Proceedings of KFIS Fall Conference, 2003. 5. 3

An Approach to Combining Classifier with MIMO Fuzzy Model

Do Wan Kim[1], Jin Bae Park[1], Yeon Woo Lee[2], Young Hoon Joo[2]

[1] Department of Electrical and Electronic Engineering, Yonsei University, Seodaemun-gu, Seoul, 120-749, Korea.
[2] School of Electronic and Information Engineering, Kunsan National University, Kunsan, Chonbuk, 573-701, Korea.

ABSTRACT

This paper presents a new design algorithm for the combination with the fuzzy classifier and the Bayesian classifier. Only few attempts have so far been made at providing an effective design algorithm combining the advantages and removing the disadvantages of two classifiers. Specifically, the suggested algorithms are composed of three steps: the combining, the fuzzy-set-based pruning, and the fuzzy set tuning. In the combining, the multi-inputs and multi-outputs (MIMO) fuzzy model is used to combine two classifiers. In the fuzzy-set-based pruning, to effectively decrease the complexity of the fuzzy-Bayesian classifier and the risk of the overfitting, the analysis method of the fuzzy set and the recursive pruning method are proposesd. In the fuzzy set tuning for the misclassified feature vectors, the premise parameters are adjusted by using the gradient decent algorithm. Finally, to show the feasibility and the validity of the proposed algorithm, a computer simulation is provided.

Keywords: Combining classifier, fuzzy-Bayesian classifier, tuning, pruning, fuzzy classifier, Bayesian classifier.

1. Introduction

In recent years, the conventional classifiers are needed to the classification capability for the highly complex real data such as the sensory data from the multiple sensors and the signal data. To correctly classify these highly complex data, a new classifier, which is able to provide the classification performance above it of the conventional classifiers, is indispensably required. The combining classifier is useful to resolve this problem. The combining classifier may generate more accurate classification than each of the constituent classifiers [6]. Particularly, the fuzzy-rule-based structure is proposed for combining the fuzzy classifier and the Bayesian classifier in [7,8]. However, these researches do not show the design algorithms combining the merits and removing the demerits of each component classifiers.

In this paper, the design algorithm of the combining classifier, *i.e.*, fuzzy-Bayesian classifier, is proposed for preserving the advantages and overcoming the limitations of the fuzzy classifier and the Bayesian classifier. The structure of the suggested combining classifier is the multi-inputs and multi-outputs (MIMO) fuzzy model, which can be implemented by the linguistic form and the discriminant function, and contains two component classifiers by using the fuzzy sets in the antecedent part of it and the Bayesian classifier in the consequent part of it. Moreover, through the suggested design algorithms, the fuzzy-bayesian classifier can be simply designed, reduced the dimensionality, and tuned.

2. Preliminaries

2.1 Bayesian Classifier

Using the prior probabilities $P(\mathbf{x})$ and the conditional densities $P(\mathbf{x}|\mathcal{C}_i)$, especially, the multivariate Gaussian model, the Bayesian classifier is designed by the following discriminant function:

$$d_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\Sigma_i^{-1}(\mathbf{x}-\mathbf{m}_i)} P(\mathcal{C}_i) \quad (1)$$

where $\mathbf{x}$ is a n-component column vector, *i.e.*, a feature vector $(x_1,\ldots,x_n)^T$, $\mathbf{m}_i$ is the n-component mean vector, Σ_i is the $n \times n$ covariance matrix, $\mathcal{C}_i, i \in \mathcal{I}_m = \{1,\ldots,m\}$, is the ith class, and $|\Sigma_i|$ and Σ_i^{-1} are its determinant and inverse, respectively. The Bayesian classifier is said to assign a feature vector $\mathbf{x}$ to class $\mathcal{C}_{i_1}, i_1 \in \mathcal{I}_m$, if

$$d_{i_1}(\mathbf{x}) > d_{i_2}(\mathbf{x}), \quad \forall i_2 \neq i_1, \quad i_2 \in \mathcal{I}_m \quad (2)$$

Specifically, the effect of d_{i_1} and d_{i_2} is to divide the feature space into two decision regions $\mathcal{R}_{i_1}$ and $\mathcal{R}_{i_2}$.

Due to the robust performance and simple implementation, the Bayesian classifier has become a popular classification tool in recent year [3].

Remark 1 *However, the Bayesian classifier has the following drawbacks:*

(i) It is difficult to determine and compute $P(\mathbf{x}|\mathcal{C}_i)$. Specifically, in the design of Bayesian classifiers, particularly in the design of Gaussian normal classifiers, a frequently made assumption about the normal form of $P(\mathbf{x}|\mathcal{C}_i)$ governing of patterns is not necessarily true for real data.

(ii) In the Bayesian classifier, it is difficult to reduce the dimensionality.

2.2 Fuzzy Classifier

There are various fuzzy model to do pattern classification, but the most popular and general [5] is

$$R_i : \text{IF } x_1 \text{ is } A_{i1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{in}, \quad \text{THEN the class is } i, \tag{3}$$

where $R_i, i \in \mathcal{I}_m$, denotes the ith fuzzy rule, $x_j, j \in \mathcal{I}_n = \{1, 2, \ldots, n\}$, is the jth feature variable, and $A_{ij}, (i,j) \in \mathcal{I}_m \times \mathcal{I}_n$, is the fuzzy set. The conjunction rule to transform the fuzzy sets into a discriminant function is

$$w_i = \mu_{A_{i1}} \times \cdots \times \mu_{A_{in}} \tag{4}$$

where $\mu_{A_{ij}}$ is the membership function. The effect of w_i is to divide the feature space $\mathbb{R}^n$ into the decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_m$. Therefore, the fuzzy classifier is said to assign a feature vector $\mathbf{x}$ to class $\mathcal{C}_{i_1}, i_1 \in \mathcal{I}_m$, if

$$w_{i_1} > w_{i_2}, \qquad \forall i_2 \neq i_1, \quad i_2 \in \mathcal{I}_m \tag{5}$$

The fuzzy approaches to pattern recognition have the following advantages: It would lie in guiding the steps by which one takes knowledge in a linguistic form and casts it into discriminant functions [4]. Generally, in the majority of classifiers, it is difficult to analyze the decision region $\mathcal{R}$ for the given high dimensional feature space $\mathbb{R}^n$. In the fuzzy classifier, however, because $A_{ij} \in [0,1]$ is able to give the information of the ith decision region $\mathcal{R}_i$ in the aspect of the jth one-dimensional feature variable, the analysis of $\mathcal{R}$ is more easy. This easy analysis of $\mathcal{R}$ helps to tune and prune A_{ij} such as [5].

Remark 2 *In spite of the above-mentioned advantages, the fuzzy classifier has the following limitations:*

(i) Fuzzy methods are cumbersome to use in high dimensions or on complex problems or in problems with dozens or hundreds of features [4].

(ii) The amount of information the designer can be expected to bring to a problem is quite limited—the number, positions, and widths of membership function [4].

3. A Novel Approach to Fuzzy-Bayesian Classifier

3.1 Combining

To perform the pattern classification for a given feature vector $\{\mathbf{x}, \mathcal{C}_i\}, i \in \mathcal{I}_m$, the MIMO fuzzy model is designed by

$$R_i : \text{IF } x_1 \text{ is } A_{i1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{in}, \quad \text{THEN } \mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{im} \end{bmatrix} \tag{6}$$

where $R_i, i \in \mathcal{I}_m$, denotes the ith fuzzy rule, $x_j, j \in \mathcal{I}_n$, is the jth feature, $A_{ij}, (i,j) \in \mathcal{I}_m \times \mathcal{I}_n$, is the fuzzy set, and $\mathbf{y}_i$ is the output vector of R_i.

To reduce the design efforts of the fuzzy sets in the problems with dozen or hundreds feature, A_{ij} in the premise part of R_i is simply identified by the mean m_{ij} and the standard deviation σ_{ij} of the jth feature values labelled as $\mathcal{C}_i$. A_{ij} is characterized by the following Gaussian membership function $\mu_{A_{ij}}$.

$$\mu_{A_{ij}} = e^{\frac{-(x_j - m_{ij})^2}{2\sigma_{ij}{}^2}} \tag{7}$$

In the consequent part identification of R_i, the Bayesian classifier (1) is used. It divides the feature space into m distinct decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_m$.

Therefore, by applying (1), $\mathbf{y}_i$ has the following form:

$$\mathbf{y}_i = \begin{bmatrix} d_1(\mathbf{x}) - d_i(\mathbf{x}) \\ \vdots \\ d_i(\mathbf{x}) \\ \vdots \\ d_m(\mathbf{x}) - d_i(\mathbf{x}) \end{bmatrix} \tag{8}$$

In the same manner, when the overall consequent parts of $R_1, \ldots, R_m$ are represented as $\mathbb{Y}_{m \times m} = [\, \mathbf{y}_1 \ \cdots \ \mathbf{y}_i \ \cdots \ \mathbf{y}_m \,]$, they are identified by

$$\mathbb{Y}_{m \times m} = \begin{bmatrix} d_1(\mathbf{x}) & \cdots & \star & \cdots & \star \\ \vdots & & \vdots & & \vdots \\ d_i(\mathbf{x}) - d_1(\mathbf{x}) & \cdots & d_i(\mathbf{x}) & \cdots & \star \\ \vdots & & \vdots & & \vdots \\ d_m(\mathbf{x}) - d_1(\mathbf{x}) & \cdots & d_m(\mathbf{x}) - d_i(\mathbf{x}) & \cdots & d_m(\mathbf{x}) \end{bmatrix} \tag{9}$$

where '$\star$' denotes the element in skew-symmetric positions.

For the given the feature vector $\mathbf{x}$, the final outputs of (6) are inferred as follows:

$$\widehat{\mathbf{y}} = \frac{\sum_{i=1}^{m} w_i \mathbf{y}_i}{\sum_{i=1}^{m} w_i} \tag{10}$$

where

$$w_i = \prod_{j=1}^{n} \mu_{A_{ij}}(x_j), \qquad \widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_i \\ \vdots \\ \widehat{y}_m \end{bmatrix}$$

From (10), the fuzzy-Bayesian classifier is said to assign a feature vector $\mathbf{x}$ to class $\widehat{y}_{i_1}, i_1 \in \mathcal{I}_m$, if

$$\widehat{y}_{i_1} > \widehat{y}_{i_2}, \quad \forall i_2 \neq i_1, \quad i_2 \in \mathcal{I}_m \tag{11}$$

The following therorem shows that, in case of $w_{i_1} > w_{i_2}$ and $d_{i_1} > d_{i_2}$ for all $i_2 \neq i_1, i_1, i_2 \in \mathcal{I}_m$, the final outputs of the fuzzy-Bayesian classifier holds $\widehat{y}_{i_1} > \widehat{y}_{i_2}$.

3.2 Fuzzy-Set-Based Pruning

Generally, the goals of the pruning are to avoid the overfitting and to reduce the complexity of the classifier. In this paper, these goals are accomplished by pruning the feature variables of the fuzzy rules. To do this, the analysis of $\mathcal{R}$ is needed.

However, it is difficult to appropriately select the feature variable to be pruned. Especially, in the Bayesian classifier, the analysis of $\mathcal{R}$ for the high dimensional feature space is nearly impossible. Compared with the Bayesian classifier, the conventional fuzzy classifier has the feasibility of the approximate analysis of $\mathcal{R}$ owing to the fuzzy sets. Specifically, A_{ij} is able to give the information of the approximate $\mathcal{R}_i$ in the aspect to the jth feature variable. Moreover, because the fuzzy sets of the proposed fuzzy-Bayesian classifier are identified by stochastic information, the analysis of the fuzzy set is also help to select the feature variable of the Bayesian classifier.

In these aspects, we newly define that a *correctness method* is

$$C(A_{ij}) = \frac{|A_{ij}|_{x_{ij}\in\mathcal{R}_{ij}}}{|A_{ij}|_{x_{ij}}} \tag{12}$$

where $|\cdot|$ denotes the cardinality of a set, x_{ij} denotes x_j labelled as $\mathcal{C}_i$ and $\mathcal{R}_{ij}$ is approximate $\mathcal{R}_i$ in aspect of A_{ij}. For example, $\mathcal{R}_{i_1 j}, i_1 \in \mathcal{I}_m$, is the region corresponding to $\mu_{A_{i_1 j}} > \mu_{A_{i_2 j}}$ for all $i_2 \neq i_1, i_2 \in \mathcal{I}_m$. By applying the definition of the cardinality, the correctness method (12) is

$$C(A_{ij}) = \frac{\sum_{x_{ij}\in\mathcal{R}_{ij}} \mu_{A_{ij}}}{\sum_{x_{ij}} \mu_{A_{ij}}} \tag{13}$$

When A_{ij} is not overlapped with the others and/or all x_{ij} fall on $\mathcal{R}_{ij}$, $C(A_{ij}) = 1$. When A_{ij} is completely overlapped with the others and/or no x_{ij} falls on $\mathcal{R}_{ij}$, $C(A_{ij}) = 0$.

The following recursive method suggests effective ways to prune the feature variable in fuzzy rule.

Step 1 Define the candidate sets $\tilde{\mathbf{x}}_i$ as $\{x_1, \ldots, x_n\}$ and the pruned sets $\bar{\mathbf{x}}_i$ as $\emptyset$.

Step 2 At iteration l, $l = 0, 1, 2, \ldots$, compute the recognition rate of the fuzzy classifier $a^{(l)}$.

Step 3 At iteration p, $p = 1, \ldots, m$, determine R_p to be pruned, where $R_1, \ldots, R_m$ are sorted by $\frac{1}{n}\sum_{j=1}^{n} C(A_{1j})^{(0)}, \ldots, \frac{1}{n}\sum_{j=1}^{n} C(A_{mj})^{(0)}$ in the descending order. If the number of the feature variable in R_p is one, $p = p + 1$ and then determine R_p.

Step 4 Sort the components of $\tilde{\mathbf{x}}_p$ by $\exists j, C(A_{pj})^{(l)}$ in the ascending order. Let n_p as the column size of $\tilde{\mathbf{x}}_p$.

Step 5 At iteration q, $q = 1, \ldots, n_p$, update $\bar{\mathbf{x}}_1^{(l+1)}, \ldots, \bar{\mathbf{x}}_m^{(l+1)}$ by using $\tilde{\mathbf{x}}_p(q) \in \bar{\mathbf{x}}_p^{(l+1)}$. Simultaneously prune the feature variables in premise and consequent parts of $R_1, \ldots, R_m$ corresponding to $\bar{\mathbf{x}}_1^{(l+1)}, \ldots, \bar{\mathbf{x}}_m^{(l+1)}$ and then compute $a^{(l+1)}$ and $\forall i, \exists j, C(A_{ij})^{(l+1)}$.

Step 6 If $\left\|a^{(l+1)} - a^{(l)}\right\| \geq 0$, then let $l = l + 1$.

Step 7 If $q < n_p$, set $q = q + 1$ and then go to Step 5; otherwise, go to Step 8.

Step 8 If $\tilde{\mathbf{x}}_p \cap \bar{\mathbf{x}}_p^{(l)} = \emptyset$, stop; otherwise, let $p = \begin{cases} p+1, & \text{if } p < m; \\ 1, & \text{if } p = m, \end{cases}$ and then go to Step 3.

3.3 Fuzzy Set Tuning for the Misclassified Feature Vectors

Because of the limitation about the normal form of $P(\mathbf{x}|\mathcal{C}_i)$, the miscalssified feature vectors occur. To resolve this problem, the appropriate compensation for the Bayesian classifier is required. In this aspect, m_{ij} and σ_{ij} of the fuzzy sets are adjusted for correctly categorizing the misclassified feature vectors.

For the misclassified feature vector $(\mathbf{x}; \mathcal{C}_{i_1})$, σ_{ij} and m_{ij} are tuned so as to satisfy the following inequality:

$$i_1, i_2 \in \mathcal{I}_m, \quad \forall i_2 \neq i_1, \qquad \widehat{y}_{i_1} > \widehat{y}_{i_2} \tag{14}$$

Transform the inequality form (14) into the following equality form:

$$\widehat{y}_{i_1} \approx \max_{i_2\in\mathcal{I}_m, i_2\neq i_1} \widehat{y}_{i_2} + \epsilon \tag{15}$$

where ϵ is a small positive scalar. From (15), define the objective function as

$$\textit{Minimize} \quad J = \frac{(\max_{i_2\in\mathcal{I}_m, i_2\neq i_1} \widehat{y}_{i_2} + \epsilon - \widehat{y}_{i_1})^2}{2} \tag{16}$$

To precisely adjust the premise parameters σ_{ij}, m_{ij} of R_i, we can use such classic iterative optimization procedures as the decent gradient algorithm.

$$\begin{aligned}\Delta\sigma_{ij} = &-\alpha(\widehat{y}_{i'_2} + \epsilon - \widehat{y}_{i_1})(y_{ii'_2} - \widehat{y}_{i'_2} - y_{ii_1} + \widehat{y}_{i_1}) \\ &\times \frac{1}{\sum_{i=1}^{m} w_i} \frac{2}{\sigma_{ij}} \left(\frac{x_j - m_{ij}}{\sqrt{2}\sigma_{ij}}\right)^2 \prod_{j=1}^{n} A_{ij}\end{aligned} \tag{17}$$

$$\begin{aligned}\Delta m_{ij} = &-\beta(\widehat{y}_{i'_2} + \epsilon - \widehat{y}_{i_1})(y_{ii'_2} - \widehat{y}_{i'_2} - y_{ii_1} + \widehat{y}_{i_1}) \\ &\times \frac{1}{\sum_{i=1}^{m} w_i} \frac{2}{\sqrt{2}\sigma_{ij}} \frac{x_j - m_{ij}}{\sqrt{2}\sigma_{ij}} \prod_{j=1}^{n} A_{ij}\end{aligned} \tag{18}$$

where α and β are the learning rates for σ_{ij} and m_{ij}, respectively, and $\widehat{y}_{i'_2}$ is $\max_{i_2\in\mathcal{I}_m, i_2\neq i_1} \widehat{y}_{i_2}$.

4. Computer Simulation: Wisconsin Breast Cancer Diagnostic Data

The Wisconsin breast cancer diagnostic data [9] consists of 699 feature vectors; 458 feature vectors belong

Table 1: Classification Results on Wisconsin Breast Cancer Diagnostic Data

Design procedure	Avg. number of premise fuzzy sets	Avg. training recognition rate	Avg. testing recognition rate
Combining	18	96.37%	95.82%
Fuzzy-set-based pruning	9.46	97.57%	96.07%
Fuzzy set tuning	9.46	97.68%	96.10%

Table 2: Comparison of classification results

Ref.	Avg. testing recognition rate
[10]	95.14%
[11]	95.7%
[12]	95.6%
Ours	96.10%

to the "benign" class, and the other 241 feature vectors are the "malignant" class. In this data set, the breast cancer is to be diagnosed on the basis of nine features. Because of 16 missing values in this data set, we use 683 feature vector to evaluated the proposed classifier. One half of 683 feature vectors are randomly selected as the training data and the other half are used as the testing data. The data set is normalized to $[0, 1]$.

Table 1 shows the simulation result of the fuzzy-Bayesian classifier, which is composed of R_1 and R_2, for each step of the design algorithm. Although the average number of the fuzzy sets reduces from 18 to 9.46, both of the average testing and training recognition rate increase from 96.37% to 97.68% and from 95.82% to 96.10%, respectively. That definitely shows that the proposed design algorithm effectively provides the robustness for the overfitting and the decline of the dimensionality. Moreover, the classification performance of the fuzzy-Bayesian classifier is better than other classifiers as shown in Table 2.

5. Conclusions

In this paper, a novel fuzzy-Bayesian classifier design algorithm has been proposed for combining advantages and overcoming the limitations of the fuzzy classifier and the Bayesian classifier, and its validity, robustness for overfitting, and applicability are verified through the computer simulation.

감사의 글: 본 연구는 정보통신부 정보통신연구진흥원에서 지원하고있는 대학기초연구지원사업 (과제번호: 2001-107-3)에 의해 지원받았습니다.

Reference

[1] Y. H. Joo, H. S. Hwang, K. B. Kim, and K. B. Woo, "Linguistic model identification for fuzzy system," *Electron. Letter*, Vol. 31, pp. 330–331, 1995.

[2] Y. H. Joo, H. S. Hwang, K. B. Kim, and K. B. Woo, "Fuzzy system modeling by fuzzy partition and GA hybrid schemes," *Fuzzy Set and Syst.*, Vol. 86, pp. 279–288, 1997.

[3] H. Huang and C. Hsu, "Bayesian classification for data from the same unknown class," *IEEE Trans. Syst. Man, Cybern. B.*, vol. 32, pp. 137–145, 2002.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification,* A wiley-interscience publishing company, inc., 2001.

[5] T. P. Wu and S. M. Chen, "A new method for constructing membership functions and fuzzy rules from training examples," *IEEE Trans. Syst. Man, Cybern. B.*, vol. 29, pp. 25–40, 1999.

[6] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, "Decision templates for multiple classier fusion: an experimental comparison," *Pattern Recognition*, vol. 34, pp. 299–314, 2001.

[7] L. I. Kuncheva, "How good are fuzzy if-then classifiers?," *IEEE Trans. Syst. Man, Cybern. B.*, vol. 30, pp. 501–509, 2000.

[8] J. van den Berg, U. Kaymak, and W. M. van den Bergh, "Fuzzy classification using probability-based rule weighting," in *proc. IEEE Int. conf. Fuzzy systems,* vol. 2, pp. 991–996, 2002.

[9] C. J. Merz and P. M. Murphy, "UCI Repository of Machine Learning Databases," http://www.ics.uci.edu/ mlearn/MLRepository.html, Irvine, Dept. of Information and Computer Science, Univ. of California, Irvine, 1996.

[10] H. M. Lee, C. M. Chen, J. M. Chen, and Y. L. Jou, "An efficient fuzzy classifier with feature selesction based on fuzzy entropy," *IEEE Trans. Syst. Man, Cybern. B.*, vol. 3, pp. 426–432, 2001.

[11] H. M. Lee, C. M. Chen, and Y. F. Lu, "A self-organizing HCMAC neural-network classifier," *IEEE Trans. Neural Networks*, vol. 14, pp. 15–27, 2003.

[12] M. Muselli and D. Liberati, "Binary rule generation via hamming clustering," *IEEE Trans. Knowledge and Data Eng.*, vol. 14, pp. 1258–1268, 2002.