FEATURE RANKING BASED ON INTERCLASS SEPARABILITY FOR FUZZY CONTROL APPLICATION*

Domonkos Tikk[1,2,†] and Tamás D. Gedeon[2]

[1] Department of Telecommunication & Telematics
Budapest University of Technology and Economics
H-1117 Budapest, Pázmány sétány 1/d, Hungary
tikk@ttt.bme.hu

[2] School of Information Technology, Murdoch University
South Street, Murdoch, 6150 W.A., Australia
tgedeon@murdoch.edu.au

Abstract

This paper presents a modified feature ranking method on interclass separability based for fuzzy control application. Existing feature selection/ranking techniques are mostly suitable for classification problems. These techniques result in a ranking of the input feature or variables. Our modification exploits an arbitrary fuzzy clustering of the control output data. Using these output clusters similar feature selection methods can be used as for classification, where the membership in a class (or cluster) will no longer be crisp, but a fuzzy value determined by the clustering. We studied the proposed method through a comparative analysis.

1 INTRODUCTION

It is well-known that fuzzy systems have exponentional time and space complexity in terms of N, the number of variables [19]. The number of rules in a rule base increases exponentially with N. Thus the resulting model of the system is very large. In practice, if N exceeds the experimental limit of about 6 variables, the rule based system becomes intractable. Due to this fact, rule base reduction emerged as an important research field in the past decade including e.g. the topics of fuzzy rule interpolation methods (see e.g. [2, 14, 19, 23]), hierarchical reasoning techniques [18, 21], and other rule base reduction methods [3, 4, 5, 7].

If the design of the modelled system is based on input-output data samples, a possible method of rule base reduction is the omission of those variables which have no relevant effect on the output. In pattern recognition and classification such methods are called feature selection [9]. Henceforth, when we use the terms feature or variable in this paper, we refer to the same notion. In these contexts the output of sample data usually indicates from which among a finite number of classes or clusters, the actual sample belongs. Practically it means that the outputs are selected from a finite set of labels or, equivalently, from a closed range of natural numbers.

Feature selection methods are of two main types: feature selection and ranking methods. The methods of the former type determine which input features are relevant in the given model, whilst the ones of the latter type result in a rank of importance, and it can be decided how many to select from the head of the rank, e.g. by a trial-and-error procedure. In this paper we modify the interclass separability feature ranking method. (The origin of this method is attributed to [11], who first applied the interclass distance concept to feature selection and extraction problems. Therefore this method is also known as Fischer's interclass separability method.)

The methods used in classification problems need to be modified (or the rule base has to be preprocessed) in fuzzy control applications when the range of the output is theoretically continuous. (We remark that in practice the range is discrete due to the accuracy of computers' representation ability concerning real numbers, however, if one would scale the range to this accuracy using as many clusters as many represented real numbers exist in the given range, the problems again became intractable and computer system dependent.) Therefore we have to somehow group the outputs of the data in order to make use of feature selection methods.

An obvious way of grouping is clustering the output by some fuzzy clustering technique. A fuzzy clustering method divides the clustered space into various regions,

*This research was supported by the Australian Research Council, the Hungarian Ministry of Education (MKM) FKFP 0235/1997 and FKFP 0422/1997 and the National Science Research Fund (OTKA) 019671 and 030655.

†Corresponding author

called clusters, and determines a vector of membership degrees for each input, which contains the amount to which a particular input data belongs to every cluster. The optimal number of clusters is determined by means of an objective function. In our model we used fuzzy c-means clustering (FCMC) because it was at hand [6], but other fuzzy clustering methods are also suitable for this purpose (e.g. subtractive clustering [16]). We exploit only the property of FCMC which guarantees that membership degrees are normalized, i.e. $\sum_{i=1}^{n} \mu_{ij} = 1$ holds for all clusters $(i = 1, \ldots, C)$, and n is number of sample data.

Based on Fischer's feature ranking method we developed its counterpart for fuzzy control application. The main algorithm is described in Section 3. In Section 4 we analyse the proposed method on a few sample data sets.

2 FEATURE SELECTION METHODS IN FUZZY APPLICATION

Fuzzy modeling based on the clustering of output data was first proposed by Sugeno and Yasukawa [22]. For reducing the number of inputs they used the regularity criterion (RC) method [15]. RC creates a tree structure from the variables, where the nodes represent particular sets of the total variable set. The nodes are evaluated according to an objective function, and the evaluation process stops if a local minimum is found. We are also working on the automatic design of fuzzy modeling systems but we found the RC method unreliable: it is very sensitive to its parameters [24], therefore we decided to look for an alternative solution.

Another solution was proposed to solve this problem by Costa Branco *et al* [8]. They used the principal component analysis (PCA) method [17] for identifying the important variables in the fuzzy model of an electromechanical system. The PCA method transforms the input data matrix of (possibly highly) correlated variables in an orthogonal system, where the variables become uncorrelated. From the transformed system the variables having eigenvalues under a certain threshold can be omitted. However, by this transformation the meaning of the variables and hence the direct linguistic interpretability of the system is lost, which we consider one of the most important features of a fuzzy system.

In the field of fuzzy classification another group of feature selection algorithms has been applied successfully [1, 9, 20] which are based on the interclass separability criterion. Let us briefly describe this method.

Let us take a given input $\{\underline{x}_1, \ldots, \underline{x}_n\}$ and the corresponding output $\{y_1, \ldots, y_n\}$ data set. $\underline{x}_i$ $(j = 1, \ldots, n)$ are N-dimensional vectors, i.e. N is the number of variables, or features. Let the matrix $\mathbf{X}$ be formed by the vectors $\underline{x}_j$ $(j = 1, \ldots, n)$. The inputs should be classified into classes $\mathcal{C}_i$ $(i = 1, \ldots, C)$ which possess *a priori* probabilities P_i and the cardinality of the classes is $|\mathcal{C}_i| = n_i$. Let $\mathbf{X}' = \{\underline{x}'_1, \ldots, \underline{x}'_n\}$ be generated by a feature selection technique from $\mathbf{X}$, where $\underline{x}'_i$ are N'-dimensional vectors $N' < N$. $\mathbf{X}'$ is generated by deleting some $N - N'$ rows of $\mathbf{X}$. A criterion for ranking the features is defined as [10]

$$J(\mathbf{X}') = \frac{1}{2} \sum_{i=1}^{C} P_i \sum_{j=1}^{C} \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{\ell=1}^{n_j} d(\underline{x}'_{ik}, \underline{x}'_{j\ell}) \quad (1)$$

where $d(\underline{x}'_{ik}, \underline{x}'_{j\ell})$ is a distance metric, usually the Euclidean norm. P_i can be estimated by $\frac{n_i}{n}$, the frequency of the occurrence of class members.

$J(\mathbf{X}')$ can be expressed after some algebraic manipulation as [10]:

$$J(\mathbf{X}') = \mathrm{tr}(\mathbf{Q}_w) + \mathrm{tr}(\mathbf{Q}_b) \quad (2)$$

where $\mathbf{Q}_w$ is the within class, and $\mathbf{Q}_b$ is the between class scatter matrices (defined below). Intuitively, for the feature selection task we wish to maximize $\mathrm{tr}(\mathbf{Q}_b)$ and at the same time minimize $\mathrm{tr}(\mathbf{Q}_b)$. It can be obtained by maximizing (2), but then the effect of within class distance of samples is unchecked. Therefore, the magnitude of the criterion function (2) is not a good indicator of class separability. A more realistic criterion function to maximize is

$$J(\mathbf{X}') = \frac{\mathrm{tr}(\mathbf{Q}_b)}{\mathrm{tr}(\mathbf{Q}_w)} \quad (3)$$

For more details see [9, 10].

The modification of this technique for fuzzy control application is presented in the next section.

3 THE FEATURE RANKING ON FUZZY CLUSTERED OUTPUT (FRFCO) ALGORITHM

Consider again the data set defined in the previous section. For brevity let $\mathcal{F}$ denote the set of all features. Let us cluster the output space by the fuzzy c-means clustering algorithm with parameter m (usually 2). Let the optimal number of clusters be C. The membership degree μ_{ij} $(i = 1, \ldots, C;\ j = 1, \ldots, n)$ denotes the degree the data $(\underline{x}_j, y_j)$ belongs to cluster i. Note that $\sum_{i=1}^{n} \mu_{ij} = 1$. With the following algorithm we can rank the input variables in order to find and omit the irrelevant ones.

As it has been shown above, Fischer's interclass separability criterion is based on the *fuzzy between-class* (4) and the *fuzzy within-class* (6) *scatter matrices* that sum up to the *total fuzzy scatter matrix* (7). Scatter matrices are also known as covariance matrices. These matrices can be defined in our context as

$$\mathbf{Q}_b = \sum_{i=1}^{C}\sum_{j=1}^{n}\mu_{ij}^m(\underline{v}_i-\underline{v})(\underline{v}_i-\underline{v})^T \tag{4}$$

$$\mathbf{Q}_i = \frac{1}{\sum_{j=1}^n \mu_{ij}^m}\sum_{j=1}^{n}\mu_{ij}^m(\underline{x}_j-\underline{v}_i)(\underline{x}_j-\underline{v}_i)^T \tag{5}$$

$$\begin{aligned}\mathbf{Q}_w &= \sum_{i=1}^{C}\mathbf{Q}_i = \\ &= \sum_{i=1}^{C}\frac{1}{\sum_{j=1}^n \mu_{ij}^m}\sum_{j=1}^{n}\mu_{ij}^m(\underline{x}_j-\underline{v}_i)(\underline{x}_j-\underline{v}_i)^T\end{aligned} \tag{6}$$

$$\begin{aligned}\mathbf{Q}_t &= \mathbf{Q}_b+\mathbf{Q}_w = \\ &= \sum_{i=1}^{C}\frac{1}{\sum_{j=1}^n \mu_{ij}^m}\sum_{j=1}^{n}\mu_{ij}^m(\underline{x}_j-\underline{v})(\underline{x}_j-\underline{v})^T\end{aligned} \tag{7}$$

where

$$\underline{v}_i = \frac{1}{\sum_{j=1}^n \mu_{ij}^m}\sum_{j=1}^{n}\mu_{ij}^m\underline{x}_j \tag{8}$$

are the fuzzy centers of the ith cluster, and

$$\underline{v} = \frac{1}{n}\sum_{j=1}^{n}\underline{x}_j \tag{9}$$

are the averages of clustered data. Here we assume that matrices $\mathbf{Q}_i$ and $\mathbf{Q}_w$ are nonsingular.

The feature interclass separability criterion is a trade-off between $\mathbf{Q}_b$ and $\mathbf{Q}_w$ as described above.

The proposed feature ranking algorithm proceeds iteratively. In each iteration it determines the most important variable based on Fischer's interclass separability criterion as follows. Let us delete temporarily the variable f ($f \in \mathcal{F}$). Calculate the matrices $\mathbf{Q}_b$ and $\mathbf{Q}_w$ for the remaining data and determine

$$J_f = \det(\mathbf{Q}_b)/\det(\mathbf{Q}_w). \tag{10}$$

Repeat this procedure for all the variables in $\mathcal{F}$. The expression $J_F = \min_{1,\dots,|\mathcal{F}|} J_f$ is minimal when the deviation between $\mathbf{Q}_b$ and $\mathbf{Q}_w$ is the least, e.g. when the most important variable is omitted. Then omitting $F \in \mathcal{F}$ permanently we can restart the algorithm with the new feature set.

1. Let $\mathcal{F} := \{1,\dots,k\}$.
2. For all $f \in \mathcal{F}$
   (a) Let $\mathcal{F} := \mathcal{F} - \{f\}$ and also update matrix $\mathbf{X}$, and vectors $\underline{v}_i$ and $\underline{v}$ by deleting temporarily its fth column or element.
   (b) Calculate matrices $\mathbf{Q}_b(\mathbf{X})$, $\mathbf{Q}_w(\mathbf{X})$ and determine J_f.
3. Let $f' = \operatorname*{argmin}_{f\in\mathcal{F}} J_f$, i.e. where J_f attains its minimal value. Delete permanently the variable(s) f' from $\mathcal{F}$ the corresponding columns from $\mathbf{X}$, $\underline{v}_i$ and $\underline{v}$. Note that f' can contain more than one variable.
4. If $|\mathcal{F}| > 1$ then back to step 2, else stop.

The order of the deleted variables gives their rank of importance.

4 BRIEF COMPARATIVE ANALYSIS OF FRFCO

We applied the proposed method to several data sets. First, we checked whether it gives consistent result on the synthetic and real data sets given in [22].

The first data set had 4 inputs. The first two were obtained from the function

$$y = f(x_1,x_2) = (1 + x_1^{-2} + x_2^{-1.5}), \quad 1 \le x_1, x_2 \le 5$$

and the last two were chosen randomly. The optimal number of clusters is 6. On this synthetic sample our proposed method gives the correct ranking: $\langle 2,1,3,4\rangle$.

The second sample data set is the model of a chemical plant with 5 inputs. In [22] the first three variables were found important. According to input contribution measures (similar to sensitivity analysis) [12] the third variable is the most important, then the first, while the remaining three were considered irrelevant. Our method gives the following ranking: $\langle 3,1,5,2,4\rangle$, again with 6 output clusters. This ranking does not change if $C = 5$.

Table 1: Ranking of the 12 input features of eye-gaze data set

method	Ranking
best of [13]	$\langle 6,4,3,2,7,5,1,10,9,8,12,11\rangle$
FRFCO with original output	$\langle 6,5,2,3,9,7,4,1,12,10,8,11\rangle$
FRFCO with network output	$\langle 5,4,6,9,8,1,11,2,3,10,12,7\rangle$

We also applied our technique to an eye-gaze data set which was analyzed thoroughly in [13], where the author analysed trained back propagation neural networks for ranking the 12 input features. We used the eye-gaze data set with its original outputs, but also with the network output of the best trained network. The results can be found in Table 1.

As it can be seen, our method on the original set is quite consistent with the one of [13]: it evaluates variable 6 as the best, and in there is just one difference in the first six most significant. Naturally, the evaluation with the network output the ranking gives a different result. Nevertheless, the six most significant variables coincide with Gedeon's result in 3 places, and the three most significant variables in 2 places.

5 CONCLUSION

We proposed in this paper a feature ranking algorithm adopted for fuzzy control application. The main idea is to cluster the output data and to use the cluster-membership degree as weights in the feature selection method. We applied our method to real world and synthetic data sets, and it was likely to find the proper or close-to-proper ranking.

REFERENCES

[1] S. Abe, R. Thawonmas, and Y. Kobayashi. Feature selection by analyzing class regions approximated by ellipsoids. *IEEE Trans. on SMC, Part C*, 28(2), 1998. http://www.info.kochi-tech.ac.jp/ruck/paper.html.

[2] P. Baranyi and L. T. Kóczy. A general and specialized solid cutting method for fuzzy rule interpolation. *BUSEFAL*, 67:13–22, 1996.

[3] P. Baranyi, A. Martinovics, D. Tikk, L. T. Kóczy, and Y. Yam. A general extension of fuzzy SVD rule base reduction using arbitrary inference algorithm. In *Proc. of IEEE Int. Conf. on System Man and Cybernetics (IEEE-SMC'98)*, pages 2785–2790, San Diego, USA, 1998.

[4] P. Baranyi and Y. Yam. Fuzzy rule base reduction. In D. Ruan and E. E. Kerre, editors, *Fuzzy IF-THEN Rules in Computational Intelligence: Theory and Applications*, pages 135–160. Kluwer, 2000.

[5] P. Baranyi, Y. Yam, C. T. Yang, P. Várlaki, and P. Michelberger. Generalised SVD fuzzy rule base complexity reduction. to appear in *International Journal of Advanced Computational Intelligence*.

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

[7] J. Bruinzeel, V. Lacrose, A. Titli, and H. B. Verbruggen. Real time fuzzy control of complex systems using rule-base reduction methods. In *Proc. of the 2nd World Automation Congress (WAC'96)*, Montpellier, France, 1996.

[8] P. J. Costa Branco, N. Lori, and J. A. Dente. New approaches on structure identification of fuzzy models: Case study in an electro-mechanical system. In T. Furuhashi and Y. Uchikawa, editors, *Fuzzy Logic, Neural Networks, and Evolutionary Computation*, pages 104–143. Springer-Verlag, Berlin, 1996.

[9] R. K. De, N. R. Pal, and S. K. Pal. Feature analysis: Neural network and fuzzy set theoretic approaches. *Pattern Recognition*, 30(10):1579–1590, 1997.

[10] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, London, 1982.

[11] R. A. Fischer. The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, 7:179–188, 1936.

[12] C. C. Fung, K. W. Wong, and H. Crocker. Determining input contributions for a neural network based porosity prediction model. In *Proc. of the Eighth Australian Conference on Neural Network (ACNN97)*, pages 35–39, Melbourne, 1997.

[13] T. D. Gedeon. Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems*, 8(2):209–218, 1997.

[14] T. D. Gedeon and L. T. Kóczy. Conservation of fuzziness in rule interpolation. In *Proc. of the Symp. on New Trends in Control of Large Scale Systems*, volume 1, pages 13–19, Herľany, 1996.

[15] J. Ihara. Group method of data handling towards a modeling of complex system IV. *Systems and Control*, 24:158–168, 1980. (In Japanese).

[16] J-S. R. Jang, C.-T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Prentice Hall, Upper Saddle River, NJ, 1997.

[17] Kleinbaum, Kupper, and Muller. *Applied Regression Analysis and Other Multivariable Methods*. PWS-Kent, Boston, Mass., 2nd edition, 1988.

[18] L. T. Kóczy and K. Hirota. Approximate inference in hierarchical structured rule bases. In *Proc. of 5th IFSA World Congress (IFSA'93)*, pages 1262–1265, Seoul, 1993.

[19] L. T. Kóczy and K. Hirota. Size reduction by interpolation in fuzzy rule bases. *IEEE Trans. on SMC*, 27:14–25, 1997.

[20] J.A. Roubos, M. Setnes, and J. Abonyi. Learning fuzzy classification rules from labeled data. *International Journal of Information Sciences*, submitted, July 2000. http://www.fmt.vein.hu/softcomp.

[21] M. Sugeno, M. F. Griffin, and A. Bastian. Fuzzy hierarchical control of an unmanned helicopter. In *Proc. of the 5th IFSA World Congress (IFSA'93)*, pages 1262–1265, Seoul, 1993.

[22] M. Sugeno and T. Yasukawa. A fuzzy logic based approach to qualitative modelling. *IEEE Trans. on FS*, 1(1):7–31, 1993.

[23] D. Tikk and P. Baranyi. Comprehensive analysis of a new interpolation method. To appear in *IEEE Trans. on Fuzzy Systems*.

[24] D. Tikk, T. D. Gedeon, L. T. Kóczy, and Gy. Biró. Improvements and critique on Sugeno and Yasukawa's qualitative modelling. In preparation.