Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Pattern Recognition Letters 24 (2003) 509–519

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Finding fuzzy classification rules using data mining techniques

Yi-Chung Hu [a], Ruey-Shun Chen [a], Gwo-Hshiung Tzeng [b,*]

[a] *Institute of Information Management, National Chiao Tung University, Hsinchu 300, Taiwan, ROC*
[b] *Institute of Management of Technology, National Chiao Tung University, Hsinchu 300, Taiwan, ROC*

Received 25 February 2002; received in revised form 29 May 2002

Abstract

Data mining techniques can be used to discover useful patterns by exploring and analyzing data, so, it is feasible to incorporate data mining techniques into the classification process to discover useful patterns or classification rules from training samples. This paper thus proposes a data mining technique to discover fuzzy classification rules based on the well-known Apriori algorithm. Significantly, since it is difficult for users to specify the minimum fuzzy support used to determine the frequent fuzzy grids or the minimum fuzzy confidence used to determine the effective classification rules derived from frequent fuzzy grids, therefore the genetic algorithms are incorporated into the proposed method to determine those two thresholds with binary chromosomes. For classification generalization ability, the simulation results from the iris data and the appendicitis data demonstrate that the proposed method performs well in comparison with other classification methods.

Keywords: Data mining; Fuzzy sets; Classification problems; Genetic algorithms

1. Introduction

Data mining is the exploration and analysis of data in order to discover meaningful patterns (Berry and Linoff, 1997). In addition, data mining problems involving classification can be viewed within a common framework of rule discovery (Agrawal et al., 1993). The goal of this paper is just to propose an effective method that can find a compact set of fuzzy rules for classification problems by data mining techniques.

Recently, the discovery of association rules from databases has become an important research topic, and association rules have been applied for analysis to help managers determine which items are frequently purchased together by customers (Berry and Linoff, 1997; Han and Kamber, 2001). The Apriori algorithm proposed by Agrawal et al. (1996) is an influential algorithm that can be used to find association rules. In this algorithm, a candidate k-itemset ($k \geqslant 1$) containing k items is frequent (i.e., frequent k-itemset) if its support is larger than or equal to a user-specified minimum

[*] Corresponding author. Tel.: +886-3-5712121x57505; fax: +886-3-5753926.
E-mail address: ghtzeng@cc.nctu.edu.tw (G.-H. Tzeng).

0167-8655/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-8655(02)00273-8

support. Significantly, the well-known Apriori property (Han and Kamber, 2001) for mining association rules shows that any subset of a frequent itemset must also be frequent. Subsequently, we use frequent itemsets to generate association rules.

A fuzzy classification rule is a fuzzy if–then rule whose consequent part is a class label. Since the comprehensibility of fuzzy rules by human users is a criterion in designing a fuzzy rule-based system (Ishibuchi et al., 1999), fuzzy classification rules with linguistic interpretations must be taken into account. To cope with this problem, we consider both quantitative and categorical attributes, which are used to describe each sample data, as linguistic variables. Then, each linguistic variable can be partitioned by its linguistic values represented by fuzzy numbers with triangular membership functions. Simple fuzzy grids or grid partitions (Ishibuchi et al., 1999; Jang and Sun, 1995) in feature space resulting from the fuzzy partition are thus obtained.

In this paper, we propose a two-phase data mining technique to discover fuzzy rules for classification problems based on the Apriori algorithm. The first phase finds frequent fuzzy grids by dividing each quantitative attribute with a pre-specified number of various linguistic values. The second phase generates effective fuzzy classification rules from those frequent fuzzy grids. The fuzzy support and the fuzzy confidence, which have been defined previously (e.g., Ishibuchi et al., 2001a; Ishibuchi et al., 2001b; Hu et al., 2002), are employed to determine which fuzzy grids are frequent and which rules are effective by comparison with the minimum fuzzy support (min FS) and the minimum fuzzy confidence (min FC), respectively.

However, both min FS and min FC are not easily user-specified for each classification problem. To solve this problem, the genetic algorithm (GA) (Goldberg, 1989) is thus incorporated into the proposed algorithm to automatically determine those two parameters. A binary chromosome with sufficiently large length used in this paper is composed of two substrings: one for the min FS, and the other for the min FC. Each generation of the GA can obtain the fitness value of each chromosome, which maximizes the classification accuracy rate and minimizes the number of fuzzy rules. When reaching the termination condition, a chromosome with the maximum fitness value is used to test the performance of the proposed method.

For classification generalization ability, the simulation results from the iris data and the appendicitis data demonstrate that proposed learning algorithm performs well in comparison with other fuzzy or non-fuzzy classification methods. Thus, the goal of acquiring an effectively compact set of fuzzy rules for classification problems can be achieved.

This paper is organized as follows. Notations used in this paper are described in Section 2. The fuzzy partition methods are detailed introduced in Section 3. In Section 4, the proposed learning algorithm incorporated with the GA is presented. In Section 5, the performance of the proposed method is examined by computer simulation on Anderson's iris data (Anderson, 1935) and the appendicitis data. Discussions and conclusions are presented in Section 6 and Section 7, respectively.

2. Notations

Notations used in this paper are as follows:

C	number of class labels
d	number of attributes used to describe each sample data, where $1 \leqslant d$
k	dimension of one fuzzy grid, where $1 \leqslant k \leqslant d$
K	number of various linguistic values defined in each quantitative attribute, where $K \geqslant 2$
$A^{x_k}_{K,i_k}$	i_kth linguistic value of K linguistic values in the linguistic variable x_k, where $1 \leqslant i_k \leqslant K$ and $1 \leqslant m \leqslant d$
$\mu^{x_k}_{K,i_k}$	membership function of $A^{x_k}_{K,i_k}$
t_p	pth training sample, where $t_p = (t_{p_1}, t_2, \ldots, t_{p_d})$, and t_{p_i} is the value with respect to the ith attribute
N_{pop}	population size in each generation of the GA
$s^{(j)}$	substring of the jth chromosome. $s^{(j)}$ encodes the min FS, where $1 \leqslant j \leqslant N_{\mathrm{pop}}$
$c^{(j)}$	substring of the jth chromosome. $c^{(j)}$ encodes the min FC, where $1 \leqslant j \leqslant N_{\mathrm{pop}}$

3. Fuzzy partition methods

The concepts of linguistic variables were proposed by Zadeh (1975a,b, 1976) and it is reasonable that we view each attribute as a linguistic variable. Formally, a linguistic variable is characterized by a quintuple (Pedrycz and Gomide, 1998; Zimmermann, 1991) denoted by $(x, T(x), U, G, M)$, in which x is the name of the variable; $T(x)$ denotes the set of names of linguistic values or terms, which are linguistic words or sentences in a natural language (Chen and Jong, 1997), of x; U denotes a universe of discourse; G is a syntactic rule for generating values of x; and M is a semantic rule for associating a linguistic value with a meaning. Using the simple fuzzy partition methods, each attribute can be partitioned by various linguistic values. The simple fuzzy partition methods have been widely used in pattern recognition and fuzzy reasoning. For example, there are the applications to pattern classification by Ishibuchi et al. (1995, 1999), to fuzzy neural networks by Jang (1993), and to the fuzzy rule generation by Wang and Mendel (1992).

In the simple fuzzy partition methods, K various linguistic values are defined in each quantitative attribute. K is also pre-specified before executing the proposed method. Triangular membership functions are usually used for the linguistic values. For example, $K = 3$ and $K = 4$ for the attribute "Width" (denoted by x_1) that ranges from 0 to 60 are shown as Figs. 1 and 2, respectively. That is, three (i.e., small, medium and large) and four (i.e., small, medium small, medium large and large) various linguistic values are defined in Figs. 1 and 2, respectively.

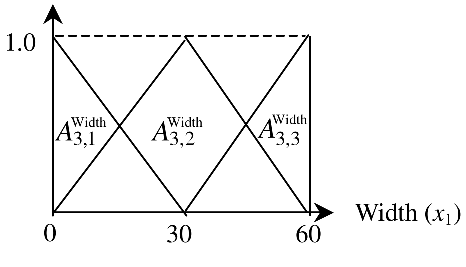

Fig. 1. $K = 3$ for "Width".

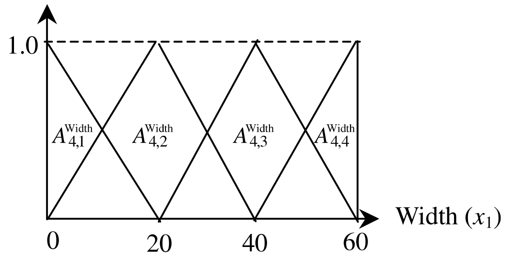

Fig. 2. $K = 4$ for "Width".

In the proposed method, each linguistic value is actually viewed as a candidate 1-dim fuzzy grid. Then, A_{K,j_1}^{Width} can be represented as follows:

$$\mu_{K,j_1}^{\text{Width}}(x) = \max\{1 - |x - a_{j_1}^K|/b^K, 0\} \tag{1}$$

where

$$a_{j_1}^K = \text{mi} + (\text{ma} - \text{mi})(j_1 - 1)/(K - 1) \tag{2}$$

$$b^K = (\text{ma} - \text{mi})/(K - 1) \tag{3}$$

where ma is the maximum domain value, and mi is the minimum value. Here, ma = 60 and mi = 0 for "Width".

If we divide both "Width" and "Length" (denoted by x_2) by three various linguistic values, then a pattern space can be divided into nine 2-dim fuzzy grids, as shown in Fig. 3. We use $A_{3,1}^{\text{Width}} \times A_{3,3}^{\text{Length}}$ to denote the shaded 2-dim fuzzy grid shown in Fig. 3, whose linguistic value is "small AND large".

As for categorical attributes, each has a finite number of possible values, with no ordering among values (e.g., sex, color) (Han and

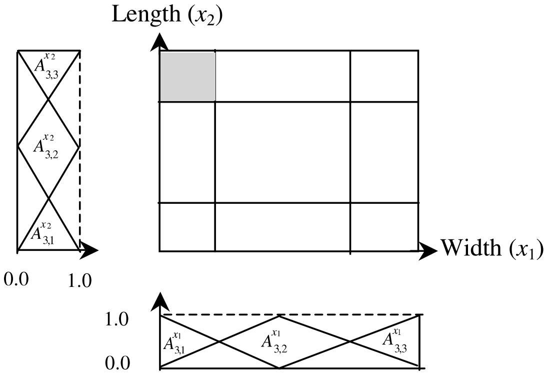

Fig. 3. $K = 3$ for "Width" and "Length".

Kamber, 2001). If the distinct attribute values are n' (n' is finite), then this attribute can only be partitioned by n' linguistic values. For example, since the attribute "Class" is categorical, the linguistic sentence of each linguistic value may be stated as follows:

$A_{2,1}^{\text{class label}}$: class 1

$A_{2,2}^{\text{class label}}$: class 2

It should be noted that the maximum number of dimensions for a single fuzzy grid is d. A significant task is how to use the candidate 1-dim fuzzy grids to generate the other frequent fuzzy grids and effective classification rules. An effective algorithm is thus described in following section.

4. Finding fuzzy classification rules

As we have mentioned above, the generation of frequent fuzzy grids and fuzzy classification rules are two significant phases of the proposed learning algorithm. In this section, we thus describe the individual phase of the proposed method in Sections 4.1 and 4.2. The proposed learning algorithm incorporated with the GA is presented in detail in Section 4.3.

4.1. Determining frequent fuzzy grids

Without loss of generality, given a candidate k-dim fuzzy grid $A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}$, where $1 \leqslant i_1, i_2, \ldots, i_k \leqslant K$, the degree to which t_p belongs to this fuzzy grid can be computed as $\sum_{p=1}^{n} \mu_{A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}}(t_p)$. The fuzzy support (Ishibuchi et al., 2001a; Ishibuchi et al., 2001b; Hu et al., 2002) of $A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}$ is defined as follows:

$$\begin{aligned}
&\mathrm{FS}\left(A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}\right) \\
&\quad = \sum_{p=1}^{n} \mu_{A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}}(t_p)/n \\
&\quad = \left[\sum_{p=1}^{n} \mu_{K,i_1}^{x_1}(t_{p_1}) \cdot \mu_{K,i_2}^{x_2}(t_{p_2}) \cdot \ldots \cdot \mu_{K,i_{k-1}}^{x_{k-1}}(t_{p_{k-1}}) \right. \\
&\qquad \left. \cdot\, \mu_{K,i_k}^{x_k}(t_{p_k}) \right] \Big/ n \qquad (4)
\end{aligned}$$

It is clear that the algebraic product, which is a t-norm operator in the fuzzy intersection, is used in Eq. (4). When $\mathrm{FS}(A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k})$ is larger than or equal to the user-specified min FS, we can say that $A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}$ is a frequent k-dim fuzzy grid. For any two frequent grids, say $A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}$ and $A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k} \times A_{K,i_{k+1}}^{x_{k+1}}$, since $\mu_{A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k} \times A_{K,i_{k+1}}^{x_{k+1}}}(t_p) \leqslant \mu_{A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}}(t_p)(1 \leqslant p \leqslant n)$ from Eq. (4), $A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k} \times A_{K,i_{k+1}}^{x_{k+1}} \subseteq A_{K,i_1}^{x_1} \times A_{K,i_2}^{x_2} \times \cdots \times A_{K,i_{k-1}}^{x_{k-1}} \times A_{K,i_k}^{x_k}$ thus holds. It is clear that any subset of a frequent fuzzy grid must also be frequent. We can observe that this is quite different from the Apriori property, but we may view it as a special property for mining frequent fuzzy grids.

Table FGTTFS is implemented to generate frequent fuzzy grids. FGTTFS consists of the following substructures:

(a) Fuzzy grid table (FG): each row represents a fuzzy grid, and each column represents a linguistic value.
(b) Transaction table (TT): each column represents t_p, and each element records the membership degree of the corresponding fuzzy grid.
(c) Column FS: stores the fuzzy support corresponding to the fuzzy grid in FG.

An initial tabular FGTTFS is shown as Table 1 as an example, from which we can see that there are two samples t_1 and t_2, with two attributes x_1 and x_2. Both x_1 and x_2 are divided into three linguistic values (i.e., $K = 3$). Assume that x_2 is the attribute of class labels. Since each row of FG is a bit string consisting of 0 and 1, FG[u] and FG[v] (i.e., uth row and vth row of FG) can be paired to generate certain desired results by applying the Boolean operations. For example, if we apply the OR operation on two rows, $\mathrm{FG}[1] = (1, 0, 0, 0, 0, 0)$ and $\mathrm{FG}[4] = (0, 0, 0, 1, 0, 0)$, then $(\mathrm{FG}[1]\,\mathrm{OR}\text{-}\mathrm{FG}[4]) = (1, 0, 0, 1, 0, 0)$ corresponding to a candidate 2-dim fuzzy grid $A_{3,1}^{x_1} \times A_{3,1}^{x_2}$ is generated. Then, $\mathrm{FS}(A_{3,1}^{x_1} \times A_{3,1}^{x_2}) = (\mathrm{TT}[1] \cdot \mathrm{TT}[4]) = [\mu_{3,1}^{x_1}(t_{1_1}) \cdot \mu_{3,1}^{x_2}(t_{1_2}) + \mu_{3,1}^{x_1}(t_{2_1}) \cdot \mu_{3,1}^{x_2}(t_{2_2})]/2$ is obtained to com-

Table 1
Initial table FGTTFS for an example

Fuzzy grid	FG						TT		FS
	$A^{x_1}_{3,1}$	$A^{x_1}_{3,2}$	$A^{x_1}_{3,3}$	$A^{x_2}_{3,1}$	$A^{x_2}_{3,2}$	$A^{x_2}_{3,3}$	t_1	t_2	
$A^{x_1}_{3,1}$	1	0	0	0	0	0	$\mu^{x_1}_{3,1}(t_{1_1})$	$\mu^{x_1}_{3,1}(t_{2_1})$	$\mathrm{FS}(A^{x_1}_{3,1})$
$A^{x_1}_{3,2}$	0	1	0	0	0	0	$\mu^{x_1}_{3,2}(t_{1_1})$	$\mu^{x_1}_{3,2}(t_{2_1})$	$\mathrm{FS}(A^{x_1}_{3,2})$
$A^{x_1}_{3,3}$	0	0	1	0	0	0	$\mu^{x_1}_{3,3}(t_{1_1})$	$\mu^{x_1}_{3,3}(t_{2_1})$	$\mathrm{FS}(A^{x_1}_{3,3})$
$A^{x_2}_{3,1}$	0	0	0	1	0	0	$\mu^{x_2}_{3,1}(t_{1_2})$	$\mu^{x_2}_{3,1}(t_{2_2})$	$\mathrm{FS}(A^{x_2}_{3,1})$
$A^{x_2}_{3,2}$	0	0	0	0	1	0	$\mu^{x_2}_{3,2}(t_{1_2})$	$\mu^{x_2}_{3,2}(t_{2_2})$	$\mathrm{FS}(A^{x_2}_{3,2})$
$A^{x_2}_{3,3}$	0	0	0	0	0	1	$\mu^{x_2}_{3,3}(t_{1_2})$	$\mu^{x_2}_{3,3}(t_{2_2})$	$\mathrm{FS}(A^{x_2}_{3,3})$

pare with the min FS. However, any two linguistic values defined in the same attribute cannot be contained in the same candidate k-dim fuzzy grid ($k \geqslant 2$). Therefore, for example, $(1,1,0,0,0,0)$ and $(0,0,0,1,0,1)$ are invalid.

In the Apriori algorithm, a candidate k-itemset can be derived by joining two frequent $(k-1)$-itemsets, and these two frequent itemsets share $(k-2)$ items. Similarly, a candidate k-dim $(2 \leqslant k \leqslant d)$ fuzzy grid is derived by joining two frequent $(k-1)$-dim fuzzy grids, and these two frequent grids share $(k-2)$ linguistic values. For example, we can use $A^{x_1}_{3,2} \times A^{x_2}_{3,1}$ and $A^{x_1}_{3,2} \times A^{x_3}_{3,3}$ to generate the candidate 3-dim fuzzy grid $A^{x_1}_{3,2} \times A^{x_2}_{3,1} \times A^{x_3}_{3,3}$ because $A^{x_1}_{3,2} \times A^{x_2}_{3,1}$ and $A^{x_1}_{3,2} \times A^{x_3}_{3,3}$ share $A^{x_1}_{3,2}$. However, $A^{x_1}_{3,2} \times A^{x_2}_{3,1} \times A^{x_3}_{3,3}$ can also be generated by joining $A^{x_1}_{3,2} \times A^{x_2}_{3,1}$ to $A^{x_2}_{3,1} \times A^{x_3}_{3,3}$. This implies that we must select one of many possible combinations to avoid redundant computations. To cope with this problem, the method we adopt here is that if there exist integers $1 \leqslant e_1 < e_2 < \cdots < e_k$, such that $\mathrm{FG}[u,e_1] = \mathrm{FG}[u,e_2] = \cdots = \mathrm{FG}[u,e_{k-2}] = \mathrm{FG}[u,e_{k-1}] = 1$ and $\mathrm{FG}[v,e_1] = \mathrm{FG}[v,e_2] = \cdots = \mathrm{FG}[v,e_{k-2}] = \mathrm{FG}[v,e_k] = 1$, where $\mathrm{FG}[u]$ and $\mathrm{FG}[v]$ correspond to frequent $(k-1)$-dim fuzzy grids and $\mathrm{FG}[v,e_k]$ stands for the e_kth element of the vth row of FG, then $\mathrm{FG}[u]$ and $\mathrm{FG}[v]$ can be paired to generate a candidate k-dim fuzzy grid.

4.2. Determining effective fuzzy rules

The general type of one fuzzy classification rule denoted by R is stated as Eq. (5).

$$\text{Rule } R : A^{x_1}_{K,i_1} \times A^{x_2}_{K,i_2} \times \cdots \times A^{x_{k-1}}_{K,i_{k-1}} \times A^{x_k}_{K,i_k} \Rightarrow A^{x_\alpha}_{C,i_\alpha} \quad \text{with } \mathrm{CF}(R) \tag{5}$$

where x_α $(1 \leqslant \alpha \leqslant d)$ is the class label and $\mathrm{CF}(R)$ is the certainty grade of R. The above rule can be interpreted as: if x_1 is $A^{x_1}_{K,i_1}$ and x_2 is $A^{x_2}_{K,i_2}$ and $\cdots$ and x_k is $A^{x_k}_{K,i_k}$, then x_α is $A^{x_\alpha}_{C,i_\alpha}$ with certainty grade $\mathrm{CF}(R)$. The left-hand-side of "⇒" is the antecedence of R, and the right-hand-side is the consequence. Since $(A^{x_1}_{K,i_1} \times A^{x_2}_{K,i_2} \times \cdots \times A^{x_{k-1}}_{K,i_{k-1}} \times A^{x_k}_{K,i_k} \times A^{x_\alpha}_{C,i_\alpha}) \subseteq (A^{x_1}_{K,i_1} \times A^{x_2}_{K,i_2} \times \cdots \times A^{x_{k-1}}_{K,ik-1} \times A^{x_k}_{K,i_k})$ holds, R can be generated by $A^{x_1}_{K,i_1} \times A^{x_2}_{K,i_2} \times \cdots \times A^{x_{k-1}}_{K,i_{k-1}} \times A^{x_k}_{K,i_k} \times A^{x_\alpha}_{C,i_\alpha}$ and $A^{x_1}_{K,i_1} \times A^{x_2}_{K,i_2} \times \cdots \times A^{x_{k-1}}_{K,i_{k-1}} \times A^{x_k}_{K,i_k}$. In addition, the fuzzy confidence (Ishibuchi et al., 2001a; Ishibuchi et al., 2001b; Hu et al., 2002) of R (i.e., $\mathrm{FC}(R)$) is defined as follows:

$$\begin{aligned}\mathrm{FC}(R) = \mathrm{FS}(&A^{x_1}_{K,i_1} \times A^{x_2}_{K,i_2} \times \cdots \times A^{x_{k-1}}_{K,i_{k-1}} \\ &\times A^{x_k}_{K,i_k} \times A^{x_\alpha}_{C,i_\alpha})/\mathrm{FS}(A^{x_1}_{K,i_1} \times A^{x_2}_{K,i_2} \\ &\times \cdots \times A^{x_{k-1}}_{K,i_{k-1}} \times A^{x_k}_{K,i_k})\end{aligned} \tag{6}$$

When $\mathrm{FC}(R)$ is larger than or equal to the user-specified min FC, we can say that R is effective. $\mathrm{FC}(R)$ can further be used as the grade of certainty of R (i.e., $\mathrm{CF}(R) = \mathrm{FC}(R)$).

We also use Boolean operations to obtain the antecedence and consequence of each rule. For example, if there exists $\mathrm{FG}[u] = (1,0,0,0,0,0)$ and $\mathrm{FG}[v] = (1,0,0,1,0,0)$ corresponding to frequent fuzzy grids L_u and L_v, where $L_v \subset L_u$, respectively; then $\mathrm{FG}[v]$ AND $\mathrm{FG}[v] = (1,0,0,0,0,0)$, corresponding to the frequent fuzzy grid $A^{x_1}_{3,1}$, is generated as the antecedent part of rule R. Then, $\mathrm{FG}[u]$ XOR $\mathrm{FG}[v] = (0,0,0,1,0,0)$, which corresponds to the frequent fuzzy grid $A^{x_2}_{3,1}$, is generated to be the consequent part of rule R. Then, $\mathrm{FC}(R) = \mathrm{FS}(A^{x_1}_{3,1} \times A^{x_2}_{3,1})/\mathrm{FS}(A^{x_1}_{3,1})$ is easily obtained by Eq. (6).

However, some redundant rules must be eliminated in order to achieve compactness. If there exist two rules, say R and S, having the same consequence, and the antecedence of R is contained in that of S, then R is redundant and can be discarded, whereas S is temporarily reserved. For example, if S is "$A^{x_1}_{K_1,i_1} \times A^{x_2}_{K_2,i_2} \times \cdots \times A^{x_{k-1}}_{K_{k-1},i_{k-1}} \Rightarrow A^{x_\alpha}_{C,i_\alpha}$", then R can be eliminated. This is because the minimization of the number of antecedent conditions should be considered.

Ishibuchi et al. (1999) and Nozaki et al. (1996) further demonstrated that the performance of fuzzy rule-based systems can be improved by adjusting the grade of certainty of each rule. Therefore, it is possible to improve the classification ability of our methods by incorporating the adaptive rules proposed by Nozaki et al. (1996) into the proposed learning algorithm. Now, we determine the class label of t_p by applying fuzzy rules derived by the proposed learning algorithm. Without loss of generality, if the antecedent part of a fuzzy associative classification rule R_τ is $A^{x_1}_{K_1,j_1} \times A^{x_2}_{K_2,i_2} \times \cdots \times A^{x_z}_{K_\tau,i_\tau}$, then we can calculate ω_τ of R_τ as Eq. (7).

$$\omega_\tau = \mu^{x_1}_{K_1,j_1}(t_{p_1}) \cdot \mu^{x_2}_{K_2,j_2}(t_{p_2}) \cdot \ldots \cdot \mu^{x_\tau}_{K,i_\tau}(t_{p_\tau}) \cdot \mathrm{FC}(R_\tau) \tag{7}$$

Then t_p can be determined to categorize to the class label which is the consequent part of R_β, when

$$\omega_\beta = \max_j \{\omega_j | R_j \in \mathrm{TR}\} \tag{8}$$

where TR is the set of fuzzy rules generated by the proposed learning algorithm. The class label of t_p is thus determined and adaptive rules can be employed to adjust the fuzzy confidence of the "firing" rule R_β. That is, if t_p is correctly classified then ω_β is increased; otherwise, ω_β is decreased.

4.3. The proposed learning algorithm

As we have mentioned above, both min FS and min FC are user-specified. However, it is difficult for users to appropriately give these two thresholds for each classification problem. Therefore GA is incorporated into the proposed algorithm to automatically determine the above-mentioned parameters (i.e., min FS and min FC).

As we have mentioned above, the type of a binary chromosome is composed of min FS and min FC. That is, as shown in Fig. 4, the jth chromosome is actually denoted by $s^{(j)}c^{(j)}$, with total length $(|s^{(j)}| + |c^{(j)}|)$, where $|s^{(j)}|$ and $|c^{(j)}|$ are the lengths of $s^{(j)}$ and $c^{(j)}$, respectively. $s^{(j)}$ (or $c^{(j)}$) can be decoded by transforming the binary representation to an integer number, and then this number is divided by $2^{s^{(j)}}$ (or $2^{c^{(j)}}$). For example, if $s^{(j)} = 10010$ (i.e. $|s^{(j)}| = 5$), then the corresponding min FS can be obtained by transformed $s^{(j)}$ to 0.5625 (i.e., 18/32). From Fig. 4, we can also see that the two-point crossover operator (Rooij et al., 1996) is used for exchanging partial information between two selected chromosomes, and two new chromosomes are thus generated at the same time to replace their parents. Two crossover points are randomly selected and lie in $s^{(j)}$ and $c^{(j)}$, respectively. It should be noted that both $|s^{(j)}|$ and $|c^{(j)}|$ should be sufficiently large (e.g., $|s^{(j)}| = |c^{(j)}| = 10$); otherwise, it may be unnecessary to employ the GA to find min FS and min FC by the proposed chromosome.

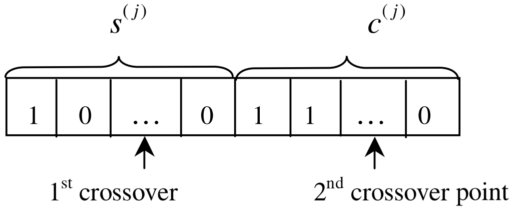

Fig. 4. The jth chromosome.

In each generation of the GA, the fitness value of each chromosome can be obtained. Moreover, the fitness value $f(V^{(j)})$ of the jth chromosome is formulated as follows:

$$f(V^{(j)}) = W_{\mathrm{CAR}} \cdot \mathrm{CAR}(V^{(j)}) - W_V \cdot |V^{(j)}| \tag{9}$$

where $V^{(j)}$ denotes a set consisting of the effective fuzzy classification rules obtained by $s^{(j)}c^{(j)}$, and W_{CAR} and W_V are relative weights of the classification accuracy rate by $V^{(j)}$ (i.e., $\mathrm{CAR}(V^{(j)})$) and the number of fuzzy rules in $V^{(j)}$ (i.e., $|V^{(j)}|$), respectively. The chromosome that has the maximum fitness value in the final generation is further used to examine the classification performance of the proposed method. That is, the acquisition of a compact fuzzy rule set with high classification accuracy rate is taken into account in the overall objective. In general, $0 < W_V \ll W_{\mathrm{CAR}}$ holds since the classification power of a classification system is more important than its compactness (Ishibuchi et al., 1995).

The proposed learning algorithm incorporated with the GA is detailed presented as follows:

Algorithm. *Finding fuzzy classification rules based on the Apriori algorithm*

Input:

a. A set of training samples;
b. K;
c. Maximum number of iterations $J_{\max}$ for performing adaptive rules;
d. Population size N_{pop};
e. Weights W_{CAR} and W_V;
f. Crossover probability $\mathrm{Prob}_{\mathrm{c}}$;
g. Mutation probability $\mathrm{Prob}_{\mathrm{M}}$;
h. Maximum number of generations $t_{\max}$.

Output: Phase I: Discover frequent fuzzy grids; Phase II: Generate fuzzy classification rules

Method:

Step1. Initialization

Generate an initial population containing N_{pop} chromosomes and insert those chromosomes into the current population P. Each gene is randomly assigned to one or zero, with probability of 0.5.

Step 2. Perform the simple fuzzy partition

Step 3. Scan the training samples, and construct FGTTFS

Step 4. Compute the fitness

4-1. Compute the fitness value for each chromosome in P
Decode $s^{(j)}c^{(j)}$ $(1 \leqslant j \leqslant N_{\mathrm{pop}})$, and generate $V^{(j)}$ through Step 5–8. Then, $f(V^{(j)})$ can be obtained by $V^{(j)}$.

4-2. Find the chromosomes denoted by $L_{\max}^{(P)}$ that has $f_{\max}(P)$ which denotes the maximum fitness value in P. Set NP to ϕ, denoting the next generation; then go to Step 9.

Step 5. Generate frequent fuzzy grids

Generate frequent k-dim $(k \geqslant 2)$ fuzzy grids, whose fuzzy support is larger than or equal to the min FS, from $(k-1)$-dim frequent fuzzy grids.

Step 6. Generate fuzzy classification rules

Generate the antecedence and the consequence of an effective fuzzy rule, whose fuzzy confidence is larger than or equal to the min FC, using the methods introduced in Section 4.2.

Step 7. Reduce redundant rules

Step 8. Employ adaptive rules to adjust fuzzy confidences (Nozaki et al., 1996)

Set J to be zero.
Repeat
$J = J + 1$
For each training sample t_p do
a. Find the "firing" fuzzy rule R_β which is marked in FG.
b. If t_p is correctly classified then

$$\omega_\beta = \omega_\beta + \eta_1 \cdot (1 - \omega_\beta) \tag{10}$$

otherwise, as

$$\omega_\beta = \omega_\beta - \eta_2 \cdot \omega_\beta. \tag{11}$$

where η_1 and η_2 are learning rates.
End
Until $J = J_{\max}$

Step 9. Selection

Select $N_{\mathrm{pop}}/2$ pairs of chromosomes from the current population P. The selection probability $\mathrm{Prob}(V^{(j)})$ of the jth $(1 \leqslant j \leqslant N_{\mathrm{pop}})$ chromosome is as follows:

$$\mathrm{Prob}(V^{(j)}) = \frac{f(V^{(j)}) - f_{\min}(P)}{\sum_{V^{(i)} \in P} [f(V^{(i)}) - f_{\min}(P)]} \tag{12}$$

where $f_{\min}(P)$ denotes the minimum fitness value in P.

Step 10. Crossover

For each selected pair, perform the two-point crossover operation with probability $\mathrm{Prob}_{\mathrm{c}}$ and insert two new chromosomes to NP.

Step 11. Mutation

For each gene of the newly generated chromosomes in NP, perform the mutation operation (i.e., change each gene's value either from one to zero or from zero to one) with probability $\mathrm{Prob}_{\mathrm{M}}$.

Step 12. Elitist strategy

12-1. From NP, find $L_{\min}^{(NP)}$ that has $f_{\min}(NP)$, where $f_{\min}(NP)$ represents the minimum fitness value in NP.

12-2. Replace $L_{\min}^{(NP)}$ with $L_{\max}^{(P)}$. That is, the best performing chromosome is retained and inserted into the next generation without change (Rooij et al., 1996).

Step 13. Termination test

If $t_{\max}$ generations have been generated, then terminate the execution of the whole learning algorithm, and $L_{\max}^{(NP)}$ with $f_{\max}(NP)$ is the chromosome that is used to test the classification performance of the proposed method; otherwise replace P with NP and return to Step 4.

Nozaki et al. (1996) also suggested that the learning rates should be specified as $0 < \eta_1 \ll \eta_2 < 1$. In later simulations, $\eta_1 = 0.001$, $\eta_2 = 0.1$ and $J_{\max} = 100$ are used. In the following section, simulation results are presented to demonstrate the effectiveness of the proposed learning algorithm.

5. Experiments

To examine the performance of the proposed learning algorithm for testing samples, we perform the leave-one-out technique, which is an almost unbiased estimator of the true error rate of a classifier (Weiss and Kulikowski, 1991). Based on the leave-one-out technique, we try to make a comparison between the proposed learning algorithm and other fuzzy or non-fuzzy classification methods.

In Sections 4.1 and 4.2, we employ the proposed learning algorithm to discover fuzzy classification rules from the iris data and the appendicitis data, respectively.

5.1. Experiment 1

The iris data consists of three classes (Class 1: Iris setosa, Class 2: Iris versicolor and Class 3: Iris virginica) and each class consists of fifty samples with four dimensions. Suppose that attribute x_1 is the sepal length, attribute x_2 is the sepal width, attribute x_3 is the petal length, attribute x_4 is the petal width, and attribute x_5 is the class label (i.e., $d = 5$) to which $t_p = (t_{p_1}, t_{p_2}, \ldots, t_{p_5})$ $(1 \leqslant p \leqslant 150)$ belongs. The pairs $(\mathrm{ma}, \mathrm{mi})$ for x_1, x_2, x_3 and x_4 are $(79, 43)$, $(44, 20)$, $(69, 10)$, and $(25, 1)$ respectively. Of course, since t_{p_i} is the value of the pth training sample with respect to the ith attribute x_i, the range of t_{p_i} $(1 \leqslant i \leqslant 4)$ is equivalent to x_i. It should be noted that only three linguistic values can be defined in x_5; they are $A_{3,1}^{\text{class label}}$: "Class 1", $A_{3,2}^{\text{class label}}$: "Class 2", and $A_{3,3}^{\text{class label}}$: "Class 3".

Parameter specifications used in the proposed method are as follows:

Maximum iterations $J_{\max} = 100$
Population size $N_{\mathrm{pop}} = 30$
$|s| = |c| = 10$
$W_{\mathrm{CAR}} = 10$
$W_V = 1$
Crossover probability $\mathrm{Prob_c} = 1.0$
Mutation probability $\mathrm{Prob_M} = 0.01$

Maximum number of generations $t_{\max} = 50$ Simulation results with various values of K (i.e., $K = 2$, 3, 4, 5 and 6) for each attribute except x_5 are shown as Table 2. From this table, we can see that the best result with the classification accuracy rate 96.00% and 12.13 fuzzy if–then rules on an average can be obtained by the proposed method when $K = 5$. It seems that K is also not an influential factor when it is larger than 2.

Some significant classification methods of fuzzy if–then rule extraction using simple fuzzy partition methods have been proposed, such as the simple-fuzzy-grid method (Ishibuchi et al., 1992), the multi-rule-table method (Ishibuchi et al., 1992), the pruning method (Nozaki et al., 1996), and the GA-based method (Ishibuchi et al., 1995). The similarity between the proposed method and these known fuzzy classification methods is that they all use the concepts of "clusters define rules" (Kosko, 1992). That is, if one fuzzy grid whose fuzzy support is larger than zero, then this fuzzy grid may be taken into account in the generation of fuzzy rules.

The simulation results of the aforementioned methods demonstrated by Nozaki et al. (1996) are shown as Table 3. The proposed method with the

Table 2
Simulation results of the proposed method by the leave-one-out technique

K	Rate (%)	Average number of rules
2	76.00	15.35
3	95.33	11.87
4	96.00	18.27
5	96.00	12.13
6	94.67	10.09

Table 3
Simulation results of different fuzzy classification methods by the leave-one-out technique

Method	Rate (%)	Average number of rules
The proposed method	96.00	12.13
Simple-fuzzy-grid	96.67	294.50
Multi-rule-table	94.67	691.11
Pruning	93.33	42.62
GA-based	94.67	12.90

best result is also shown in this table. In comparison with these known methods, we can see that the proposed method can use the minimum average number of fuzzy rules to perform classification. From the viewpoint of classification rates, although the proposed method performs slightly worse than the simple-fuzzy-grid method, it outperforms the other methods, which may suffer from overfitting to the training samples.

Previously, the classification accuracy rate of the nine fuzzy methods (i.e., fuzzy integral with perceptron criterion, fuzzy integral with quadratic criterion, minimum operator, fast heuristic search with Sugeno integral, simulated annealing with Sugeno integral, fuzzy k-nearest neighbor, fuzzy c-means, fuzzy c-means for histograms and hierarchical fuzzy c-means) were reported by Grabisch and Dispot (1992). In addition, the 10 non-fuzzy methods (i.e., the linear discriminant, the quadratic discriminant, the nearest neighbor, the Bayes independence, the Bayes second order, the neural networks with the BP algorithm, the neural networks with the ODE algorithm, the PVM rule, the optimal rule with size two and the CART tree) for the iris data estimated by the leaving-one-out technique were reported by Weiss and Kulikowski (1991). The classification results of various classification methods are summarized in Table 4. The average classification accuracy rates for the nine fuzzy methods and ten non-fuzzy methods are 94.47% and 95.20%, respectively. Although the best result (i.e., 96.00%) from the proposed method is slightly worse than those of the above-mentioned methods, it is clear that the best result of the proposed method outperforms these two average results.

5.2. Experiment 2

The appendicitis data consists of 106 samples classified into two classes with seven attributes. The same parameter specifications used in last section of the GA are also employed here. Based on the performance summarized in Table 2 according to various K, $K = 5$ is thus considered in the classification of the appendicitis data. By using the above-mentioned parameter specifications, the classification accuracy rate obtained by the proposed learning algorithm is 87.7% using an average of 42.78 fuzzy if–then rules.

The classification accuracy rate of aforementioned various methods for the appendicitis data as estimated by the leave-one-out technique were also reported by Grabisch and Dispot (1992) and Weiss and Kulikowski (1991), respectively. The classification results are summarized in Table 5.

Table 4
Classification accuracy rates of various classification methods for the iris data

Fuzzy methods				
Perceptron criterion	Quadratic criterion	Minimum operator	Fast heuristic search	Simulated annealing
95.33%	96.67%	96.00%	92.00%	91.33%
Fuzzy k-nearest neighbor	Fuzzy c-means	Fuzzy c-means for histograms	Hierarchical fuzzy c-means	
96.67%	93.33%	93.33%	95.33%	
Non-fuzzy methods				
Linear discriminant	Quadratic discriminant	Nearest neighbor	Bayes independence	Bayes second order
98.00%	97.33%	96.00%	93.33%	84.00%
BP algorithm	PVM rule	ODE algorithm	Optimal rule with size 2	CART tree
96.67%	97.33%	96.00%	98.00%	95.33%

Table 5
Classification accuracy rates of various classification methods for the appendicitis data

Fuzzy methods				
Perceptron criterion	Quadratic criterion	Minimum operator	Fast heuristic search	Simulated annealing
79.2%	86.8%	86.8%	84.9%	81.1%
Fuzzy *k*-nearest neighbor	Fuzzy *c*-means	Fuzzy *c*-means for histograms	Hierarchical fuzzy *c*-means	
86.8%	71.2%	78.3%	80.2%	
Non-fuzzy methods				
Linear discriminant	Quadratic discriminant	Nearest neighbor	Bayes independence	Bayes second order
86.8%	73.6%	82.1%	83.0%	81.1%
BP algorithm	PVM rule	ODE algorithm	Optimal rule with size 2	CART tree
85.8%	86.8%	89.6%	89.6%	84.9%

We can see that the best classification accuracy rates for the nine fuzzy methods and 10 non-fuzzy methods are 86.8% and 89.69%, respectively. It is obvious that the classification rate of the proposed method slightly outperforms the best rate of the nine fuzzy methods, and is slightly worse than the best rate of the 10 non-fuzzy methods. In addition, the result of the proposed method also outperforms the average rates over the nine fuzzy methods (i.e., 82.1%) and 10 non-fuzzy methods (i.e., 84.3%), respectively.

On the other hand, a significant genetics-based learning method proposed by Ishibuchi et al. (1999) was employed to extract fuzzy if–then rules from various classification data. Using the leave-one-out technique to examine the performance of the genetics-based learning method for the appendicitis data, the classification accuracy rate was 84.9% using 100 fuzzy if–then rules. However, due to the intrinsic limits of this method, it is difficult to minimize the number of fuzzy rules.

6. Discussions

The performance of the proposed method is examined by the iris data and appendicitis data. From Table 2, we may conclude that K is not an influential factor when it is larger than 2.

We also find that the fuzzy classification methods proposed by Ishibuchi et al. (2001a,b) employed each $(d-1)$-fuzzy grid, which does not contain the dimension of the class label, as an antecedence of one fuzzy rule, whose consequence can be determined by computing the fuzzy confidence for each class label. The class label with the maximum fuzzy confidence then serves as the consequence. However, since there are no min FS and min FC to determine the frequent fuzzy grids and the effective fuzzy rules, respectively, thus the method of Ishibuchi et al. may suffer from the curse of dimensionality (Ishibuchi et al., 1999). For example, in the appendicitis data (i.e., $d = 8$), if each dimension is partitioned into five various linguistic values, then a large number (i.e., 5^7) fuzzy rules will be generated. However, based on the well-known apriori algorithm, our method searches for a compact set of fuzzy rules by using the GA to automatically find the appropriately min FS and min FC.

Since fuzzy knowledge representation can facilitate interaction of the expert system and users (Zimmermann, 1991), it is feasible to extend the proposed learning algorithm to discover other types of fuzzy association rules to ease the fuzzy knowledge acquisition bottleneck in building prototype expert systems. That is, the proposed algorithm may be further viewed as a knowledge acquisition tool to discover fuzzy association rules to perform the market basket analysis, which can help users make decisions. For example, in a supermarket, the manager may design a particular store layouts using the analytic results (Han and Kamber, 2001).

In the fuzzy data mining, Hong et al. (2001) presented discussions on the relations between the computation time and the number of rules. Their study can further provide useful suggestions to further improve the proposed method.

7. Conclusions

In this paper, we propose a learning algorithm that can find fuzzy rules for classification problems based on the processing of the Apriori algorithm. Significantly, our method tries to find a compact set of fuzzy rules by using the GA to automatically find the appropriately min FS and min FC.

Simulation results on the iris data and the appendicitis data demonstrate that the classification accuracy rates of the proposed method are comparable to the other fuzzy or non-fuzzy methods. Thus, the goal of acquiring an effectively compact set of fuzzy rules for classification problems can be achieved.

References

Agrawal, R., Imielinski, T., Swami, A., 1993. Database mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering 5 (6), 914–925.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I., 1996. Fast discovery of association rules. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in knowledge discovery and data mining. AAAI Press, Menlo Park, pp. 307–328.

Anderson, E., 1935. The irises of the gaspe peninsula. Bull. Amer. Iris Soc. 59, 2–5.

Berry, M., Linoff, G., 1997. Data Mining Techniques: for Marketing, Sales, and Customer Support. John Wiley & Sons, New York.

Chen, S.M., Jong, W.T., 1997. Fuzzy query translation for relational database systems. IEEE Transactions on Systems, Man, and Cybernetics 27 (4), 714–721.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, MA.

Grabisch, M., Dispot, F., 1992. A comparison of some methods of fuzzy classification on real data. In: Proceedings of the 2nd International Conference on Fuzzy Logic and Neural Networks, Iizuka, Japan, pp. 659–662.

Han, J.W., Kamber, M., 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco.

Hong, T.P., Kuo, C.S., Chi, S.C., 2001. Trade-off between computation time and number of rules for fuzzy mining from quantitative data. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 9 (5), 587–604.

Hu, Y.C., Chen, R.S., Tzeng, G.H., 2002. Mining fuzzy association rules for classification problems. Computers and Industrial Engineering 43 (4), 735–750.

Ishibuchi, H., Nozaki, K., Tanaka, H., 1992. Distributed representation of fuzzy rules and its application to pattern classification. Fuzzy Sets and Systems 52 (1), 21–32.

Ishibuchi, H., Nozaki, K., Yamamoto, N., Tanaka, H., 1995. Selecting fuzzy if–then rules for classification problems using genetic algorithms. IEEE Transactions on Fuzzy Systems 3 (3), 260–270.

Ishibuchi, H., Nakashima, T., Murata, T., 1999. Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems. IEEE Transactions on Systems, Man, and Cybernetics 29 (5), 601–618.

Ishibuchi, H., Yamamoto, T., Nakashima, T., 2001a. Fuzzy data mining: effect of fuzzy discretization. In: Proceedings of the 1st IEEE International Conference on Data Mining, San Jose, USA, pp. 241–248.

Ishibuchi, H., Nakashima, T., Yamamoto, T., 2001b. Fuzzy association rules for handling continuous attributes. In: Proceedings of IEEE International Symposium on Industrial Electronics, Pusan, Korea, pp. 118–121.

Jang, J.S.R., 1993. ANFIS: adaptive-network-based fuzzy inference systems. IEEE Transactions on Systems, Man, and Cybernetics 23 (3), 665–685.

Jang, J.S.R., Sun, C.T., 1995. Neuro-fuzzy modeling and control. Proceedings of the IEEE 83 (3), 378–406.

Kosko, B., 1992. Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence. Prentice Hall, NJ.

Nozaki, K., Ishibuchi, H., Tanaka, H., 1996. Adaptive fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems 4 (3), 238–250.

Pedrycz, W., Gomide, F., 1998. An Introduction to Fuzzy Sets: Analysis and Design. MIT Press, Cambridge.

Rooij, A.J.F., Jain, L.C., Johnson, R.P., 1996. Neural Network Training Using Genetic Algorithms. World Scientific, Singapore.

Wang, L.X., Mendel, J.M., 1992. Generating fuzzy rules by learning from examples. IEEE Transactions on Systems, Man, and Cybernetics 22 (6), 1414–1427.

Weiss, S.M., Kulikowski, C.A., 1991. Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann, CA.

Zadeh, L.A., 1975a. The concept of a linguistic variable and its application to approximate reasoning. Information Science (part 1) 8 (3), 199–249.

Zadeh, L.A., 1975b. The concept of a linguistic variable and its application to approximate reasoning. Information Science 8 (4), 301–357.

Zadeh, L.A., 1976. The concept of a linguistic variable and its application to approximate reasoning. Information Science (part 3) 9 (1), 43–80.

Zimmermann, H.-J., 1991. Fuzzy Sets, Decision Making, and Expert Systems. Kluwer, Boston.