds.), Aspects of Vagueness. Reidel, Dordrecht (1984), 231–256.

[13] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Transactions on Pattern Analysis 13 (1991), 841–847.

[14] L.A. Zadeh, Similarity Relations and Fuzzy Orderings, Information Sciences 3 (1971), 177–200.