A Novel Approach to Feature Selection Based on Analysis of Class Regions

Ruck Thawonmas, *Member, IEEE*, and Shigeo Abe, *Senior Member, IEEE*

***Abstract*— This paper presents a novel approach to feature selection based on analysis of class regions which are generated by a fuzzy classifier. A measure for feature evaluation is proposed and is defined as the exception ratio. The exception ratio represents the degree of overlaps in the class regions, in other words, the degree of having exceptions inside of fuzzy rules generated by the fuzzy classifier. It is shown that for a given set of features, a subset of features that has the lowest sum of the exception ratios has the tendency to contain the most relevant features, compared to the other subsets with the same number of features. An algorithm is then proposed that performs elimination of irrelevant features. Given a set of remaining features, the algorithm eliminates the next feature, the elimination of which minimizes the sum of the exception ratios. Next, a terminating criterion is given. Based on this criterion, the proposed algorithm terminates when a significant increase in the sum of the exception ratios occurs due to the next elimination. Experiments show that the proposed algorithm performs well in eliminating irrelevant features while constraining the increase in recognition error rates for unknown data of the classifiers in use.**

I. Introduction

IN PATTERN recognition, feature reduction has long been an important topic and has been studied by many authors because of its impact on the complexity of classifiers. It is also known that a good feature reduction method must have the ability to constrain the increase in recognition error rates for unknown data of the classifiers in use, due to the reduction in dimensionality.

There are two different approaches to achieve feature reduction: feature extraction and feature selection. In the feature extraction approach, all of the original features are mapped into a lower-dimensional feature space. Principal component analysis (PCA) [1] (or the Karhunen–Loeve transform in signal processing) performs a linear transformation of an input feature vector. The first component of the transformed feature vector represents the component of the original input feature vector in the direction of its largest eigenvector of the feature covariance matrix, the second component of the transformed feature vector in the direction of the second largest, and so on. In [2], this technique is applied so that training of a neural net classifier is initiated in the direction of the major eigenvectors of the covariance matrix of training patterns. Discriminant analysis is another technique discussed in [1], which finds the set of transformed features that gives the greatest class separation. In [3] and [4], class regions are analyzed directly to retain informative features and to eliminate redundant features.

In the feature selection approach, relevant features are selected from the original features. In [5], various well-known measures, such as Bhattacharyya probabilistic distance, are given for selecting the set of features that maximizes class separability. In [6], a feature-selection algorithm is proposed that exploits some fuzzy parameters, which represent fuzziness in a set, to measure class separability. In [7], features are selected based on the mutual information criterion.

Compared with the feature selection approach, the feature extraction approach has a higher degree of freedom in finding a set of features, especially when the best set in terms of classification cannot be selected directly from the original features. However, the feature selection approach does have some advantages over its counterpart, as elaborated in the following. After a set of features is selected, nonselected features will no longer be used. To collect new data, only collection of the selected features is necessary, which may reduce costs. Furthermore, the physical meaning of each selected feature is retained. For some rule based classifiers, a characteristic that the features are perceivable by human experts is indispensable.

Motivated by the above considerations, we adopt the feature selection approach in this paper. We propose an algorithm to eliminate irrelevant features. The proposed algorithm is based on analysis of class regions which are generated by a fuzzy classifier [8]. The degree of overlaps in the class regions, or the degree of having exceptions inside of fuzzy rules generated by the fuzzy classifier, is defined as the exception ratio and is used as a measure for feature evaluation. The idea of using the exception ratio for feature evaluation derives from the fact that for a given set of features, a subset of features that has the lowest sum of the exception ratios has the tendency to contain the most relevant features, compared to the other subsets with the same number of features. Given a set of remaining features, the proposed algorithm eliminates the next feature, the elimination of which minimizes the sum of the exception ratios. Next, a terminating criterion is proposed. Based on this criterion, the proposed algorithm terminates when additional elimination of a single feature results in a significant increase in the sum of the exception ratios, which implies a remarkable rise in recognition error rates for unknown data of the classifiers in use.

In the following section, we review the fuzzy rule representation and inference scheme of the fuzzy classifier in [8]. We then propose the exception ratio based feature elim-

Manuscript received April 11, 1995; revised December 23, 1995.

R. Thawonmas is with the Laboratory for Artificial Brain Systems, Institute of Physicial and Chemical Research (RIKEN), Saitama-ken 351–01, Japan.

S. Abe is with Hitachi Research Laboratory, Hitachi, Ltd., Ibaraki-ken 319-12, Japan.

Publisher Item Identifier S 1083-4419(97)00139-8.

ination algorithm in Section III. Finally, we demonstrate in Section IV the effectiveness of the proposed algorithm using two classifiers on four classification problems from different domains.

II. Fuzzy Rule Representation and Inference Scheme

A. Fuzzy Rule Representation

In this section, we briefly describe the method for generating fuzzy rules discussed in [8]. The basic idea behind the method is as follows. First, class regions in the input space are approximated by means of hyperboxes. If there exists an overlap between the regions of any two classes, the method attempts to resolve the overlapping region in a recursive fashion. Finally, fuzzy rules are defined for all generated regions.

Suppose we have a training data set consisting of input–output pairs. Let X_i denote the set of input data for class i, where $i = 1, \cdots, n$. We generate fuzzy rules by which a given m-dimensional input vector $\mathbf{x}$ can be classified into one of n classes as shown in Fig. 1, where $j' = i$ and $i' = j$ for $l = 1$, $j' = j$, and $i' = i$ for $l \geq 2$. To do this, for each class, say, class i, we first define an activation hyperbox of level 1, denoted as $A_{ii}(1)$, by finding the minimum and maximum values of each input variable from X_i. Then, if $A_{ii}(1)$ and $A_{jj}(1)$ overlap, the overlapping region is defined as the inhibition hyperbox of level 1, denoted as $I_{ij}(1)$. To increase recognition rates for unknown data of the classifiers in use, expansion of $I_{ij}(1)$ in $A_{ii}(1)$ and $A_{jj}(1)$ is next performed, resulting in expanded inhibition hyperboxes $J_{ij}(1)$ and $J_{ji}(1)$, respectively. If data of classes i and/or j exist in the corresponding expanded inhibition hyperbox, a second activation hyperbox will be defined and denoted as $A_{ij}(2)$ or $A_{ji}(2)$. Moreover, if these two activation hyperboxes still overlap with each other, an inhibition hyperbox $I_{ij}(2)$ and the associated expanded inhibition hyperboxes $J_{ij}(2)$ and $J_{ji}(2)$ will be defined. The recursion procedure terminates when either there is no overlap between $A_{ij'}(l)$ and $A_{i'j}(l)$ or the condition $A_{ij'}(l) = A_{i'j}(l) = I_{ij}(l-1)$ holds. In the latter case, the overlap cannot be resolved by the recursive procedure. Therefore, for each datum residing in $I_{ij}(l-1)$, an activation hyperbox is defined which includes only that datum.

Let $r_{ij}(l)$ denote a fuzzy rule for class i which is defined at level $l\,(\geq 1)$ by resolving overlaps with class j. Fuzzy rules without and with inhibition are, respectively, given by

$$\text{If } \mathbf{x} \text{ is in } A_{ij'}(l), \text{ then } \mathbf{x} \text{ belongs to class } i \tag{1}$$

$$\text{If } \mathbf{x} \text{ is in } A_{ij'}(l) \text{ and } \mathbf{x} \text{ is not in } J_{ij}(l), \text{ then } \mathbf{x} \text{ belongs to class } i \tag{2}$$

where $j' = i$ for $l = 1$ and $j' = j$ for $l \geq 2$.

Here, the activation hyperbox $A_{ij'}(l)$ is defined as

$$A_{ij'}(l) = \{\mathbf{x} |\, v_{ij'k}(l) \leq x_k \leq V_{ij'k}(l), \quad k = 1, \cdots, m\} \tag{3}$$

where x_k: the kth element of input vector $\mathbf{x}$,

$v_{ij'k}(l)$: the minimum value of x_k where $x \in X_i$ and $\mathbf{x}$ is in $J_{ij}(l-1)$ if $l \geq 2$,

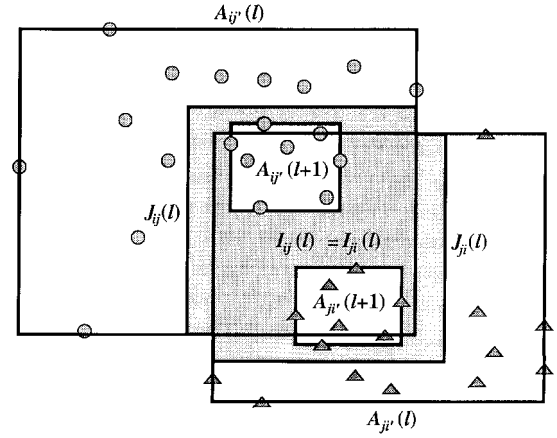


Fig. 1. Concept of the recursive procedure for generating activation and inhibition hyperboxes.

$V_{ij'k}(l)$: the maximum value of x_k where $x \in X_i$ and $\mathbf{x}$ is in $J_{ij}(l-1)$ if $l \geq 2$.

The inhibition hyperbox $I_{ij}(l)$ is defined as

$$I_{ij}(l) = \{\mathbf{x} |\, w_{ijk}(l) \leq x_k \leq W_{ijk}(l), \;\; k = 1, \cdots, m\} \tag{4}$$

where $v_{ij'k}(l) \leq w_{ijk}(l) \leq W_{ijk}(l) \leq V_{ij'k}(l)$.

The expanded inhibition hyperbox $J_{ij}(l)$ is defined as

$$J_{ij}(l) = \{\mathbf{x} |\, u_{ijk}(l) \leq x_k \leq U_{ijk}(l), \quad k = 1, \cdots, m\} \tag{5}$$

where $v_{ij'k}(l) \leq u_{ijk}(l) \leq w_{ijk}(l) \leq W_{ijk}(l) \leq U_{ijk}(l) \leq V_{ij'k}(l)$. To regulate the expansion, an expansion parameter α is introduced. For the case shown in Fig. 1, the expansion process is done according to the following definitions:

1) For $v_{ji'k}(l) \leq v_{ij'k}(l) \leq V_{ji'k}(l) \leq V_{ij'k}(l)$

$$\begin{aligned} u_{ijk}(l) &= v_{ij'k}(l) \\ U_{ijk}(l) &= V_{ji'k}(l) + \alpha[V_{ij'k}(l) - V_{ji'k}(l)]. \end{aligned} \tag{6}$$

2) For $v_{ij'k}(l) \leq v_{ji'k}(l) \leq V_{ij'k}(l) \leq V_{ji'k}(l)$

$$\begin{aligned} u_{ijk}(l) &= v_{ji'k}(l) - \alpha[v_{ji'k}(l) - v_{ij'k}(l)] \\ U_{ijk}(l) &= V_{ij'k}(l). \end{aligned} \tag{7}$$

B. Fuzzy Rule Inference

For a given $\mathbf{x}$, the membership degree with respect to a fuzzy rule given by (1) is 1 if $\mathbf{x}$ is inside of the activation hyperbox $A_{ij'}(l)$. If $\mathbf{x}$ is outside of $A_{ij'}(l)$, it has a lower membership degree. As the distance between $\mathbf{x}$ and $A_{ij'}(l)$ increases, the membership degree of $\mathbf{x}$ decreases and vice versa. We can realize these characteristics by using the following function:

$$m_{A_{ij'}}(\mathbf{x}) = \min_{k=1,\cdots,m} m_{A_{ij'}(l)}(\mathbf{x}, k) \tag{8}$$

$$\begin{aligned} m_{A_{ij'}(l)}(\mathbf{x}, k) &= [1 - \max(0, \min\{1, \gamma_i[v_{ij'k}(l) - x_k]\})] \\ &\quad \times [1 - \max(0, \min\{1, \gamma_i[x_k - V_{ij'k}(l)]\})] \end{aligned} \tag{9}$$

where γ_i is the sensitivity parameter for class i and it is used to regulate the membership degree.

The membership degree of $\mathbf{x}$ with respect to the fuzzy rule $r_{ij}(l)$ given by (1) is defined as

$$d_{r_{ij}(l)}(\mathbf{x}) = m_{A_{ij'}(l)}(\mathbf{x}). \tag{10}$$

The membership degree of $\mathbf{x}$ with respect to a fuzzy rule given by (2) is 1 if $\mathbf{x}$ is inside of the activation hyperbox but not inside of the expanded inhibition hyperbox, i.e., $\mathbf{x}$ is inside of $\overline{A_{ij'}(l) - J_{ij}(l)}$, where $\overline{S}$ denotes the closure of set S and $j' = i$ for $l = 1$ and $j' = j$ for $l \geq 2$. If $\mathbf{x}$ is outside of this region, we set contour surfaces of the membership degree to be parallel to, and to lie at, an equal distance from the surface of $\overline{A_{ij'}(l) - J_{ij}(l)}$, as shown in Fig. 2. To realize this, the membership degree is calculated according to whether or not $\mathbf{x}$ is included in the region $H_{ij}(l)$. $H_{ij}(l)$ is associated with $A_{ij'}(l)$ and $J_{ij}(l)$, and is defined as

$$\begin{aligned} H_{ij}(l) = \{\mathbf{x} |\, & x_k \leq U_{ijk}(l) \\ & \text{for } v_{ji'k}(l) \leq v_{ij'k}(l) \leq V_{ji'k}(l) \leq V_{ij'k}(l), \\ & x_k \geq u_{ijk}(l) \\ & \text{for } v_{ij'k}(l) \leq v_{ji'k}(l) \leq V_{ij'k}(l) \leq V_{ji'k}(l), \\ & -\infty \leq x_k \leq \infty \\ & \text{for } v_{ji'k}(l) \leq v_{ij'k}(l) \leq V_{ij'k}(l) \leq V_{ji'k}(l), \\ & u_{ijk}(l) \leq x_k \leq U_{ijk}(l) \\ & \text{for } v_{ij'k}(l) \leq v_{ji'k}(l) \leq V_{ji'k}(l) \leq V_{ij'k}(l), \\ & k = 1, \cdots, m\} \end{aligned} \tag{11}$$

where $j' = i$ and $i' = j$ for $l = 1$, $j' = j$, and $i' = i$ for $l \geq 2$. It is noted that $H_{ij}(l)$ and $H_{ji}(l)$ are in general different. The region $H_{ij}(l)$, as shown in Fig. 2, defines an input region where the expanded inhibition hyperbox has an effect on the membership degree with respect to the rule given by (2). Thus the membership degree with respect to $r_{ij}(l)$ given by (2) is calculated by

$$\begin{aligned} d_{r_{ij}(l)}(\mathbf{x}) &= m_{A_{ij'}(l)}(\mathbf{x}) && \text{for } \mathbf{x} \notin H_{ij}(l) \\ &= \min\left[m_{A_{ij'}(l)}(\mathbf{x}), m_{J_{ij}(l)}(\mathbf{x})\right] && \text{for } \mathbf{x} \in H_{ij}(l) \end{aligned} \tag{12}$$

where $m_{J_{ij}(l)}(\mathbf{x})$ is given by

$$m_{j_{ij}(l)}(\mathbf{x}) = \max_{k=1,\cdots,m} m_{J_{ij}(l)}(\mathbf{x}, k). \tag{13}$$

The definition of $m_{J_{ij}(l)}(\mathbf{x}, k)$ has a form similar to (9) [8].

The final membership degree of $\mathbf{x}$ with respect to a set of fuzzy rules $\{r_{ij}(l) |\, l = 1, \cdots, l_{ij}\}$, denoted as $d_{r_{ij}(l)}(\mathbf{x})$, is given by

$$d_{r_{ij}}(\mathbf{x}) = \max_{l=1,\cdots,l_{ij}} \left[d_{r_{ij}(l)}(\mathbf{x})\right] \tag{14}$$

where l_{ij} is the deepest level of the overlaps between classes i and j.

Here, we take the maximum because the activation hyperbox $A_{ij}(l+1)$, if it exists, is included in the expanded inhibition hyperbox $J_{ij}(l)$, and hence each fuzzy rule in $\{r_{ij}(l) | l = 1, \cdots, l_{ij}\}$ is exclusive of any others.

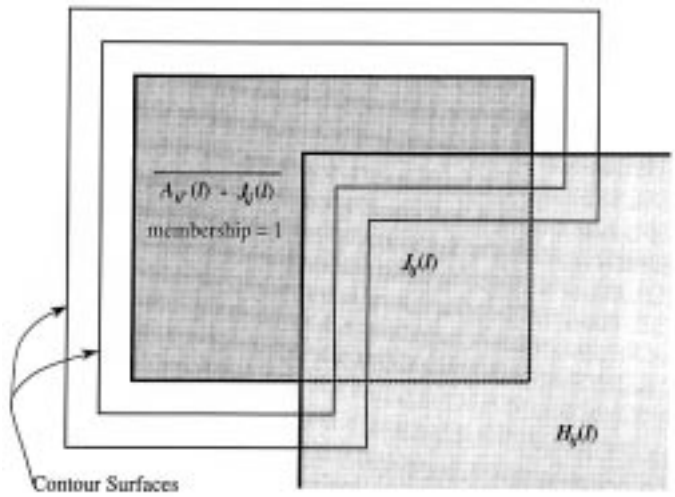


Fig. 2. Contour surfaces of the membership degree for a fuzzy rule in which an inhibition hyperbox exists.

Finally, the membership degree of $\mathbf{x}$ for class i, denoted as $d_i(\mathbf{x})$, is given by

$$d_i(\mathbf{x}) = \min_{\substack{j \neq i,\, j=1,\cdots,n, \\ A_{ii}(1) \cap A_{jj}(1) \neq \emptyset}} \left[d_{r_{ij}}(\mathbf{x})\right]. \tag{15}$$

When the activation hyperbox of class i overlaps with those of classes j and k, we resolve the conflict, independently, first between classes i and j, then between classes i and k. This process is accomplished by taking the minimum in (15). The input $\mathbf{x}$ is then classified as class i if $d_i(\mathbf{x})$ is the maximum among $d_j(\mathbf{x})$, $j = 1, \cdots, n$.

III. Feature Elimination Based on Class Regions

A. Exception Ratio

An activation hyperbox of a given class is a region where data points of the class in the feature space are generalized in the form of hyperrectangles. Similar methods can be found in [9] and [10]. Unlike the method described in Section II, these two methods operate incrementally in the learning process, i.e., the generalization process takes into account one training datum at a time. In addition, in [9] generalization is constrained by a user-defined parameter controlling the size of the hyperboxes, while in [10] it is constrained by a matching process.

Following the interpretation in [10], the inhibition hyperbox $I_{ij}(l)$ can be regarded as an exception of the activation hyperbox $A_{ij}(l)$. It is noted that, for feature evaluation, we use here the inhibition hyperbox $I_{ij}(l)$, rather than the expanded inhibition hyperbox $J_{ij}(l)$, to exactly represent a region defined from the data for generating rules. If a given datum $\mathbf{x}$ is located inside of $A_{ij}(l)$ but outside of $I_{ij}(l)$, supposing that $I_{ij}(l)$ exists, the inference of $A_{ij}(l)$ will exclusively contribute to the membership degree of $\mathbf{x}$ with respect to the rule $r_{ij}(l)$. On the contrary, if $\mathbf{x}$ is located inside of $I_{ij}(l)$, the membership degree of $\mathbf{x}$ with respect to the rule $r_{ij}(l)$ results from the inference of both $A_{ij}(l)$ and $I_{ij}(l)$.

Consequently, the larger the size of the inhibition hyperbox, or the exception, in a given activation hyperbox, the less is the contribution of the activation hyperbox to the classification of the corresponding class and vice versa. This indicates that for a given set of features, a subset of features upon which the generated rules have the lowest number of exceptions has the tendency to contain the most relevant features, compared to the other subsets with the same number of features.

We determine the number of exceptions for a pair of classes based on the exception ratio, the computation of which is described in the following. First, at each overlapping level the ratio of the size of the inhibition hyperbox to the size of the activation hyperbox is computed. Next, since the deeper the level of a rule, the less is the contribution of that rule to the classification, the ratio computed at each level is weighted by the probability to find a datum of the corresponding class inside of the inhibition hyperbox. This probability corresponds to the frequency at which inference of rules inside of the inhibition hyperbox has an effect on the classification of the corresponding class. Finally, the exception ratio is computed by taking the sum for all levels of the weighted ratio.

Let F denote a set of features upon which rules are generated. To be more precise, we define the exception ratio $o_{ij}(F)$ as follows:

$$o_{ij}(F) = \sum_{l=1,\cdots,l_{ij}} p_{ij}(l)\,\frac{B_{I_{ij}}(F,l)}{B_{A_{ij'}}(F,l)} \tag{16}$$

where

$$\begin{aligned} B_{X_{ij}}(F,l) &= \prod_{f\in F} b_{X_{ij}}(f,l),\\ b_{I_{ij}}(f,l) &= W_{ijf}(l) - w_{ijf}(l),\\ &\quad \text{for } W_{ijf}(l) - w_{ijf}(l) > \varepsilon\\ &\quad = \varepsilon \text{ otherwise},\\ b_{A_{ij'}}(f,l) &= V_{ij'f}(l) - v_{ij'f}(l),\\ &\quad \text{for } V_{ij'f}(l) - v_{ij'f}(l) > \varepsilon\\ &\quad = \varepsilon \text{ otherwise}, \end{aligned}$$

ε is a small number

and

$$p_{ij}(l) = \frac{\text{number of class } i \text{ training data in } I_{ij}(l)}{\text{total number of training data}}.$$

In the above formula, it is necessary to limit the smallest value of $b_{X_{ij}}(f,l)$ to ε. This is done to allow the computation of $o_{ij}(F)$ for the case where there exists a feature f^* in F such that $V_{ijf^*}(l) - v_{ijf^*}(l) = 0$.

To verify the aforementioned formula, iris data [11], consisting of three classes and four features, are considered here. Details of the iris data are discussed in Section IV. Here we show the projection of the original data for the three classes on two different sets of two features $\{f_1, f_2\}$ and $\{f_3, f_4\}$ in Fig. 3(a) and (b), respectively. The activation hyperboxes obtained when half of the original data are used for generating rules are also superimposed on the figure. It is apparent from the figure that the set $\{f_3, f_4\}$ is better or more relevant than its counterpart in that class regions are clearer, or in other words, better separated. The exception ratio for each set is shown in Table I where $\hat{o}_{ij}(F) = o_{ij}(F) + o_{ji}(F)$. From this table, the result that the sum of $\hat{o}_{ij}(\{f_3, f_4\})$ is less than that of $\hat{o}_{ij}(\{f_1, f_2\})$ substantiates well the idea of using the exception ratio as a measure for evaluating features.

TABLE I
THE EXCEPTION RATIO OF TWO DIFFERENT SETS OF TWO FEATURES FOR THE IRIS DATA

Exception Ratio	Set of Features $\{f_1, f_2\}$	Set of Features $\{f_3, f_4\}$
$\hat{o}_{12}(F)$	0.71×10^{-2}	0
$\hat{o}_{13}(F)$	5.76×10^{-2}	0
$\hat{o}_{23}(F)$	81.10×10^{-2}	5.72×10^{-2}

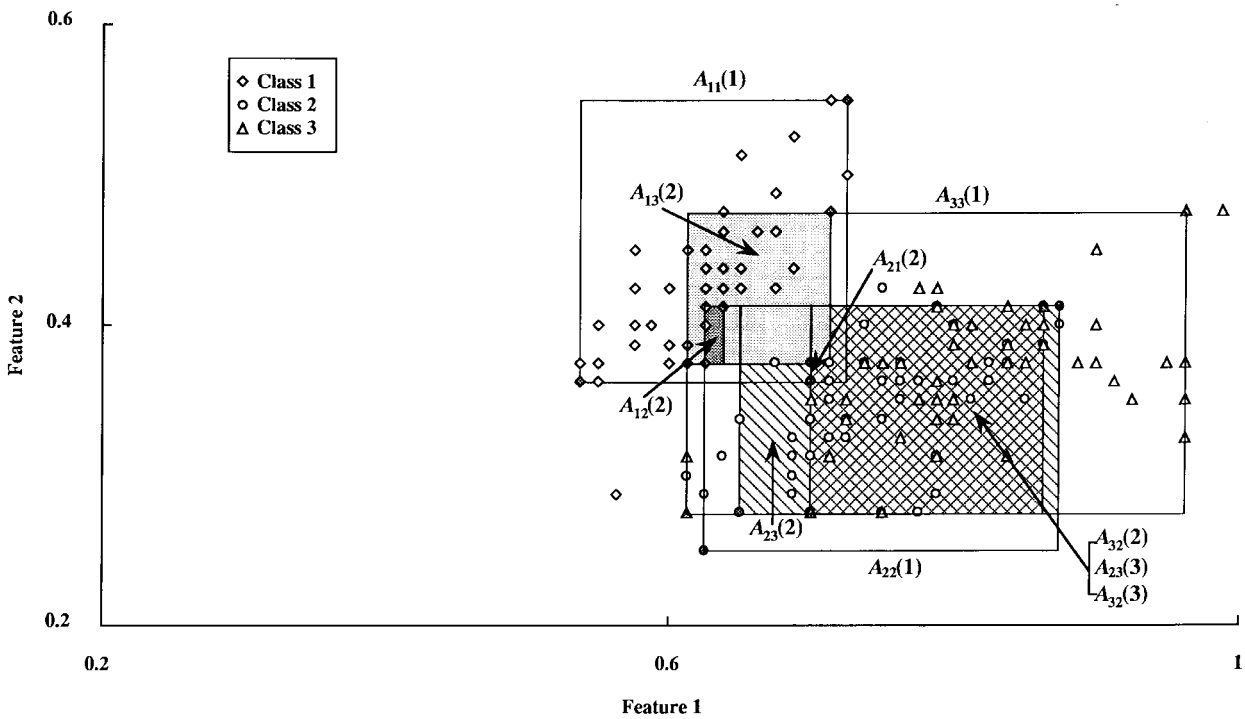
The exception ratio given by (16) can be computed very quickly using concise representation of axis-parallel hyperboxes generated by the fuzzy classifier. This type of representation is ideal for approximating class regions in domains that are horizontally or vertically oriented. The concept of the exception ratio can also be applied to other domains where axis-nonparallel hyperboxes or other types of representation, e.g., ellipsoidal regions, are more appropriate. This, however, essentially requires modification of the method in [8] that we use for generating rules, which is beyond the scope of this paper.

B. Exception Ratio Based Feature Elimination Algorithm

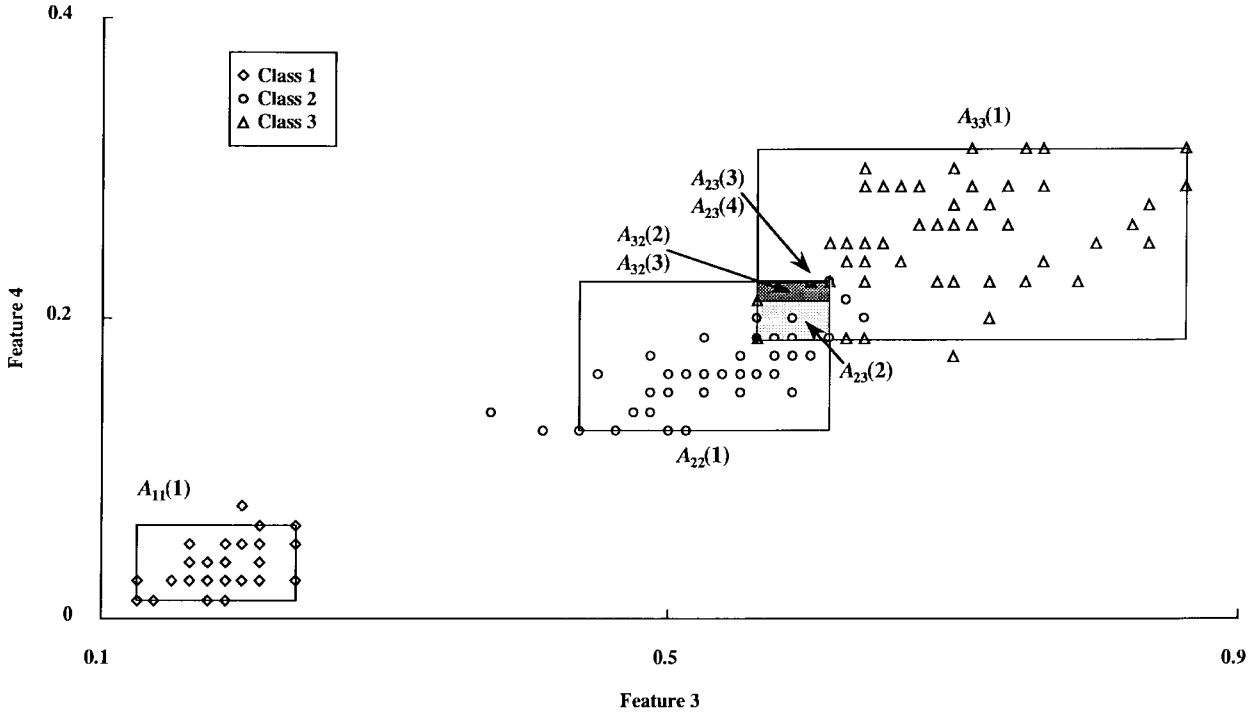
To select features, we take the backward selection search technique [1], which begins with all the features and eliminates the most irrelevant feature. To find this most irrelevant feature, each of the features is temporarily eliminated. The sum of the exception ratios after each temporary elimination is then computed. The feature the elimination of which minimizes the sum of the exception ratios is chosen. The chosen feature is the most irrelevant to classification, compared with the other features. This is because after the elimination of the chosen feature, rules generated based upon the set of remaining features have the lowest number of exceptions. The procedure then continues to eliminate the next most irrelevant feature.

In addition, monitoring an increase in the sum of the exception ratios, we can heuristically give a terminating criterion. When only relevant features are left in the remaining features, additional elimination of one single feature results in much more complicated class regions, hence many more overlaps or exceptions. Therefore, a terminating threshold can be given. The procedure then terminates if after additional elimination, the increase in the sum of the exception ratios of the set of remaining features, compared to that of the set of the original features, is beyond the given threshold. With this terminating criterion, it is expected that the performance of a classifier built upon the selected features does not degrade much compared to that of a classifier built upon the original features.

From the above ideas, now we propose the following algorithm to eliminate irrelevant features based on the exception ratio. Let $O(F)$ denote the sum of the exception ratios and be defined as $\sum_{\substack{i,j\\ i\neq j}} o_{ij}(F)$. Then let F^m denote the set



(a)



(b)

Fig. 3. Projection in two dimensions of the iris data superimposed with activation hyperboxes when using half of the data to generate rules.

of m remaining features and F_i^m be defined as $F^{m+1} - \{f_i^{m+1}\}$ where f_i^{m+1} is the ith element in F^{m+1}. It is noted that F_i^m is the set of m features obtained by temporarily eliminating f_i^{m+1} from F^{m+1}. Let F_{org} denote the set of the original M features where $M \geq 2$. The exception ratio based feature elimination (ERFE) algorithm can be described by the following procedure:

Step 1: Initialize F^m by setting $F^m \leftarrow F_{org}$, hence $m = M$.

Step 2: Compute $O(F_i^{m-1})$ for $i = 1, \cdots, m$.

Step 3: Find the feature f_j^m that $O(F_j^{m-1}) = \min_i [O(F_i^{m-1})]$.

Step 4: If $[O(F_j^{m-1}) - O(F_{org})]/O(F_{org}) < \beta$ go to step 5; otherwise terminate.

Step 5: Set $F^{m-1} \leftarrow F_j^{m-1}$. ($f_j^m$ is permanently eliminated from F^m.)

Step 6: Set $m = m - 1$. If $m = 1$, terminate; otherwise go to Step 2.

Step 4 checks if the increase rate in the sum of the exception ratios exceeds the terminating threshold β. In the following sections, this rate is referred to as the exception increase rate. If β is set to zero, representing the most conservative criterion, only elimination of features that does not at all increase the complexity of class regions is performed. Doing so, we can expect that the increase in recognition error rates for unknown data of the classifiers in use is constrained, provided that the characteristic of the unknown data is similar to that of the data used for generating fuzzy rules. In practice, from many problem domains tested in the next section, we find that we can loosen the criterion by allowing a value of β up to 0.5.

We note here that it is also possible to implement an algorithm that performs forward selection search [1] based on the exception ratio. In this algorithm, given a set of already selected features, the next feature to be selected is the one the addition of which minimizes the sum of the exception ratios. Similarly, a terminating threshold can be given. The algorithm terminates if after further addition, the decrease in the sum of the exception ratios of the set of selected features, compared to that of the set of the original features, is less than the terminating threshold. Below we refer to this algorithm as the Exception Ratio based Feature Addition (ERFA) algorithm. Further discussion on the ERFA algorithm is given in Section IV-E.

Of course, due to its heuristic characteristic, we cannot guarantee the optimal selection of subsets of features using either backward selection search or forward selection search. As far as the optimal subset of features is concerned, some other search techniques, e.g., those with backtracking mechanisms, can also be applied, but they need more computational effort. In this paper, we therefore consider the backward and forward selection searches that have computational advantages over the other search techniques.

IV. Experimental Results

The ERFE algorithm is tested using real data of four problems from different domains: 1) iris data, 2) thyroid data, 3) numeral data, and 4) blood cell data. The first two data sets are well-known benchmark data for classification. The others are data sets used in our original applications. For each data set, all of the available data are divided into training data and test data. The training data are used both for eliminating features and training classifiers. The test data are used for evaluating the recognition rate of the classifiers.

Two classifiers are used, namely, the fuzzy classifier [8] described in Section II, and a back propagation neural net classifier [12]. Unless explicitly specified, the following sets of parameters are used for the fuzzy classifier and the neural net classifier, respectively.

Fuzzy Classifier:

$$\text{expansion parameter} = 0.001$$

and

$$\text{sensitivity parameter} = 1.$$

Neural Net Classifier:

$$\text{learning rate} = 1$$

and

$$\text{momentum} = 0.$$

In addition, for the neural net classifier, a three-layered net is used. To obtain a recognition rate, a set of 10 runs is executed, each run having initial connection weights randomly assigned between -0.1 and 0.1. The average value of the results from the 10 runs is then taken. For each data set, we choose the number of hidden units and the number of training epochs so that satisfactory performance can be achieved from the neural net classifier on the original features.

The parameters used in the ERFE algorithm are as follows: $\varepsilon = 0.001$ and $\beta = 1 \times 10^6$. The value of β is intentionally set to such a large value in order to let the algorithm proceed until the number of remaining features becomes 1. However, it is shown in the following sections that a robust value of β can be determined which serves well as the terminating threshold for the data sets mentioned above.

A. Iris Data

The iris data [11] consist of 150 data with four input features and three classes. Training and test data are composed of the first 25 data and the remaining 25 data of each class, respectively. For the neural net classifier, a net with 3 hidden units is used and the net is trained for 1000 epochs.

Fig. 4(a) and (b) plot the recognition rates of the fuzzy classifier against a wide range of expansion parameter values and the learning curve of the neural net classifier, respectively. The rates of all possible combinations of two features are shown in the figure. Binary presentation is used to present each combination, i.e., the ith digit is 1 if the ith feature is present; otherwise 0.

The combination of features obtained by the ERFE algorithm is "0011," upon which class regions are well separated [cf., Fig. 3(b)]. Both classifiers have better performance using this combination than the other combinations.

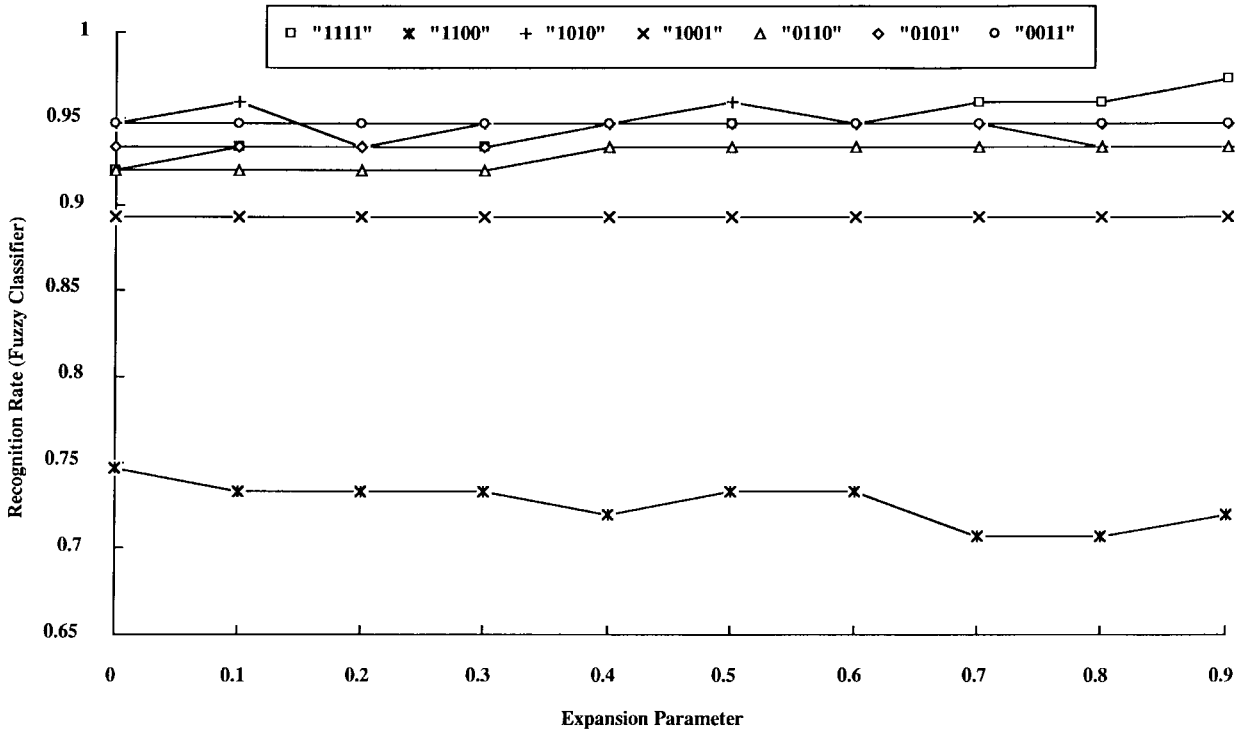
B. Thyroid Data

The thyroid data [11] consist of 3772 training data and 3428 test data with 21 input features, among which 15 are binary and 6, analog. These data belong to one of the three classes. Since there is one class which occupies over 92% of the collected data, any acceptable classifier must have the recognition rate of more than 92%. For the neural net classifier, a net with 3 hidden units is used and the net is trained for 10000 epochs.

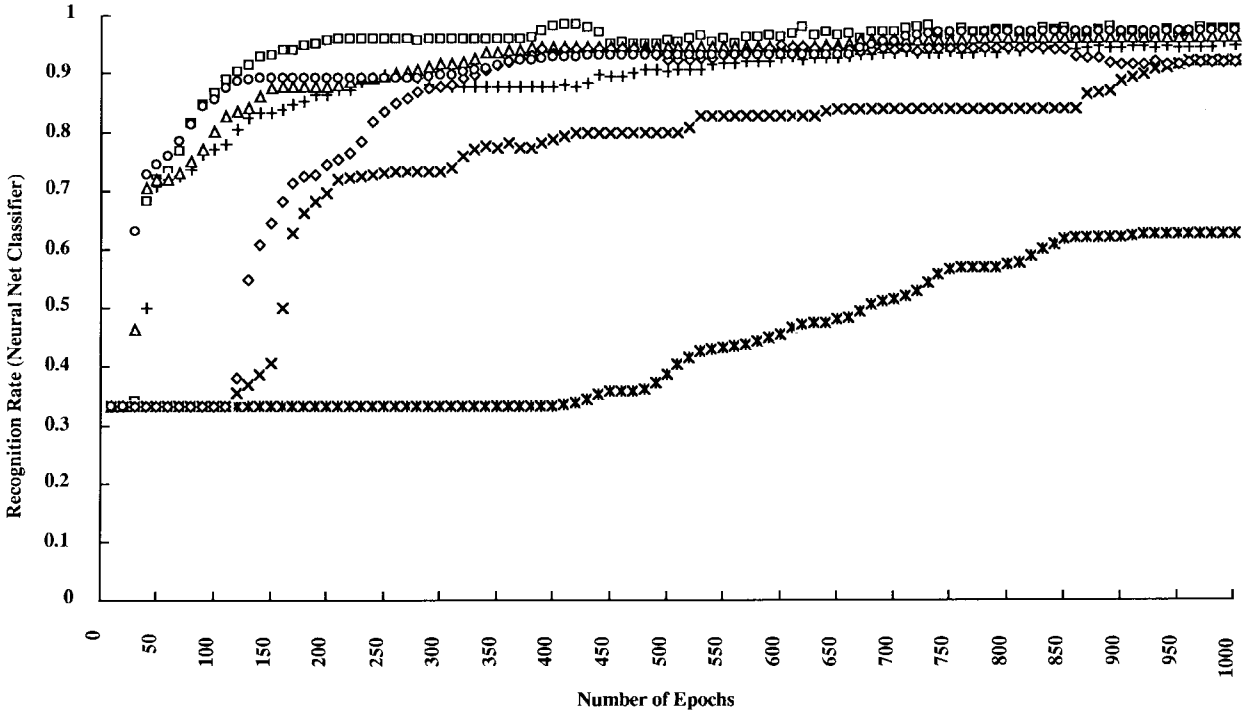
Fig. 5 plots the recognition rates of the two classifiers and the exception increase rate against the number of eliminated features. The maximum recognition rate is achieved for a wide range of numbers of eliminated features. When 15 features are eliminated, the recognition rates of both classifiers start to drop. At the same time, the exception increase rate starts to rise.

C. Numeral Data

This set of data was initially used in a system for number recognition of license plates using a decision-tree algorithm



(a)



(b)

Fig. 4. The iris data. (a) Recognition rates of the fuzzy classifier and (b) the neural net classifier for all the possible combinations of two features out of four features.

[13] and [14]. The task of the system is to recognize 10 numbers using 12 input features extracted from images of moving cars taken by a TV camera. In our study, 1630 data are divided into a combination of 810 training data and 820 test data. For the neural net classifier, a net with six hidden units is used and the net is trained for 4000 epochs.

Fig. 6 plots the recognition rates of the two classifiers and the exception increase rate against the number of eliminated features. For the fuzzy classifier, the maximum recognition rate can be achieved for a wide range of numbers of eliminated features. The same tendency in the recognition rate can be seen for the neural net classifier. For

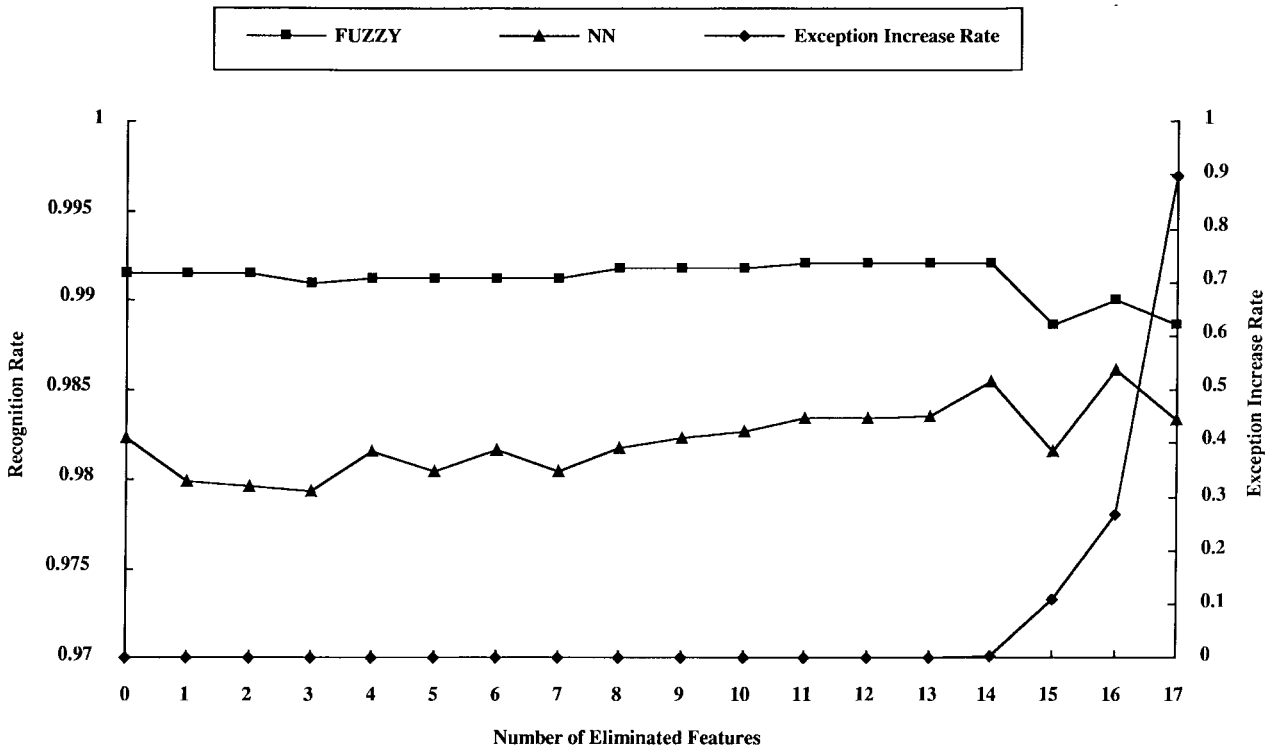


Fig. 5. The thyroid data. Recognition rates of the fuzzy classifier and the neural net classifier plotted together with the exception increase rate.

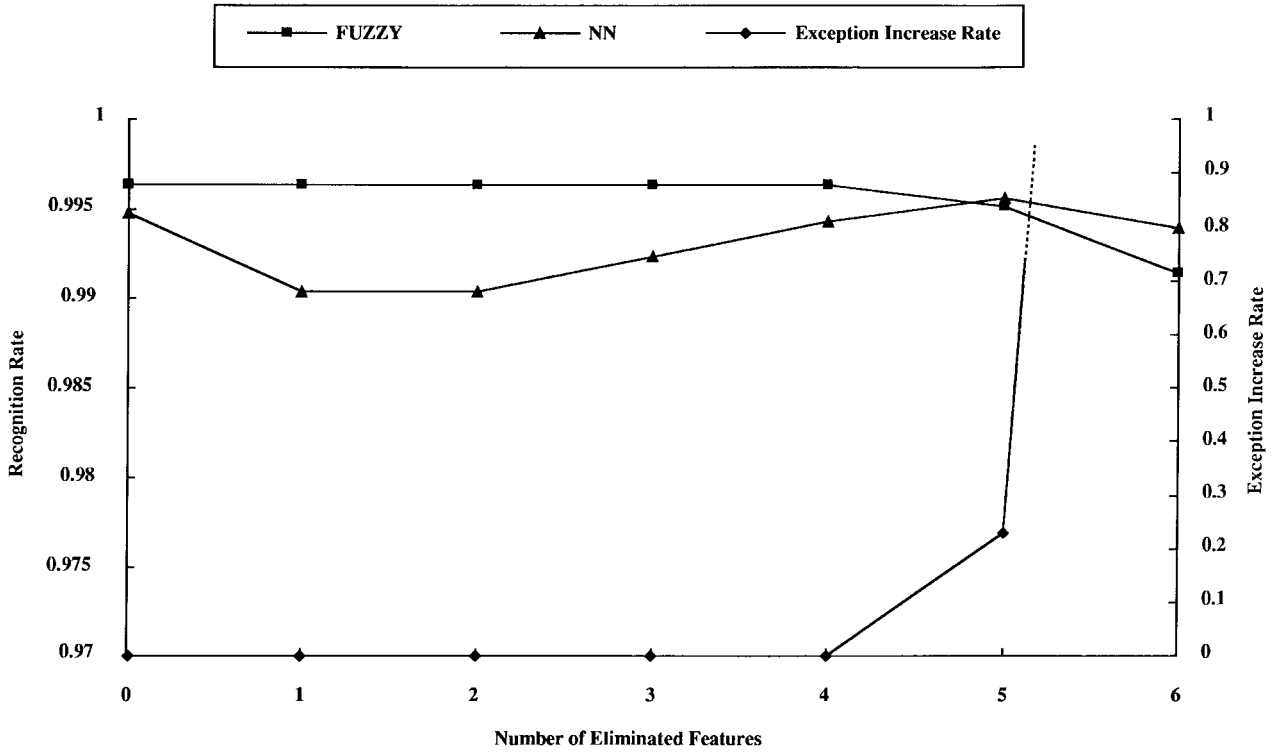


Fig. 6. The numeral data. Recognition rates of the fuzzy classifier and the neural net classifier plotted together with the exception increase rate.

both classifiers, the recognition rates start to drop when five features are eliminated. A rise in the exception increase rate can be seen at this number of eliminated features.

D. Blood Cell Data

The task in this last application [15] is to classify optically screened white blood cells into 12 classes of mature and

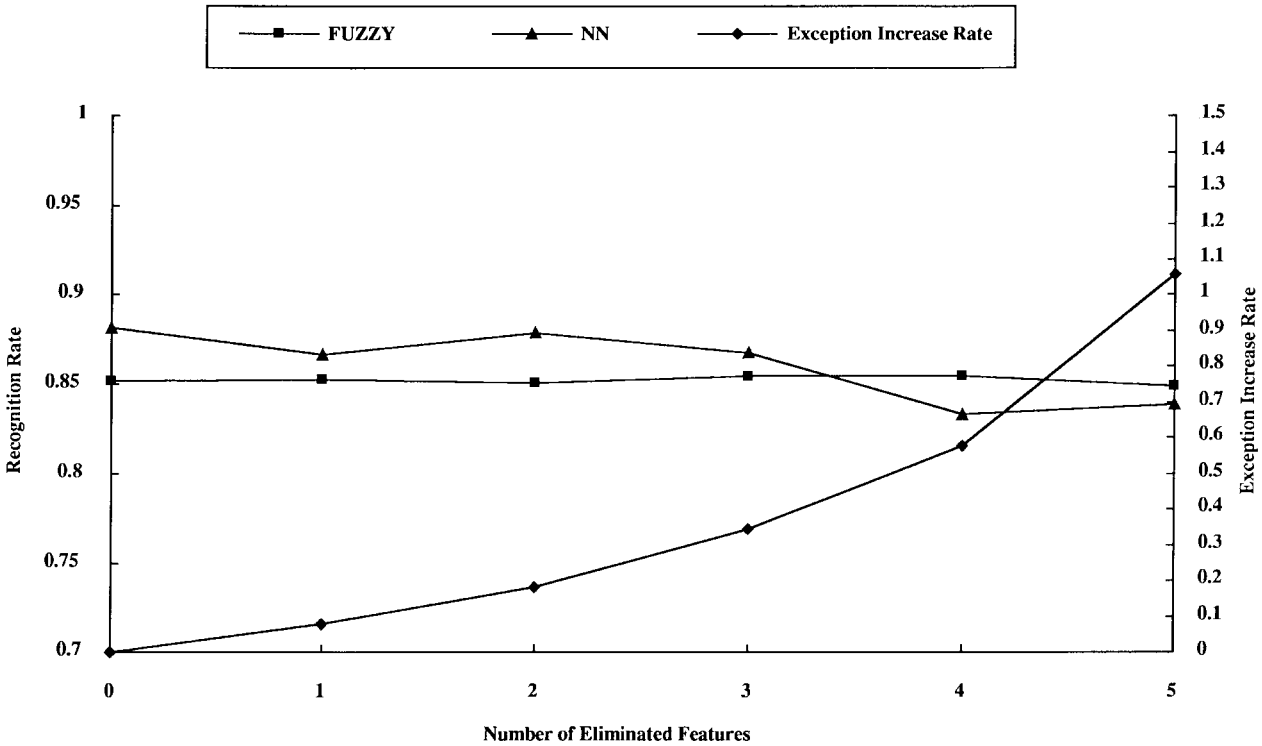


Fig. 7. The blood cell data. Recognition rates of the fuzzy classifier and the neural net classifier plotted together with the exception increase rate.

immature cells using 13 features such as area and perimeter of a kernel. Five of the classes are mature and the others immature. The blood cell classification is known to be a very hard problem. In our study, there are 6196 blood cell data. These data are divided into 3097 training data and 3100 test data. For the neural net classifier, a net with 18 hidden units is used and the net is trained for 6000 epochs.

Fig. 7 plots the recognition rates of the two classifiers and the exception increase rate against the number of eliminated features. For the fuzzy classifier, the recognition rate starts to drop when the number of eliminated features is five. It is three for the neural net classifier. The exception increase rate starts to rise at the first elimination.

E. Comparison with Other Methods and Discussion

In this section, we show comparison results with some other methods. For nonparametric classification, recent methods such as decision boundary feature extraction in [4] or mutual information feature selection in [7] have some parameters that must be appropriately chosen for each classification problem, namely, the decision boundary searching threshold in the former method and the number of quantization levels in the latter method. Fair empirical comparison of our method with them is nontrivial, if not infeasible. This leads to the need of theoretical understanding of these three methods so that unbiased comparison can be performed, which is left as a challenging open problem.

From the above considerations, we conduct additional tests using three popular conventional feature reduction methods that do not have sensitive parameters influencing the performance. The first two are feature extraction methods: principal component analysis (PCA) and discriminant analysis (DA). The last one is a feature selection method that performs backward selection search using interclass Euclidean distance as the class separability measure (EDFE). It is known that the PCA method is not optimal because it does not take into account the information about the individual classes. In addition, the DA method is not reliable if the class means are near to one another. On the contrary, our method does not have these drawbacks.

In the tests, the number of features is set to 2 for the iris data. For the other data sets, we use the numbers of features that are selected by the ERFE algorithm with $\beta = 0.5$. Namely, the numbers of features for the thyroid data, numeral data, and blood cell data are 5, 7, and 10, respectively. From the previous results, satisfactory performance of the classifiers is retained if they are built upon the features selected by the ERFE algorithm with these numbers. The accumulation of eigenvalues of the PCA method for each data set is shown in Table II. Since the accumulation of eigenvalues corresponds to classification performance, it can be expected that the performance of the PCA method is good for the blood cell data but unfavorable for the thyroid data. For training and testing the fuzzy and neural net classifiers, we use the same conditions as those elaborated in the previous sections.

Table III summarizes the results. As can be seen, the ERFE algorithm has the most preferable performance, especially for the thyroid data. The DA method cannot be applied for the thyroid data because of zero diagonal elements of the covariance matrices. There is only one case where the PCA method outperforms the ERFE algorithm, i.e., the case where the neural net classifier is used for the blood cell data. For

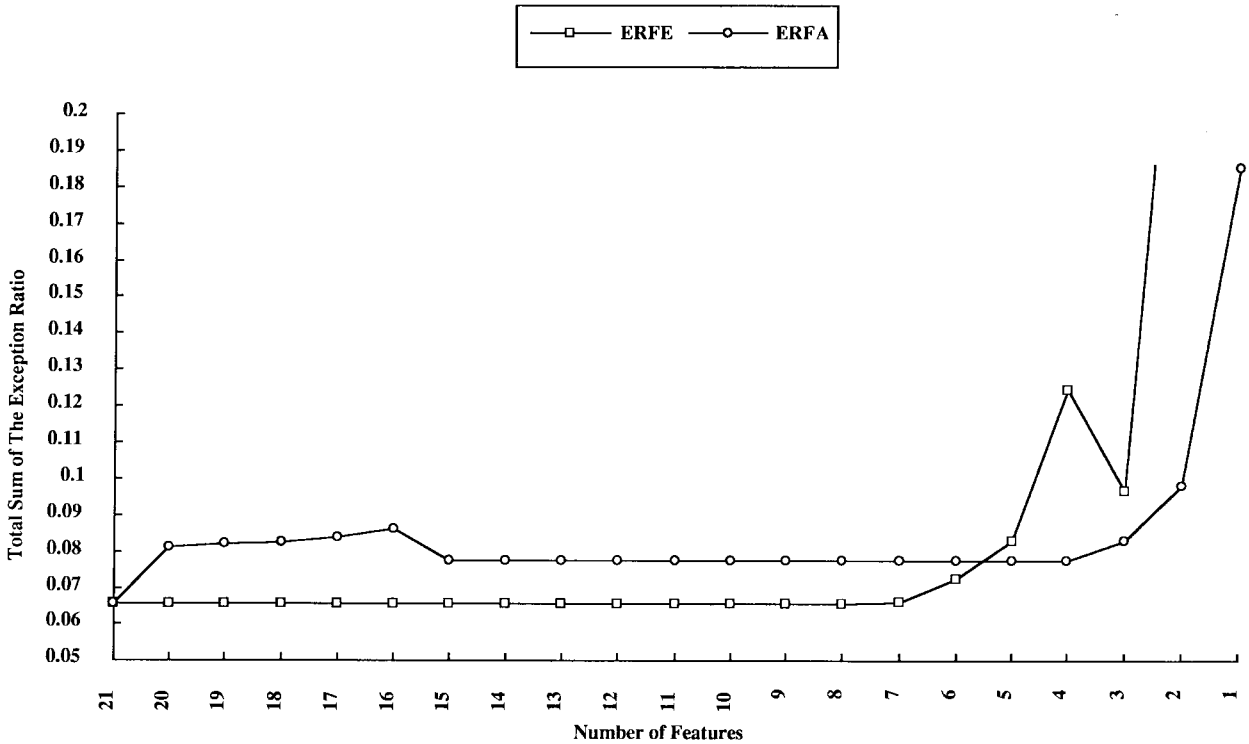


Fig. 8. The thyroid data. The sum of the exception ratios obtained from the ERFE and ERFA algorithms.

TABLE II
THE ACCUMULATION OF EIGEN VALUES (ACC. EV.) OF PRINCIPAL COMPONENT ANALYSIS (PCA) FOR EACH DATA SET

Data set	Number of Features	Acc. Ev. (%)
Iris	2	97.79
Thyroid	5	73.33
Numeral	7	93.45
Blood Cell	10	99.53

TABLE III
COMPARISON OF FEATURE REDUCTION TECHNIQUES

Data Set	Original Features	Reduced Features			
		ERFE	PCA	DA	EDFE
	FUZZY/NN	FUZZY/NN	FUZZY/NN	FUZZY/NN	FUZZY/NN
Iris	92.00/97.47	94.67/97.33	90.67/97.33	68.00/39.60	94.66/94.80
Thyroid	99.15/98.23	99.01/98.61	85.82/92.63	-----	42.50/92.71
Numeral	99.63/99.48	99.51/99.56	98.90/99.28	96.22/96.72	95.61/93.93
Blood	85.16/88.16	85.45/86.77	83.23/88.15	77.39/34.96	82.39/83.89

the iris and blood cell data, the neural net classifier converges very slowly when trained with the features extracted by the DA method. By extending the number of epochs to 10000, the rate is increased to 65.73% for the iris data, while significant improvement of the rate cannot be obtained for the blood cell data. For the thyroid data, the rate of the fuzzy classifier trained with the features selected by the EDFE method is significantly lower than the corresponding rate of the neural net classifier. This result indicates that the class regions in the selected features are not horizontally or vertically oriented. As a result, the axis-parallel hyperboxes generated by the fuzzy classifier do not fit well to the class regions in this case.

It is interesting to compare the ERFE algorithm with the ERFA algorithm. From our experience, both algorithms under the same terminating threshold have compatible performance for most of the data sets, namely, the iris data, numeral data, and the blood cell data, though the ERFE algorithm performs slightly better. For the thyroid data, however, the ERFA algorithm yields notably worse results. To explain this case, the sum of the exception ratios obtained from the ERFE and ERFA algorithms are plotted against the number of features in Fig. 8. As can be seen, the ERFA algorithm has higher values for a wide range of numbers of features. This means that within that range, class regions in the feature space of the features selected by the ERFA algorithm are more complicated, resulting in inferior classification performance. In addition, until the last feature is added, the sum of the exception ratios obtained from the ERFA is larger than that of the original features. This result indicates that the lastly added feature is a relevant feature but it cannot be detected by the ERFA algorithm. As a consequence, we conjecture that, in general, features selected by the ERFE algorithm are more reliable than those selected by the ERFA algorithm. The reason for this conjecture is that in the ERFE algorithm, the interdependence of features with higher dimensionality are taken into account in order to determine features with lower dimensionality while the opposite is done in the ERFA algorithm.

In [8], another algorithm is discussed that eliminates features subject to a constraint that the number of fuzzy rules generated by the fuzzy classifier does not increase at each elimination.

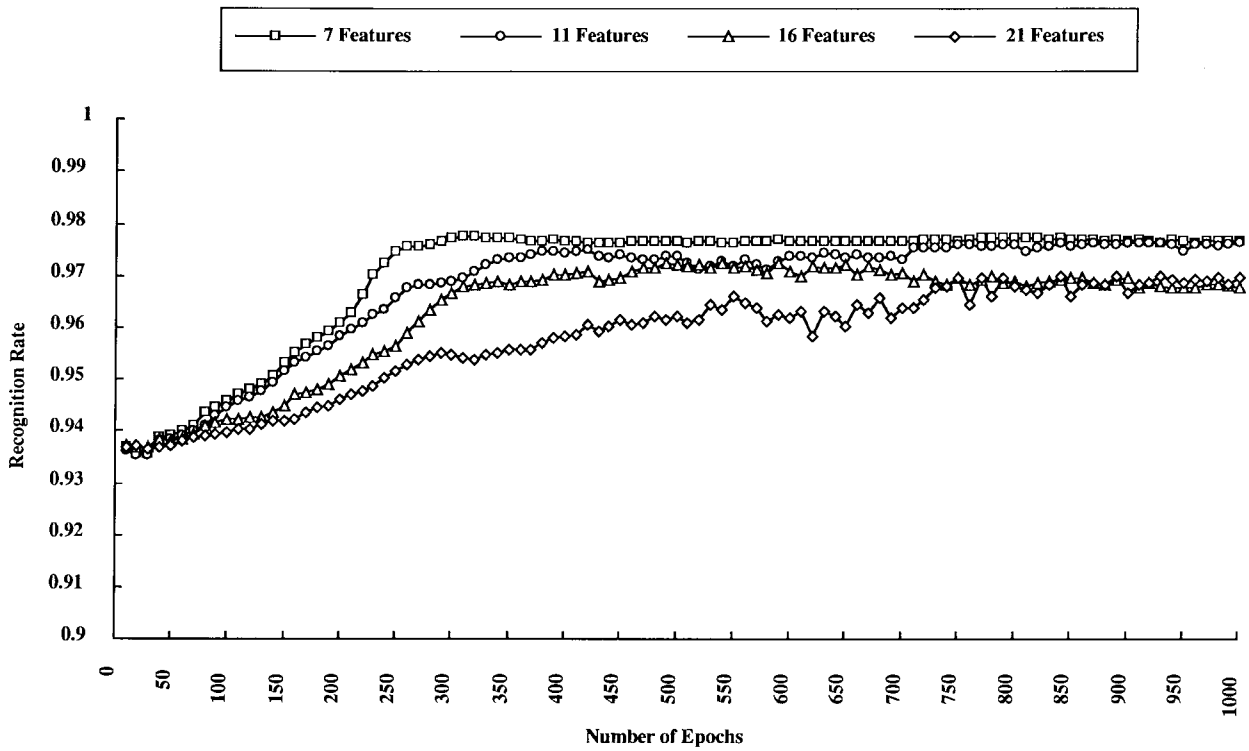


Fig. 9. The thyroid data. The learning curve of the neural net classifier for different sets of features selected by the ERFE algorithm for which the exception increase rates remain zero.

In practice, this algorithm is too conservative in terms of the number of features that can be eliminated. For the numeral data, the maximum number of features that can be eliminated by the algorithm is three.

The above results substantiate well the fact that the ERFE algorithm can successfully select features by eliminating irrelevant features. In the case of the numeral data and thyroid data, satisfactory classification performance is retained for a wide range of numbers of eliminated features. Within this range, the exception increase rate remains zero, implying that the complexity of class regions does not change. This can be validated by the learning curve of the neural net classifier. If the same complexity of class regions can be achieved, the lower the number of selected features, the lower is the number of epochs required to reach the peak performance, due to the decrease in the number of weights which need to be updated in the net. Fig. 9 shows the learning curve for the thyroid data with different sets of features for which the exception increase rates remain zero.

In addition, it can be seen that the classification performance of the classifiers drops as there is a rise in the exception increase rate. In the case of the blood data, known to be hard to classify, this tendency also holds. Though the complexity of class regions increases at the first elimination, this does not mean that elimination of features is not at all possible. Analyzing the results of all the data sets, we find that the terminating threshold β can be heuristically set to 0.5 in order to retain satisfactory performance of the classifiers. More experiments are needed to obtain a better understanding on the relationship between the classification performance and the exception increase rate.

V. Conclusions

We have presented a novel approach to feature selection based on analysis of class regions which are generated by a fuzzy classifier. For feature evaluation, we proposed the exception ratio as a measure of the degree of overlaps in the class regions, in other words, the degree of having exceptions inside of fuzzy rules generated by the fuzzy classifier. The idea of using the exception ratio derived from the fact that for a given set of features, a subset of features that has the lowest sum of the exception ratios has the tendency to contain the most relevant features, compared to the other subsets with the same number of features. An algorithm was proposed that performs elimination of features based on the exception ratios and terminates when further elimination of a single feature degrades the classification performance. Extensive experiments were conducted using four types of data namely, iris data, thyroid data, numeral data, and blood cell data. The fuzzy classifier and a back-propagation neural net classifier were used to evaluate the features selected by the proposed algorithm. It was shown by the experiments that the proposed algorithm could successfully select features by eliminating irrelevant features.

Acknowledgment

The authors are grateful to Prof. N. Matsuda, Kawasaki Medical School, for providing the blood cell data. The thyroid data set was obtained from the machine learning databases at the University of California, Irvine (available FTP at ftp.ics.uci.edu/pub/machine-learning-databases). Thanks are also due to P. M. Murphy and D. W. Aha, who organized

the databases. The authors would like to thank T. W. Rauber of Universidade Nova de Lisboa for providing his pattern recognition software "tooldiag" that contains many conventional methods for feature reduction. They would also like to thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* 2nd ed. San Diego, CA: Academic, 1990.

[2] H. A. Malki and A. Moghaddamjoo, "Using the Karhunen–Love transformation in the back-propagation training algorithm," *IEEE Trans. Neural Networks,* vol. 2, no. 1, pp. 162–165, 1991.

[3] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 15, no. 4, pp. 388–400, 1993.

[4] ——, "Decision boundary feature extraction for nonparametric classification," *IEEE Trans. Syst., Man, Cybern.,* vol. 23, no. 2, pp. 433–444, 1993.

[5] J. Kittler, "Feature selection and extraction," in *Handbook of Pattern Recognition and Image Proc.,* T. Y. Young and K. S. Fu, Eds. San Diego, CA: Academic, 1986, pp. 59–83.

[6] S. K. Pal and B. Chakraborty, "Fuzzy set theoretic measure for automatic feature evaluation," *IEEE Trans. Syst., Man, and Cybern.,* vol. 16, no. 5, pp. 754–760, 1986.

[7] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks,* vol. 5, no. 4, pp. 537–550, 1994.

[8] S. Abe and M.-S. Lan, "A method for fuzzy rules extraction directly from numerical data and its application to pattern classification," *IEEE Trans. Fuzzy Syst.,* vol. 3, no. 1, pp. 18–28, 1995.

[9] P. K. Simpson, "Fuzzy min-max neural networks—Part 1: Classification," *IEEE Trans. Neural Networks,* vol. 3, pp. 776–786, Sept. 1992.

[10] S. Salzberg, "A nearest hyperrectangle learning method," *Mach. Learn.,* vol. 6, pp. 251–276, 1991.

[11] S. M. Weiss and I. Kapouleas, "An empirical comparison of pattern recognition, neural nets, and machine learning classification methods," in *Proc. IJCAI-89,* 1989, pp. 781–787.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition Foundations,* D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: MIT Press, 1986, vol. 1, pp. 318–362.

[13] M. Takatoo *et al.,* "Gray scale image processing technology applied to vehicle license number recognition system," in *Proc. Int. Workshop Industrial Appl. Machine Vision and Machine Intel.,* Feb. 1987, pp. 76–79.

[14] H. Takenaga *et al.,* "Input layer optimization of neural networks by sensitivity analysis and its application to recognition of numerals," *Trans. IEE Jpn.,* vol. 111-D, no. 1, pp. 36–44, 1991, (in Japanese); *Elec. Eng. Japan,* vol. 111, no. 4, pp. 130–138, 1991, (translated into English by Scripta Technica, Inc.).

[15] A. Hashizume, J. Motoike, and R. Yabe, "Fully automated blood cell differential system and its application," in *Proc. IUPAC 3rd Int. Congr. Auto. New Technology Clinical Lab.,* Sept. 1988, pp. 297–302.

Ruck Thawonmas (M'97) received the B.Eng. degree in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 1987, the M.Eng. degree in information science from Ibaraki University, Hitachi, Japan, in 1990, and the D.Eng. degree in information engineering from Tohoku University, Sendai, Japan, in 1994.

Since April 1996, he has been with the Brain Information Processing Group, Institute of Physical and Chemical Research (RIKEN), Wako, Japan, as a Special Postdoctoral Researcher under the Special Postdoctoral Researchers Program. From January 1994 to March 1996, he was a Visiting Researcher under the Hitachi Research Visit Programs (HIVIPS), Hitachi Research Laboratory, Hitachi, Ltd., Hitachi, Japan. His research interests include neural networks, fuzzy systems, and computational intelligence models. He is an author and coauthor of more than 20 peer-reviewed international journal and conference papers.

During his graduate study in Japan, Dr. Thawonmas was the recipient of the Japanese Government (Monbusho) Scholarship from April 1987 to March 1993, and the Asahi Glass Company Scholarship, as well as the Tohoku Kaihatsu Memorial Foundation Research Grant, from April 1993 to January 1994.

Shigeo Abe (M'79–SM'83) received the B.S. degree in electronics engineering, the M.S. degree in electrical engineering, and the Doctor of Engineering degree, all from Kyoto University, Kyoto, Japan, in 1970, 1972, and 1984, respectively.

Since 1972, he has been with Hitachi Research Laboratory, Hitachi, Ltd., and has been engaged in power system analysis, development of a vector processor, a prolog processor, neural network theories, and fuzzy system models. From 1978 to 1979, he was a visiting research associate at the University of Texas, Arlington, TX. He is the author of *Neural Networks and Fuzzy Systems: Theory and Applications* (Norwell, MA: Kluwer).

Dr. Abe was awarded an outstanding paper prize from the Institute of Electrical Engineers of Japan in 1984 and 1995. He is a member of the International Neural Network Society, the Institute of Electrical Engineers of Japan, the Information Processing Society of Japan, the Institute of Electronics, Information, and Communication Engineers of Japan, the Society of Instrument and Control Engineers of Japan, and the National Geographic Society.