Derivation of Fuzzy Classification Rules from Multidimensional Data

F. Klawonn and R. Kruse
Department of Computer Science
University of Braunschweig
Bültenweg 74/75
38106 Braunschweig, Germany
E-Mail klawonn@ibr.cs.tu-bs.de

Abstract

This paper describes techniques for deriving fuzzy classification rules based on special modified fuzzy clustering algorithms. The basic idea is that each fuzzy cluster induces a fuzzy classification rule. The fuzzy sets appearing in a rule associated with a fuzzy cluster are obtained by projecting the cluster to the one–dimensional coordinate spaces. In order to allow clusters of varying shape and size we derive special fuzzy clustering algorithms which are searching for clusters in the form of axes–parallel hyper–ellipsoids. Our method can be applied to classification tasks where the classification of the sample data is known as well as when it is not known.

Keywords. fuzzy cluster analysis, classification, rule induction

1 Introduction

Fuzzy systems provide the possibility to transform linguistic descriptions into a mathematical framework in which suitable computations for processing data and formal inference can be carried out (see f.e. [7]). However, the acquisition of knowledge is often a very tedious task and the translation of linguistic rules to the framework of fuzzy sets, i.e. the choice of adequate fuzzy sets, is also a severe problem. In many cases no structural knowledge about the data is available, so that neither linguistic rules nor a fuzzy system can be specified.

Our aim is to reverse the procedure of acquiring linguistic classification rules and translating them into a fuzzy system. We start from a set of data (real vectors) which have to be classified. The goals are the construction of a fuzzy system for the classification task and the automatic generation of linguistic rules for classification. The principle idea is to apply fuzzy clustering to the data and to derive fuzzy sets and finally linguistic rules from the fuzzy clusters.

In section 2 we briefly review basic fuzzy clustering techniques and introduce a special modifications of fuzzy clustering algorithms which are adapted to rule induction. Section 3 discusses the connection between fuzzy clusters and fuzzy rules. Section 4 is devoted to unsupervised and supervised classification where the classification of the sample data is unknown or known in advance, respectively.

2 Fuzzy Cluster Analysis

Nearly all fuzzy clustering algorithms try to find an adequate prototype for each fuzzy cluster and suitable membership degrees for the data to each cluster. Usually, the cluster algorithm aims at minimizing the objective function

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

under the constraints

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

and

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

$X = \{x_1,\ldots,x_n\} \subseteq \mathbb{R}^p$ 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 $1 < m$ 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 (1) to be minimized uses the sum over the quadratic distances of the data to the prototypes weighted with their membership degrees. (2) guarantees that no cluster is completely empty, (3) 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 (1) one obtains

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

as a necessary condition for (1) to have a (local) minimum. The equations in (4) 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 [3] 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

$$d^2(v_i, x_k) \;=\; (\rho_i \det C_i)^{1/p}(x_k - c_i)^\top C^{-1}(x_k - c_i), \tag{5}$$

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

$$d^2(v_i, x_k) \;=\; \frac{(\det C_i))^{1/2}}{p_i} \exp\left(\frac{(x_k - c_i)^\top C^{-1}(x_k - c_i)}{2}\right) \tag{6}$$

where $p_i \;=\; \frac{\sum_{k=1}^{n}(u_{ik})^m}{\sum_{j=1}^{c}\sum_{k=1}^{n}(u_{jk})^m}$ 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 of each cluster has to be specified implicitly in advance by the value ρ_i, whereas in the GG the sizes the cluster need not be known in advance.

For our purpose to generate rules from data by fuzzy cluster analysis the FCM is too restrictive, since it concentrates on spherical clusters of approximately the same size. As we will see in the next section, the GK and the GG cause trouble, since they are looking for clusters in the form of arbitrary hyper–ellipsoids, whereas for rule induction it would be more advantageous to have axes–parallel hyper–ellipsoids. 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 (5) and (6). We obtain

$$c_\nu^{(i)} \;=\; \frac{\left(\rho_i \prod_{\alpha=1}^{p}\sum_{k=1}^{n}(u_{ik})^m(x_{k\alpha} - v_{i\alpha})^2\right)^{1/p}}{\sum_{k=1}^{n}(u_{ik})^m(x_{k\nu} - v_{i\nu})^2} \qquad \text{and} \qquad c_\nu^{(i)} \;=\; \frac{\sum_{k=1}^{n}(u_{ik})^m}{\sum_{k=1}^{n}(u_{ik})^m(x_{k\nu} - v_{i\nu})^2}$$

as the updating schemes of the modified versions of the GK and GG, respectively, where $c_\nu^{(i)}$ denotes the νth diagonal element of the diagonal matrix C_i and $x_{k\alpha}$ and $v_{i\alpha}$ are the αth coordinates of the vectors x_k and v_i (see [6] for details). Note that neither the inverse nor the determinant of a matrix has to be computed so that the corresponding algorithms are much simpler and faster than the original GK and GG.

3 Fuzzy Clusters and Fuzzy Rules

The principal idea of inducing classification rules based on fuzzy cluster analysis is the following. Each fuzzy cluster is assumed to be assigned to one class for classification. The membership grades of the data to the clusters determine the degree to which they can be classified as a member of the corresponding class. With a fuzzy cluster that is assigned to the class $\mathcal{C}$ we associate a linguistic classification rule in the following way. The fuzzy cluster i is projected into each single dimension leading to a fuzzy set on the real numbers. The correct, but computationally inefficient method of projecting a fuzzy cluster would lead to the fuzzy set

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

as the νth projection of cluster i. Therefore, we use an approximation of this fuzzy set by projecting only the data set and computing the convex hull of this (discrete) projected fuzzy set or approximating it by a trapezoidal or triangular membership function as for instance proposed in [8].

To these fuzzy sets we assign suitable linguistic labels like *approximately zero* or *positive small* etc. (for example, one could assign the linguistic label approximately x_0 where x_0 is the value where the fuzzy set has its maximal membership degree).

The premise of the corresponding classification rule is the conjunction of these linguistic labels, the conclusion the class to which the cluster is assigned. The problem that arises is that the premise of the classification rule in the form of a conjunction leads to the Cartesian product of the corresponding one–dimensional fuzzy sets as the description of (a part of) the corresponding class. Unfortunately, the Cartesian product of projections is in general larger than the original fuzzy set (cluster) so that we have to accept a certain loss of information. As mentioned above, the FCM is very limited according to the restriction to spherical clusters of approximately the same size. On the other hand, the hyper–ellipsoidal clusters produced by the GK and the GG can result in misleading rules, since the projection of a hyper–ellipsoid may cause a strong loss of information. Therefore, we developed the modifications of the GK and GG in the previous section that are looking for hyper–ellipsoidal clusters with varying sizes whose axes are parallel to the coordinate axes. The advantage of these fuzzy clustering techniques is the greater flexibility compared to the FCM, the small loss of information when projected compared to the GK and GG and a simpler computation than for the GK and GG, since matrix inversion can be avoided.

4 Unsupervised and Supervised Classification

Our algorithms can be used for two types of classification tasks. In the case that the classification is not known in advance, we apply our modified version of the GK or the GG to the data and find out the optimal number of clusters by applying a suitable validity measure as it was already proposed for the GG in [2]. The number of cluster then coincides with the number of classes. The classification rules are derived as described in section 3.

When the classification for the given data set is known, we start the fuzzy clustering algorithm with the number of clusters equal to the number of classes. To each cluster the class of the prototype or the class of the datum with the highest membership degree is assigned. We then determine for each cluster the rate of misclassifications. For each cluster in which the rate of misclassifications exceeds a given upper bound, a new prototype in the neighbourhood of the original prototype of the cluster is introduced. After that we apply the fuzzy clustering algorithm again with increased number of clusters, i.e. the original number of clusters plus the number of newly introduced prototypes. As initialization the result of the clustering with the lower number of clusters incorporating the additional prototypes is used. We iterate this procedure until the number of misclassifications is small enough for each cluster.

5 Conclusions

The advantage of fuzzy clustering over hard classification techniques is the information inherent in the membership degrees so that we are able to judge how well the system is able to classify a datum. Another advantage is the possible use of intermediate classes in the case of classes in the form of real numbers. In this way our technique can be used to construct a fuzzy controller from data, since a fuzzy controller can be viewed as an interpolation technique for vague inputs and outputs [5, 4].

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 Trans. Pattern Analysis and Machine Intelligence 11 (1989), 773–781.

[3] D.E. Gustafson, W.C. Kessel, Fuzzy Clustering with a Fuzzy Covariance Matrix, Proc. IEEE CDC, San Diego (1979), 761–766.

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

[5] F. Klawonn, R. Kruse, Fuzzy Control as Interpolation on the Basis of Equality Relations, Proc. 2nd IEEE International Conference on Fuzzy Systems 1993, IEEE, San Francisco (1993), 1125–1130.

[6] F. Klawonn, R. Kruse, Constructing a Fuzzy Controller from Data, (to appear in) Fuzzy Sets and Systems.

[7] R. Kruse, J. Gebhardt, F. Klawonn, Foundations of Fuzzy Systems, Wiley, Chichester (1994).

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