Cybernetics and Systems: An International Journal, 34: 33–57, 2003

0196-9722/03 $12.00 + .00
DOI: 10.1080/01969720390168522

Taylor & Francis
Taylor & Francis Group

A NEW METHOD TO GENERATE FUZZY RULES FROM RELATIONAL DATABASE SYSTEMS FOR ESTIMATING NULL VALUES

SHYI-MING CHEN

Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology,
Taipei, Taiwan, R.O.C.

SHIH-WEI LEE

Department of Electronic Engineering,
National Taiwan University of Science and Technology,
Taipei, Taiwan, R.O.C.

Fuzzy decision trees can be used to generate fuzzy rules from training instances to deal with forecasting and classification problems. We propose a new method to construct fuzzy decision trees from relational database systems and to generate fuzzy rules from the constructed fuzzy decision trees for estimating null values, where the weights of attributes are used to derive the values of certainty factors of the generated fuzzy rules. We use the concept of "coefficient of determination" of the statistics to derive the weights of the attributes in relational database systems and use the normalized weights of the attributes to derive the values of certainty factors of the generated fuzzy rules. Furthermore, we also use regression equations of the statistics to construct a complete fuzzy decision tree for generating better fuzzy rules. The proposed method obtains a higher average estimated accuracy rate than the existing methods for estimating null values in relational database systems.

This work was supported in part by the National Science Council, Republic of China, under Grant NSC 90-2213-E-011-053.

Address correspondence to Professor Shyi-Ming Chen, Ph.D., Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C.

INTRODUCTION

Generally speaking, the knowledge residing in the knowledge base of a rule-based system is obtained from the process of knowledge acquisition. In recent years, many researchers focused on the research topic of automatically generating rules from training instances (Chang and Chen 2000; Chen and Lin 2000; Chen et al. 2001; Chen and Chen 2000; Hunt et al. 1966; Jeng and Liang 1993; Lee and Chen 2001; Lin and Chen 2000; Quinlan 1979, 1986; Sudkamp and Hammell 1994; Wang and Mendel 1992; Wu and Chen 1999; Yasdi 1991). The decision tree method is a well-known method of inductive learning (Quinlan 1979). The decision tree method can generate useful rules from a set of training data. The ID3 algorithm (Quinlan 1979, 1986) and the CLS algorithm (Hunt et al. 1966) are useful to construct decision trees for rules generation. Chang and Chen (2000) presented a method to generate fuzzy rules from numerical data based on the exclusion of attribute terms. Chen and Yeh (1997) presented a fuzzy concept learning system (FCLS) algorithm for generating fuzzy rules from relational database systems for estimating null values. Chen and Lin (2000) presented a method for constructing fuzzy decision trees and generated fuzzy classification rules from training examples. Chen et al. (2001) presented a method for generating fuzzy rules from numerical data for handling classification problems. Chen and Chen (2000) presented a method to generate fuzzy rules for fuzzy classification systems. Lin and Chen (2000) presented a method to generate weighted fuzzy rules from training data. Sudkamp and Hammell (1994) presented the techniques of interpolation, completion, and learning fuzzy rules. Wang and Mendel (1992) presented a method for constructing membership functions and fuzzy rules from training examples. Wu and Chen (1999) presented a method for constructing fuzzy rules and membership functions from training examples. Yeung and Tsang (1995, 1997) proposed the concepts of weighted fuzzy rules and weighted inference techniques, where weighted fuzzy rules consider the importance of attributes appearing in the antecedent portions of the rules.

In this article, we extend the FCLS algorithm we presented previously (Chen and Yeh 1997) to present a new method to generate fuzzy rules from relational database systems for estimating null values, where the attributes appearing in the antecedent portions of the generated fuzzy rules have different weights. Furthermore, we also apply the weights of the attributes to derive the certainty factor (CF) value of each generated

fuzzy rule to generate better fuzzy rules for estimating null values in relational database systems. First, we use the concept of "coefficient of determination" of the statistics (Berenson et al. 1983; Mendenhall and Beaver 1994) to calculate the coefficient of determination of related attributes in relational database systems, and then we normalize them and use the normalized values as the weights of the attributes to derive the certainty factor values of the generated fuzzy rules. We also apply the regression equations of the statistics to present a method to derive the CF values of the hypothetical certainty factor (HCF) nodes by constructing a complete fuzzy decision tree for generating better fuzzy rules (based on Mendenhall and Beaver 1994; Neter et al. 1999). The proposed method can obtain a higher average estimated accuracy rate than the existing methods.

BASIC CONCEPTS OF FUZZY SET THEORY

In 1965, Zadeh proposed the theory of fuzzy sets (Zadeh 1965). A fuzzy set A of the universe of discourse U can be described by a membership function μ_A, where $\mu_A : U \rightarrow [0, 1]$. Let U be the universe of discourse, $U = \{u_1, u_2, \ldots, u_n\}$. A fuzzy set A of the universe of discourse U can be represented as follows:

$$A = \mu_A(u_1)/u_1 + \mu_A(u_2)/u_2 + \cdots + \mu_A(u_n)/u_n \tag{1}$$

where $\mu_A(u_i)$ indicates the grade of membership of u_i in the fuzzy set A, $\mu_A(u_i) \in [0, 1]$, and $1 \leq i \leq n$. If the universe of discourse U is an infinite set, then the fuzzy set A can be expressed as follows:

$$A = \int_U \mu_A(u)/u \qquad u \in U. \tag{2}$$

Let A be a triangular fuzzy set of the universe of discourse U:

$$A = \int_a^b \left(\frac{x-a}{b-a}\right)\Big/x + \int_b^c \left(\frac{c-x}{c-b}\right)\Big/x \qquad \forall x \in U. \tag{3}$$

The membership function curve of the triangular fuzzy set A is shown in Figure 1.

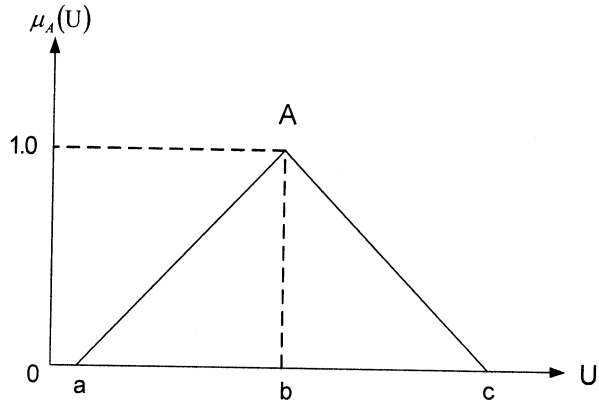


Figure 1. A triangular fuzzy set.

The triangular fuzzy set A shown in Figure 1 can be parametrized as (a, b, c), where b is the center of the triangular fuzzy set A, and a and c are the left vertex and the right vertex, respectively, of the triangular fuzzy set A. Figure 2 shows a trapezoidal fuzzy set A, where the trapezoidal fuzzy set A can be parametrized as (a, b, c, d).

We (Chen 1994) presented the defuzzification techniques of trapezoidal fuzzy sets based on previous work (Kandel 1986). Let A be a trapezoidal fuzzy set, where $A = (a, b, c, d)$. Then, the defuzzified value DEF(A) of the trapezoidal fuzzy set A is as follows:

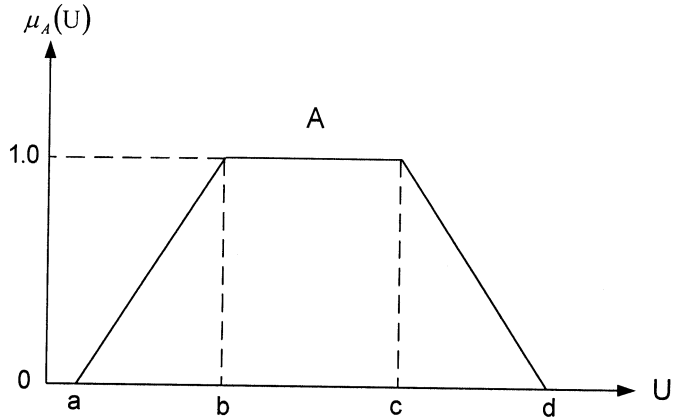


Figure 2. A trapezoidal fuzzy set.

$$\mathrm{DEF}(A) = \frac{a + b + c + d}{4}. \tag{4}$$

It is obvious that a triangular fuzzy set can be regarded as a special case of a trapezoidal fuzzy set. Let A be a triangular fuzzy set, where $A = (a,\ b,\ c)$. Based on previous work (Chen 1994; Kandel 1986), the defuzzified value $\mathrm{DEF}(A)$ of the triangular fuzzy set A is as follows:

$$\mathrm{DEF}(A) = \frac{a + 2b + c}{4}. \tag{5}$$

FUZZY DECISION TREES AND FUZZY RULES

The concept of fuzzy decision trees (Chen and Yeh 1997) is an extension of the concept of Quinlan's decision trees (Quinlan 1986). In a fuzzy decision tree, a nonterminal node is called a decision node. There are two kinds of terminal nodes in a fuzzy decision tree, namely CF nodes and HCF nodes, which associate with real values between zero and one. From the root node to each terminal node (CF node or HCF node) a fuzzy rule is formed. Figure 3 shows an example of a fuzzy decision tree, where the CF nodes are denoted by ◯, and the HCF nodes are denoted by ⬚; X, Y, and Z are attributes in a relational database system, and X_i, Y_j, and Z_k $(1 \le i \le n,\ 1 \le j \le m,$ and $1 \le k \le p)$ are linguistic terms represented by fuzzy sets. Consider the path $X \xrightarrow{X_1} Y \xrightarrow{Y_1} Z \xrightarrow{Z_1} CF_i$ in the fuzzy

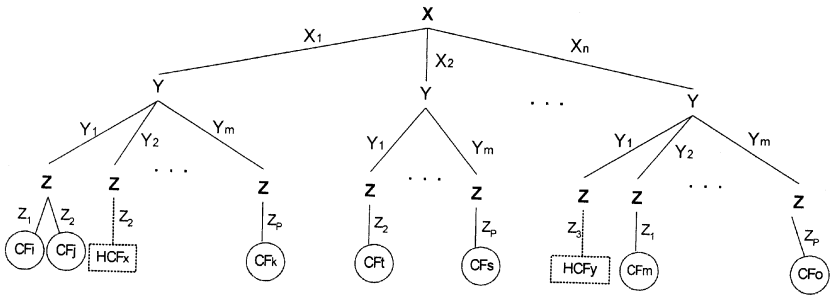


Figure 3. A fuzzy decision tree.

decision tree shown in Figure 3, where the path forms the following fuzzy rule:

IF X is X_1 AND Y is Y_1 THEN Z is Z_1 $(CF = CF_i)$

where X, Y, and Z are linguistic variables (Zadeh 1975) and X_1, Y_1, and Z_1 are linguistic terms represented by fuzzy sets; CF denotes the certainty factor of the rule and $CF_i \in [0, 1]$. The larger the value of CF_i, the more the rule is believed in. A null path is a path whose terminal node is an HCF node. A nonnull path is a path whose terminal node is a CF node. For example, the path $X \xrightarrow{X_1} Y \xrightarrow{Y_2} Z \xrightarrow{Z_2} HCF_x$ shown in Figure 3 forms a null path (Chen and Yeh 1997). It indicates there is the following virtual fuzzy rule in the knowledge base:

IF X is X_1 AND Y is Y_2 THEN Z is Z_2 $(CF = HCF_x)$.

For example, Figure 4 shows a subtree of a fuzzy decision tree. From Figure 4, we can see that there are five fuzzy rules to be generated, shown as follows:

IF A is Low **AND** B is Low **THEN** C is Low $(CF = 0.69)$
IF A is Low **AND** B is Medium **THEN** C is Medium $(CF = 0.51)$
IF A is Low **AND** B is High **THEN** C is High $(CF = 0.78)$
IF A is Medium **AND** B is Medium **THEN** C is High $(CF = 0.75)$
IF A is High **AND** B is High **THEN** C is High $(CF = 0.85)$

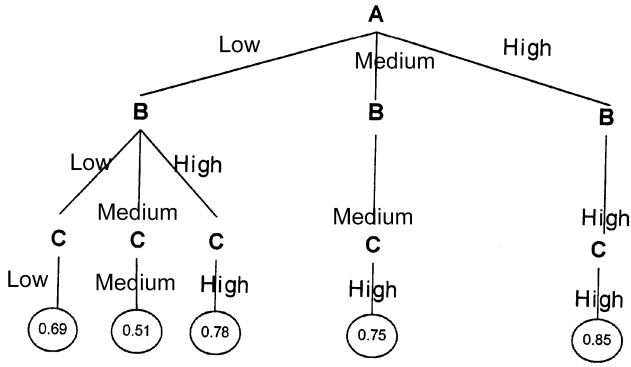


Figure 4. A subtree of a fuzzy decision tree.

FUZZINESS OF ATTRIBUTES AND CERTAINTY FACTOR VALUES OF FUZZY RULES

The FCLS algorithm we presented previously (Chen and Yeh 1997) can generate fuzzy rules from a relational database system for estimating null values. The FCLS algorithm selected an attribute that has the smallest degree of fuzziness to be a decision node. The definition of fuzziness of the degree of an attribute is reviewed from Chen and Yeh (1997) as follows.

Definition 1: Let S be a set of antecedent attributes, $S = \{X, Y, \ldots, W\}$, determining the consequent attribute Z. Let $t_j(X)$ be the value of the attribute X of the jth training instance (that is, jth tuple of a relation), then the degree of fuzziness $FA(X)$ of the attribute X is defined by

$$FA(X) = \frac{\sum_{j=1}^{c} (1 - \mu_{Xi}(t_j(X)))}{c}, \tag{6}$$

where c is the number of training instances.

In a fuzzy decision tree constructed by the FCLS algorithm (Chen and Yeh 1997), the value of every CF node in the tree is calculated as follows. Assume that there is a path $\boldsymbol{D}_1 \xrightarrow{F_1} \boldsymbol{D}_2 \xrightarrow{F_2} \boldsymbol{D}_3 \xrightarrow{F_3} (CF)$ in a fuzzy decision tree, where F_1, F_2, and F_3 are linguistic terms, then

$$CF = \min\{\text{Avg}(F_1), \text{Avg}(F_2), \text{Avg}(F_3)\}, \tag{7}$$

where $\text{Avg}(F_1)$, $\text{Avg}(F_2)$, and $\text{Avg}(F_3)$ are the average values of the linguistic terms F_1, F_2, and F_3, respectively, defined as follows:

$$\text{Avg}(F_i) = \frac{\sum_{j=1}^{s} \mu_{F_i}(t_j(\boldsymbol{D}_i))}{s}, \tag{8}$$

where $t_j(\boldsymbol{D}_i)$ denotes the values of the attribute $\boldsymbol{D}_i$ of the jth tuple of a relation, $\mu_{F_i}(t_j(\boldsymbol{D}_i))$ denotes the grade of membership of the value of the attribute $\boldsymbol{D}_i$ of the jth tuple of the relation belonging to the linguistic term F_i, s is the number of training instances (i.e., the number of tuples in the relation), and $1 \leq i \leq 3$. The FCLS algorithm we presented previously (Chen and Yeh 1997) applied formula (6) to calculate the degree of fuzziness of the attributes, and then to select the attribute that has the smallest degree of fuzziness as a decision node to sprout the

fuzzy decision tree. Furthermore, the FCLS algorithm we presented previously (Chen and Yeh 1997) applied formulas (7) and (8) to calculate the value of each CF node in the process of constructing a fuzzy decision tree. Moreover, the FCLS algorithm we presented previously sets the value of each hypothetical certainty factor node in the constructed fuzzy decision tree to 0.5. (For more details, please refer to Chen and Yeh 1997.)

A NEW ALGORITHM FOR GENERATING FUZZY RULES FROM RELATIONAL DATABASE SYSTEMS FOR ESTIMATING NULL VALUES

In the following, we present a new algorithm called the extended fuzzy concept learning system (EFCLS) algorithm to generate fuzzy rules from a relational database system for estimating null values. Consider the relation of a relational database shown in Table 1 (Chen and Yeh 1997). We can see that the relation shown in Table 1 has three attributes, namely, "Degree," "Experience," and "Salary," where the attributes Degree and Experience determine the attribute Salary (i.e., Degree and Experience are independent variables, and Salary is a dependent variable). Thus, we can apply the equations of the statistics to analyze the relationship between attributes. We also can apply the concept of coefficient of determination of the statistics (Mendenhall and Beaver 1994) to derive the weights of the attributes.

Definition 2: Assume that there are two variables X and Y, where X is an independent variable and Y is a dependent variable, then

$$\text{Coefficient of Determination from } X \text{ to } Y = \left(\frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \right)^2 \tag{9}$$

where X_i denotes the ith value of the variable X, Y_i denotes the ith value of the variable Y, $1 \leq i \leq n$, $\bar{X}$ denotes the mean value of the variable X, and $\bar{Y}$ denotes the mean value of the variable Y.

First, we assign the ranking values to the values of the attribute Degree (i.e., "Bachelor," "Master," and "Ph.D."). For example, we set

Table 1. A relation in a relational database system (Chen and Yeh 1997)

Emp-ID	Degree	Experience	Salary
S1	Ph.D.	7.2	63000
S2	Master	2	37000
S3	Bachelor	7	40000
S4	Ph.D.	1.2	47000
S5	Master	7.5	53000
S6	Bachelor	1.5	26000
S7	Bachelor	2.3	29000
S8	Ph.D.	2	50000
S9	Ph.D.	3.8	54000
S10	Bachelor	3.5	35000
S11	Master	3.5	40000
S12	Master	3.6	41000
S13	Master	10	68000
S14	Ph.D.	5	57000
S15	Bachelor	5	36000
S16	Master	6.2	50000
S17	Bachelor	0.5	23000
S18	Master	7.2	55000
S19	Master	6.5	51000
S20	Ph.D.	7.8	65000
S21	Master	8.1	64000
S22	Ph.D.	8.5	70000

the ranking value of Bachelor to 1, set the ranking value of Master to 2, and set the ranking value of Ph.D. to 3. Then, based on formula (9), we calculate the coefficients of determination from Degree to Salary and from Experience to Salary, respectively, and assign the coefficient of determination from Salary to Salary to 1. Finally, we normalize these three values, and let the three normalized values be the weights of the attributes Degree, Salary, and Experience, respectively. The weights of the attributes will be used to derive the CF values of the generated fuzzy rules.

Definition 3: Assume that there is a path in a fuzzy decision tree shown as follows:

$$\boldsymbol{D}_1 \xrightarrow{F_1} \boldsymbol{D}_2 \xrightarrow{F_2} \boldsymbol{D}_3 \xrightarrow{F_3} (CF)$$

where D_1, D_2, and D_3 are attributes, and F_1, F_2, and F_3 are linguistic terms. Then,

$$CF = \text{Avg}(F_1) \times weight1 + \text{Avg}(F_2) \times weight2 + \text{Avg}(F_3) \times weight3 \tag{10}$$

where *weight1*, *weight2*, and *weight3* are the weights of the attributes D_1, D_2, and D_3, respectively, $weight1 \in [0, 1], weight2 \in [0, 1], and$ $weight3 \in [0, 1]$.

If there exist some null paths in a constructed fuzzy decision tree, these paths will contain HCF nodes. For example, assume that there is a null path in the constructed fuzzy decision tree shown as follows:

$$\text{Degree} \xrightarrow{Master} \text{Experience} \xrightarrow{\text{H}} \text{Salary}.$$

In this situation, the above null path will generate the following virtual fuzzy rule:

IF Degree is Master **AND** Experience is H
THEN Salary is Z_1 $(CF = C_1)$,

where Z_1 is a linguistic term; C_1 is a CF value between zero and one. To calculate the value C_1 of the hypothetical CF node, we must first assume the value of the attribute Salary, and then calculate the hypothetical CF value C_1 of the generated virtual fuzzy rule. In this article, we apply the regression equations of the statistics (Mendenhall and Beaver 1994; Neter et al. 1999) to obtain the relationships among the attributes.

Definition 4: Let

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{11}$$

where Y is a dependent variable, X_1 and X_2 are independent variables, and β_0, β_1, and β_2 are the regression coefficients, where β_0 denotes the distance between the intercept of the Y axis and the origin; β_1 and β_2 denote the average varying values of Y by varying the values of X_1 and X_2, respectively. The values of β_0, β_1, and β_2 are obtained by the following equations:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} X_{1i} + \beta_2 \sum_{i=1}^{n} X_{2i} = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} X_{1i} + \beta_1 \sum_{i=1}^{n} X_{1i}^2 + \beta_2 \sum_{i=1}^{n} X_{1i}X_{2i} = \sum_{i=1}^{n} X_{1i}Y_i \tag{12}$$

$$\beta_0 \sum_{i=1}^{n} X_{2i} + \beta_1 \sum_{i=1}^{n} X_{1i}X_{21} + \beta_2 \sum_{i=1}^{n} X_{2i}^2 = \sum_{i=1}^{n} X_{2i}Y_i,$$

where X_{1i} denotes the ith data of the variable X_1, X_{2i} denotes the ith data of the variable X_2, Y_i denotes the ith data of the variable Y, $1 \leq i \leq n$, and n is the number of data.

The proposed EFCLS algorithm is now presented as follows:

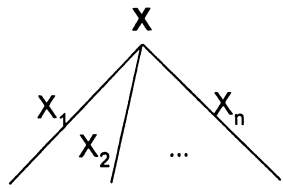
EFCLS Algorithm

Step 1: Apply formula (9) to calculate the coefficient of determination between attributes to obtain the weights of the attributes.

Step 2: Fuzzify the relation into a fuzzy relation.

Step 3: Select an attribute among the set S of antecedent attributes which has the smallest degree of fuzziness. Assume that attribute $\boldsymbol{X}$ has the smallest degree of fuzziness, then partition the set $\boldsymbol{T}$ of the training instances into subsets $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, ..., and $\boldsymbol{T}_n$ according to the fuzzy domain $\{X_1, X_2, \ldots, X_n\}$ of the attribute $\boldsymbol{X}$, respectively. Compute the average value $\text{Avg}(X_i)$ of X_i based on formula (8), where $1 \leq i \leq n$.

Step 4: Let the attribute $\boldsymbol{X}$ be a decision node, and sprout the tree according to the fuzzy domain of the attribute $\boldsymbol{X}$ shown the follows:



where $X_1, X_2, \ldots, X_n$ are linguistic terms represented by fuzzy sets and the set $\{X_1, X_2, \ldots, X_n\}$ is the fuzzy domain of the attribute $\boldsymbol{X}$.

Step 5: Let $\boldsymbol{S} = \boldsymbol{S} - \boldsymbol{X}$, where $-$ is the set difference operator.

Step 6: For $i = 1$ to n do

{
Let $\boldsymbol{T} \leftarrow T_i$;
If $\boldsymbol{S} = \phi$ then
{
create a decision node for the consequent attribute $\boldsymbol{Z}$; partition the training instances $\boldsymbol{T}$ into T_1, $T_2, \ldots,$ and T_k according to the fuzzy domain $\{Z_1, Z_2, \ldots, Z_k\}$ of the attribute $\boldsymbol{Z}$;
compute the average value $\text{Avg}(Z_i)$ of Z_i, where $1 \leq i \leq \text{k}$; create a terminal node for every T_i with $\text{Avg}(Z_i) \neq 0$ and compute the certainty factor value CF_i based on formula (10) associated with the created CF node for each nonnull path in the constructed fuzzy decision tree
}
else
go to Step 3.
}

Step 7: Find the null paths in the constructed fuzzy decision tree and generate the virtual fuzzy rules with their CF values, where the CF values of the HCF nodes in the constructed fuzzy decision trees are derived by applying the regression equations of the statistics to obtain the relationship among the attributes.

In the following, we use the relation shown in Table 1 to illustrate the fuzzy decision tree construction process and the null values estimation process.

Because the attribute Salary is determined by the attributes Degree and Experience, we calculate the coefficients of determination from the attribute Degree to the attribute Salary and from the attribute Experience to the attribute Salary using formula (9), where the results are 0.5376 and 0.6204, respectively. Then, we assign the coefficient of determination from Salary to Salary to 1. Then, after normalizing the values 0.5376, 0.6204, and 1, we can get the weights of the attributes

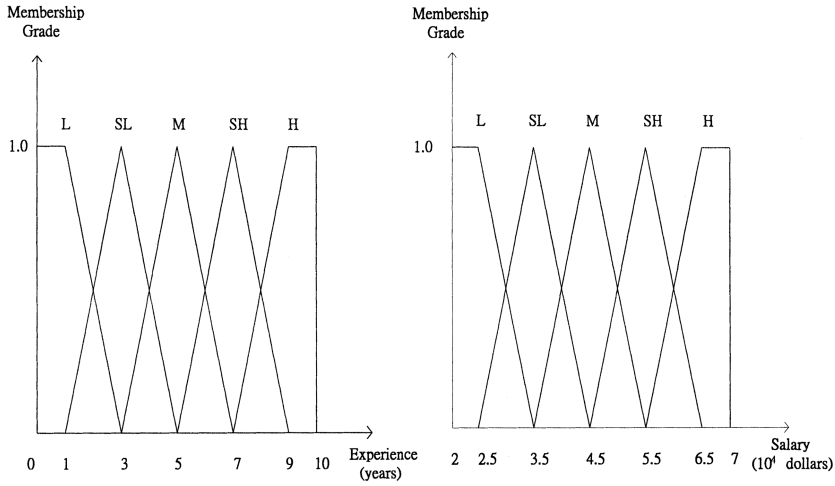


Figure 5. Membership functions of the attributes Experience and Salary (Chen and Yeh 1997).

Degree, Experience, and Salary, which are 0.25, 0.29, and 0.46, respectively.

Let S be a set of antecedent attributes and let Z be a set of consequent attributes, where $S = \{\text{Degree, Experience}\}$ and $Z = \{\text{Salary}\}$, that is, the attributes Degree and Experience determine the attribute Salary. From Table 1, we let the domains of the attribute Degree be {Ph.D. (P), Master (M), Bachelor (B)}, and let the fuzzy domain of the attributes Experience and Salary be {High (H), Somewhat-High (SH), Medium (M), Somewhat-Low (SL), Low (L)}, respectively, where the membership functions of the linguistic terms H, SH, M, SL, and L of the attributes Experience and Salary are shown in Figure 5 (Chen and Yeh 1997).

Based on Figure 5 and Wang and Mendel (1992), we can fuzzify the relation shown in Table 1. The fuzzified relation of Table 1 is shown in Table 2. For example, for every fuzzifiable value x, we can get the membership grade $\mu_{x_i}(x)$ and $\mu_{x_j}(x)$ corresponding to the linguistic terms X_i and X_j, respectively, where $\mu_{X_i}(x) \in [0, 1]$, $\mu_{X_i}(x) \in [0, 1]$, and $\mu_{X_i}(x) + \mu_{X_j}(x) = 1$. If $\mu_{X_i}(x) \geq \mu_{x_j}(x)$, then we fuzzify the value x into $\{X_i/\mu_{X_i}(x)\}$. For an unfuzzifiable value Y, where $Y \in \{\text{Bachelor, Master, Ph.D.}\}$, we let the fuzzified result of Y be $\{Y/1.0\}$.

Table 2. Fuzzified relation of Table 1 (Chen and Yeh 1997)

Emp-ID	Degree	Experience	Salary
S1	{Ph.D./1.0}	{SH/0.9}	{H/0.8}
S2	{Master/1.0}	{L/0.5}	{SL/0.8}
S3	{Bachelor/1.0}	{SH/1.0}	{SL/0.5}
S4	{Ph.D./1.0}	{L/0.9}	{M/0.8}
S5	{Master/1.0}	{SH/0.75}	{SH/0.8}
S6	{Bachelor/1.0}	{L/0.75}	{L/0.9}
S7	{Bachelor/1.0}	{SL/0.65}	{L/0.6}
S8	{Ph.D./1.0}	{L/0.5}	{M/0.5}
S9	{Ph.D./1.0}	{SL/0.6}	{SH/0.9}
S10	{Bachelor/1.0}	{SL/0.75}	{SL/1.0}
S11	{Master/1.0}	{SL/0.75}	{SL/0.5}
S12	{Master/1.0}	{SL/0.7}	{M/0.6}
S13	{Master/1.0}	{H/1.0}	{H/1.0}
S14	{Ph.D./1.0}	{M/1.0}	{SH/0.8}
S15	{Bachelor/1.0}	{M/1.0}	{SL/0.9}
S16	{Master/1.0}	{SH/0.6}	{M/0.5}
S17	{Bachelor/1.0}	{L/1.0}	{L/1.0}
S18	{Master/1.0}	{SH/0.9}	{SH/1.0}
S19	{Master/1.0}	{SH/0.75}	{SH/0.6}
S20	{Ph.D./1.0}	{SH/0.6}	{H/1.0}
S21	{Master/1.0}	{H/0.55}	{H/0.9}
S22	{Ph.D./1.0}	{H/0.75}	{H/1.0}

Then, we calculate the degree of fuzziness of each attribute in the antecedent set S using formula (6), where $S = \{\text{Degree, Experience}\}$. The calculation result is shown as follows (Chen and Yeh 1997):

$$FA(\text{Degree}) = 0,$$

$$\begin{aligned} FA(\text{Experience}) = {} & [(1-0.9)+(1-0.5)+(1-1.0)+(1-0.9) \\ & +(1-0.75)+(1-0.75)+(1-0.65)+(1-0.5) \\ & +(1-0.6)+(1-0.75)+(1-0.75)+(1-0.7) \\ & +(1-1.0)+(1-1.0)+(1-1.0)+(1-0.6) \\ & +(1-1.0)+(1-0.9)+(1-0.75)+(1-0.6) \\ & +(1-0.55)+(1-0.75)]/22 \\ = {} & 0.23. \end{aligned}$$

Because FA(Degree) has the smallest value, the attribute Degree is selected as the root node to construct the fuzzy decision tree.

The proposed EFCLS algorithm uses the weights of the attributes to calculate the CF values of the generated fuzzy rules. For example, consider the following nonnull path in the constructed fuzzy decision tree:

$$\text{Degree} \xrightarrow{\text{Ph.D.}} \text{Experience} \xrightarrow{\text{L}} \text{Salary} \xrightarrow{\text{M}} \boxed{CF}.$$

This indicates that there is the following fuzzy rule in the knowledge base:

IF Degree is Ph.D. **AND** Experience is L **THEN** Salary is M,

where we want to calculate the CF of the fuzzy rule. Let k be an attribute and let t_i be the ith tuple of a relation in a relational database system. Furthermore, let $t_i(k)$ denote the value of the attribute k of the ith tuple of a relation. From Table 2, we can see that the tuples t_1, t_4, t_8, t_9, t_{14}, t_{20}, and t_{22} whose attribute Degree have the fuzzified value "Ph.D." Thus, based on formula (8), we can get the following result:

$$\begin{aligned}\text{Avg(Degree)} = {} & (\mu_{\text{Ph.D.}}(t_1(\text{Degree})) + \mu_{\text{Ph.D.}}(t_4(\text{Degree})) \\ & + \mu_{\text{Ph.D.}}(t_8(\text{Degree})) + \mu_{\text{Ph.D.}}(t_9(\text{Degree})) \\ & + \mu_{\text{Ph.D.}}(t_{14}(\text{Degree})) + \mu_{\text{Ph.D.}}(t_{20}(\text{Degree})) \\ & + \mu_{\text{Ph.D.}}(t_{22}(\text{Degree})))/7 \\ = {} & 1.\end{aligned}$$

We also can see the tuples t_4 and t_8 whose attribute Degree has the fuzzified value Ph.D. and whose attribute Experience has the fuzzified value L. Thus, based on formula (8), we can obtain the following result:

$$\begin{aligned}\text{Avg(Experience)} &= \frac{\mu_L(t_4(\text{Experience})) + \mu_L(t_8(\text{Experience}))}{2} \\ &= \frac{0.9 + 0.5}{2} \\ &= 0.7.\end{aligned}$$

We also can see the tuples t_4 and t_8 whose attribute Degree has the fuzzified value Ph.D., whose attribute Experience has the fuzzified value

L, and whose attribute Salary has the fuzzified value M. Thus, based on formula (8), we can obtain the following result:

$$\begin{aligned}\text{Avg(Salary)} &= \frac{\mu_M(t_4(\text{Salary})) + \mu_M(t_8(\text{Salary}))}{2} \\ &= 0.65.\end{aligned}$$

Then, based on formula (10), we can calculate the CF value of the rule shown as follows:

$$\begin{aligned}CF &= 1 \times 0.25 + 0.7 \times 0.29 + 0.65 \times 0.46 \\ &\approx 0.75,\end{aligned}$$

where the weights of the attributes Degree, Experience, and Salary are 0.25, 0.29, and 0.46, respectively, as derived previously. Therefore, we can get the fuzzy rule as follows:

IF Degree is Ph.D. **AND** Experience is L
THEN Salary is M ($CF = 0.75$).

After repeatedly performing Steps 3, 4, 5, and 6 of the proposed EFCLS algorithm, we can get the constructed fuzzy decision tree as shown in Figure 6.

From Figure 6, we can obtain sixteen fuzzy rules, shown as follows:

Rule 1: **IF** Degree is Ph.D. **AND** Experience is L **THEN** Salary is M ($CF = 0.75$)

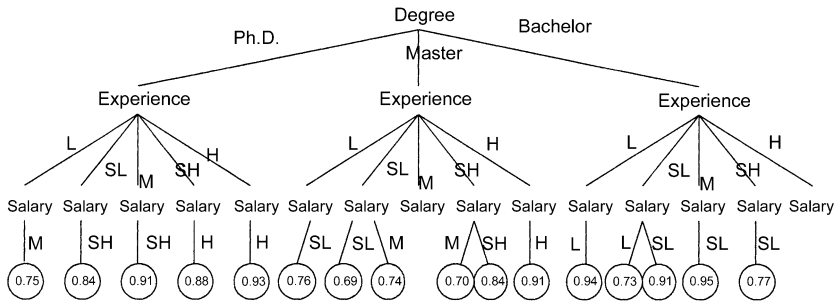


Figure 6. The constructed fuzzy decision tree.

Rule 2: **IF** Degree is Ph.D. **AND** Experience is SL **THEN** Salary is SH ($CF = 0.84$)
Rule 3: **IF** Degree is Ph.D. **AND** Experience is M **THEN** Salary is SH ($CF = 0.91$)
Rule 4: **IF** Degree is Ph.D. **AND** Experience is SH **THEN** Salary is H ($CF = 0.88$)
Rule 5: **IF** Degree is Ph.D. **AND** Experience is H **THEN** Salary is H ($CF = 0.93$)
Rule 6: **IF** Degree is Master **AND** Experience is L **THEN** Salary is SL ($CF = 0.76$)
Rule 7: **IF** Degree is Master **AND** Experience is SL **THEN** Salary is SL ($CF = 0.69$)
Rule 8: **IF** Degree is Master **AND** Experience is SL **THEN** Salary is M ($CF = 0.74$)
Rule 9: **IF** Degree is Master **AND** Experience is SH **THEN** Salary is M ($CF = 0.70$)
Rule 10: **IF** Degree is Master **AND** Experience is SH **THEN** Salary is SH ($CF = 0.84$)
Rule 11: **IF** Degree is Master **AND** Experience is H **THEN** Salary is H ($CF = 0.91$)
Rule 12: **IF** Degree is Bachelor **AND** Experience is L **THEN** Salary is L ($CF = 0.94$)
Rule 13: **IF** Degree is Bachelor **AND** Experience is SL **THEN** Salary is L ($CF = 0.73$)
Rule 14: **IF** Degree is Bachelor **AND** Experience is SL **THEN** Salary is SL ($CF = 0.91$)
Rule 15: **IF** Degree is Bachelor **AND** Experience is M **THEN** Salary is SL ($CF = 0.95$)
Rule 16: **IF** Degree is Bachelor **AND** Experience is SH **THEN** Salary is SL ($CF = 0.77$).

From Figure 6, we also can see that there are two null paths in the constructed fuzzy decision tree

$$\text{Degree} \xrightarrow{\text{Master}} \text{Experience} \xrightarrow{\text{M}} \text{Salary},$$

$$\text{Degree} \xrightarrow{\text{Bachelor}} \text{Experience} \xrightarrow{\text{H}} \text{Salary}.$$

In this situation, the above two null paths in the constructed fuzzy decision tree will generate the following two virtual fuzzy rules:

IF Degree is Master **AND** Experience is M **THEN** Salary is Z_1 $(CF = C_1)$,

IF Degree is Bachelor **AND** Experience is H
THEN Salary is Z_2 $(CF = C_2)$,

where Z_1 and Z_2 are linguistic terms, and C_1 and C_2 are CF values between zero and one.

First, based on formulas (11) and (12) and Table 1, we can derive the regression equation of the attributes Degree, Experience, and Salary as shown:

$$\text{Salary} = 10587.7 + 10147.06 \times \text{Degree} + 3074.9 \times \text{Experience} \tag{13}$$

where $\beta_0 = 10587.7$, $\beta_1 = 10147.06$, and $\beta_2 = 3074.9$ are obtained by solving formula (12) using Microsoft Excel Version 2000 on a Pentium III PC. Consider the following two fuzzy virtual rules:

Rule 17: **IF** Degree is Master **AND** Experience is M **THEN** Salary is Z_1 $(CF = c_1)$,
Rule 18: **IF** Degree is Bachelor **AND** Experience is H **THEN** Salary is Z_2 $(CF = c_2)$.

In the following, we illustrate how to get the values of Z_1, Z_2, C_1, and C_2, respectively. Let the ranking value of Bachelor be 1, the ranking value of Master be 2, and the ranking value of Ph.D. be 3. First, consider the virtual fuzzy rule: **IF** Degree is Master **AND** Experience is M **THEN** Salary is Z_1 $(CF = c_1)$. Because the value of Master is 2, and because from Figure 4 we can see that the linguistic term M has the maximum membership value when Experience is equal to 5.0 years, based on formula (13), we let the value of Degree be equal to 2 and let the value of Experience be equal to 5.0 to calculate the value of Z_1 as shown:

$$\begin{aligned} Z_1 &= 10158 + 10147.06 \times 2 + 3074.9 \times 5.0 \\ &\approx 45826. \end{aligned}$$

After fuzzifying the value of 45826, we can see that it belongs to the linguistic term Medium (M) with the membership value 0.91. Then, based on formula (10), we can see that the value of c_1 can be calculated as follows:

$$c_1 = 1 \times 0.25 + 1 \times 0.29 + 0.91 \times 0.46$$
$$\approx 0.95.$$

Thus, we can get the following virtual fuzzy rule:

Rule 17: IF Degree is Master **AND** Experience is M **THEN** Salary is M ($CF = 0.95$).

Then, consider the following virtual fuzzy rule:

IF Degree is Bachelor **AND** Experience is H
THEN Salary is Z_2 $(CF = C_2)$.

Because the ranking value of Bachelor is 1, and considering Figure 5, we can see that the linguistic term H has the maximum membership value when Experience is equal to 9.0 years. Based on formula (13), we let the value of Degree be equal to 1 and let the value of Experience be equal to 9.0 to calculate the value of Z_2 as shown:

$$Z_2 = 10158.7 + 10147.06 \times 1 + 3074.9 \times 9.0$$
$$\approx 47979.$$

After fuzzifying the value of 47979, we can see that it belongs to the linguistic term Medium (M) with the membership value 0.70. Then, based on formula (10), we can see that the value of c_2 can be calculated as follows:

$$c_2 = 1 \times 0.25 + 1 \times 0.29 + 0.70 \times 0.46$$
$$\approx 0.86.$$

Thus, we can get the following virtual fuzzy rule:

Rule 18: IF Degree is Bachelor **AND** Experience is H **THEN** Salary is M ($CF = 0.86$).

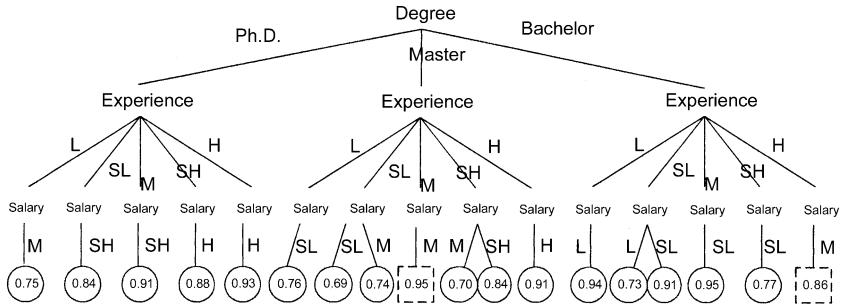


Figure 7. A complete fuzzy decision tree.

Thus, a complete fuzzy decision tree has been constructed as shown in Figure 7.

After constructing a complete fuzzy decision tree and generating fuzzy rules from the constructed fuzzy decision tree, we can apply the generated fuzzy rules to estimate null values in relational database system based on previous work (Chen and Yeh 1997). Consider the following fuzzy rules:

IF X is X_a **AND** Y is Y_b **THEN** Z is Z_{M_1} $(CF = C_1)$

IF X is X_a **AND** Y is Y_b **THEN** Z is Z_{M_2} $(CF = C_2)$

IF X is X_c **AND** Y is Y_d **THEN** Z is Z_{N_1} $(CF = D_1)$

IF X is X_c **AND** Y is Y_d **THEN** Z is Z_{N_2} $(CF = D_2)$,

where X and Y are antecedent attributes of the fuzzy rules; Z is the consequent attribute of the fuzzy rules; X_a, X_c, Y_b, Y_d, Z_{M1}, Z_{M2}, Z_{N1}, and Z_{N2} are linguistic terms represented by fuzzy sets; C_1, C_2, D_1, and D_2 are real values between zero and one. Assume that x and y are crisp domain values of the attributes X and Y in some tuples of a relational database, respectively, and assume that z is a null value of attribute Z. Let the fuzzified value of x be $\{X_a/\mu_{X_a}(x), X_b/\mu_{X_b}(x)\}$, and let the fuzzified value of y be $\{Y_a/\mu_{y_a}(y), Y_b/\mu_{y_b}(y)\}$. Then, the null value z can be calculated as follows:

$$z = \frac{\mu_{X_a}(x) \times \mu_{Y_c}(y) \times \frac{\sum_{i=1}^{2} C_i \times \mathrm{DEF}(Z_{Mi})}{\sum_{i=1}^{2} C_i} + \mu_{X_b}(x) \times \mu_{Y_d}(y) \times \frac{\sum_{i=1}^{2} D_i \times \mathrm{DEF}(Z_{Ni})}{\sum_{i=1}^{2} D_i}}{\mu_{X_a}(x) \times \mu_{Y_c}(y) + \mu_{X_b}(x) \times \mu_{Y_d}(y)}, \tag{14}$$

where $\mathrm{DEF}(Z_{M_i})$ and $\mathrm{DEF}(Z_{N_i})$ are the defuzzified values of the fuzzy sets Z_{M_i} and Z_{N_i}, respectively.

For example, consider the tuple whose Emp-ID = S7 shown in Table 1, where the value of the attribute Degree of the tuple is Bachelor and the value of the attribute Experience of the tuple is 2.3. Then, based on Figure 5, we can see that the degree of membership that 2.3 belongs to for SL and L are 0.65 and 0.35, respectively, that is, $\mu_{\mathrm{SL}}(2.3) = 0.65$ and $\mu_L(2.3) = 0.35$. Therefore, the fuzzified value of the attribute Degree of the tuple whose Emp-ID = S7 is {Bachelor/1.0, Bachelor/1.0}, and the fuzzified value of the attribute Experience of the tuple whose Emp-ID = S7 is {SL/0.65, L/0.35}. Based on formula (14) and the generated fuzzy rules 12, 13, and 14, the value of the attribute Salary of the tuple whose Emp-ID = S7 can be estimated, where Rules 12, 13, and 14 are shown as follows:

Rule 12: IF Degree is Bachelor **AND** Experience is L **THEN** Salary is L ($CF = 0.94$),

Rule 13: IF Degree is Bachelor **AND** Experience is SL **THEN** Salary is L ($CF = 0.73$),

Rule 14: IF Degree is Bachelor **AND** Experience is SL **THEN** Salary is SL ($CF = 0.91$).

Based on formula (14), the null value z of the attribute Salary of the tuple whose Emp-ID = S7 can be estimated as follows:

$$\frac{1 \times 0.65 \times \frac{0.73 \times 25000 + 0.91 \times 35000}{0.73 + 0.91} + 1 \times 0.35 \times \frac{0.94 \times 25000}{0.94}}{1 \times 0.65 + 1 \times 0.35} \approx 28606,$$

where DEF(L) = 25000 and DEF(SL) = 35000 are calculated based on Figure 5, formula (4), and formula (5).

Table 3. A comparison of the estimated error rates of the proposed method with the existing methods

				Chen and Yeh's method (1997)		Chen and Chen's method (2000)		Proposed method	
Emp-ID	Degree	Experience	Salary	Salary (estimated)	Estimated error rate	Salary (estimated)	Estimated error rate	Salary (estimated)	Estimated error rate
S1	Ph.D.	7.2	63000	65000	+3.17%	63000	+0.00%	65000	3.17%
S2	Master	2	37000	30704	−17.02%	33711	−8.89%	37587	1.59%
S3	Bachelor	7	40000	35000	−12.50%	46648	+16.62%	35000	−12.50%
S4	Ph.D.	1.2	47000	46000	−2.13%	36216	−22.94%	46000	−2.13%
S5	Master	7.5	53000	54500	+2.83%	56200	+6.04%	54090	2.06%
S6	Bachelor	1.5	26000	26346	+1.33%	27179	+4.53%	26387	1.49%
S7	Bachelor	2.3	29000	28500	−1.72%	29195	+0.67%	28606	−1.36%
S8	Ph.D.	2	50000	50000	+0.00%	39861	−20.28%	50000	0.00%
S9	Ph.D.	3.8	54000	55000	+1.85%	48061	−11.00%	55000	1.85%
S10	Bachelor	3.5	35000	31538	−9.89%	32219	−7.95%	31661	−9.54%
S11	Master	3.5	40000	41590	+3.98%	40544	+1.36%	41381	3.45%
S12	Master	3.6	41000	45159	+10.14%	41000	+0.00%	41622	1.52%
S13	Master	10	68000	65000	−4.41%	64533	−5.10%	65000	−4.41%
S14	Ph.D.	5	57000	55000	−3.51%	55666	−2.34%	55000	−3.51%
S15	Bachelor	5	36000	35000	−2.78%	35999	+0.00%	35000	−2.78%
S16	Master	6.2	50000	48600	−2.80%	51866	+3.73%	48272	−3.46%
S17	Bachelor	0.5	23000	25000	+8.70%	24659	+7.21%	25000	8.70%
S18	Master	7.2	55000	52400	−4.73%	55200	+0.36%	51909	−5.62%
S19	Master	6.5	51000	49500	−2.94%	52866	+3.66%	49090	−3.75%
S20	Ph.D.	7.8	65000	65000	+0.00%	65000	+0.00%	65000	0.00%
S21	Master	8.1	64000	58700	−8.28%	58200	−9.06%	58454	−8.67%
S22	Ph.D.	8.5	70000	65000	−7.14%	67333	−3.81%	65000	−7.14%
Average Estimated Error Rate (%)					5.08%		6.14%		4.03%

In the same way, we can estimate the value of the attribute Salary of each tuple shown in Table 1. The estimated results are shown in Table 3. Table 3 also shows the estimated results of the methods we presented previously (Chen and Yeh 1997; Chen and Chen 2000). A comparison of the estimated error rates of the methods we presented in previous work and the proposed method is also shown in Table 3, where the estimated error rate is calculated as follows:

$$\text{Estimated Error Rate} = \frac{\text{Estimated Value} - \text{Original Value}}{\text{Original Value}} \times 100\%. \tag{15}$$

From Table 3, we can see that the average estimated error rate of the proposed method is less than the ones we presented previously (Chen and Yeh 1997; Chen and Chen 2000). That is, the estimated accuracy rate of the proposed method is better than the ones we presented previously.

CONCLUSIONS

In this article, we presented a new method to construct fuzzy decision trees and to generate fuzzy rules from relational database systems for estimating null values. We use the concept of coefficient of determination of the statistic to calculate the weights of the attributes in a relational database system and use the normalized weights to derive the CF values of the generated fuzzy rules. Furthermore, we also use the regression equations of the statistics to construct a complete fuzzy decision tree for generating better fuzzy rules for estimating null values in relational database systems. The proposed method can obtain a higher average estimated accuracy rate than the ones we presented previously (Chen and Yeh 1997; Chen and Chen 2000) to estimate null values in relational database systems.

REFERENCES

Berenson, M. L., D. M. Levine, and M. Goldstein. 1983. *Intermediate statistical methods and applications*. Englewood Cliffs, NJ: Prentice-Hall.

Chang, C. H., and S. M. Chen. 2000. A new method to generate fuzzy rules from numerical data based on the exclusion of attribute terms. Proceedings of the

2000 International Computer Symposium: Workshop on the Artificial Intelligence, pp. 57–64, Chiayi, Taiwan, Republic of China.

Chen, S. M. 1994. Using fuzzy reasoning techniques for fault diagnosis of the J-85 jet engines. Proceedings of the Third National Conference on Science and Technology of National Defense, Vol. 1, pp. 29–34, Taoyuan, Taiwan, Republic of China.

Chen, S. M., and H. H. Chen. 2000. Estimating null values in the distributed relational databases environment. *Cybernetics and Systems: An International Journal*, 31(8):851–871.

Chen, S. M., and S. Y. Lin. 2000. A new method for constructing fuzzy decision trees and generate fuzzy classification rules from training examples. *Cybernetics and Systems: An International Journal*, 31(7):763–785.

Chen, S. M., and M. S. Yeh. 1997. Generating fuzzy rules from relational database systems for estimating null values. *Cybernetics and Systems: An International Journal*, 28(2):695–723.

Chen, S. M., S. H. Lee, and C. H. Lee. 2001. A new method for generating fuzzy rules from numerical data for handling classification problems. *Applied Artificial Intelligence: An International Journal*, 15(7):645–664.

Chen, Y. J., and S. M. Chen. 2000. A new method to generate fuzzy rules for fuzzy classification systems. Proceedings of the 2000 Eighth National Conference on Fuzzy Theory and Its Applications, 1–2 December 2003, Taipei, Taiwan, Republic of China.

Hunt, E. B., J. Marin, and P. J. Stone. 1966. *Experience in induction*. New York: Academic Press.

Jeng, B., and T. P. Liang. 1993. Fuzzy indexing and retrieval in case-based systems. Proceedings of the 1993 Pan Pacific Conference on Information Systems, pp. 258–266, Taiwan, Republic of China.

Kandel, A. 1986. *Fuzzy mathematical techniques with applications*. Reading, MA: Addision–Wesley.

Lee, S. W., and S. M. Chen. 2001. A new method to generate fuzzy rules from relational database systems. Proceedings of the 2001 Ninth National Conference on Fuzzy Theory and Applications, 23–24 November 2001, Chungli, Taoyuan, Taiwan, Republic of China.

Lin, H. L., and S. M. Chen. 2000. Generating weighted fuzzy rules from training data for handling fuzzy classification problems. Proceedings of the 2000 International Computer Symposium: Workshop on Artificial Intelligence, pp. 11–18, Chiayi, Taiwan, Republic of China.

Mendenhall, W., and R. J. Beaver. 1994. *Introduction to probability and statistics*. Belmont, CA: Wadsworth.

Neter, J., H. Micael, C. J. Nachtsheim, and W. Wasserman. 1999. *Applied linear statistical models*. New York: McGraw–Hill.

Quinlan, J. R. 1979. Discovering rules by induction from large collection of examples. In *Expert systems in the micro electronic age*, edited by D. Michie. Edinburgh: Edinburgh University Press.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

Sudkamp, T., and R. J. Hammell, II. 1994. Interpolation, completion, and learning fuzzy rules. *IEEE Transactions on Systems, Man, and Cybernatics*, 24(2):332–342.

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

Wu, T. P., and S. M. Chen. 1999. A new method for constructing membership functions and fuzzy rules from training examples. *IEEE Transactions on Systems, Man, and Cybernatics-Part B: Cybernatics*, 29(1):25–40.

Yasdi, R. 1991. Learning classification rules from database in the context of knowledge acquisition and representation. *IEEE Transactions on Knowledge and Data Engineering*, 3(3):293–306.

Yeung, D. S., and E. C. C. Tsang. 1995. A weighted fuzzy production rule evaluation method. Proceedings of 1995 Fourth IEEE International Conference on Fuzzy Systems, Vol. 2, pp. 461–468, Yokohama, Japan.

Yeung, D. S., and E. C. C. Tsang. 1997. Weighted fuzzy production rules. *Fuzzy Sets and Systems*, 88(3):299–313.

Zadeh, L. A. 1965. Fuzzy sets. *Information and Control*, 8:338–353.

Zadeh, L. A. 1975. The concept of a linguistic variable and its application to approximate reasoning (I). *Informational Sciences*, 8:199–249.