International Journal of Applied Science and Engineering
2005. 3, 1: 37- 49

A New Approach for Handling the Iris Data Classification Problem

Shyi-Ming Chen[a*] and Yao-De Fang[b]

[a] *Department of Computer Science and Information Engineering,*
National Taiwan University of Science and Technology,
Taipei 106, Taiwan, R.O.C.
[b] *Department of Electronic Engineering,*
National Taiwan University of Science and Technology,
Taipei 106, Taiwan, R.O.C.

Abstract: In this paper, we present a new method to deal with the Iris data classification problem based on the distribution of training instances. First, we find two useful attributes of the Iris data from the training instances that are more suitable to deal with the classification problem. It means that the distribution of the values of these two useful attributes of the three species (i.e., Setosa, Versicolor and Virginica) has less overlapping. Then, we calculate the average attribute values and the standard deviations of these two useful attributes. We also calculate the overlapping areas formed by the values of these two useful attributes between species of the training instances, the average attribute values, and the standard deviations of the values of these two useful attributes of each species. Then, we calculate the difference between the values of these two useful attributes of a testing instance to be classified and the values of these two useful attributes of each species of the training instances. We choose the species that has the smallest difference between the values of these two useful attributes of the testing instance and the values of these two useful attributes of each species of the training instances as the classification result of the testing instance. The proposed method gets a higher average classification accuracy rate than the existing methods.

Keywords: Iris data; maximum attribute value; minimum attribute value; standard deviation; average classification accuracy rate.

1. Introduction

It is obvious that how to deal with classification problems is a very important research topic of fuzzy classification systems [2, 3, 4, 7, 8, 13, 14, 15, 16, 20, 25, 26]. In [2], Castro et al. presented a method to learn maximal structure rules in fuzzy logic to deal with the Iris data [11] classification problem. In [8], Chen et al. presented a method to generate fuzzy rules from training instances based on genetic algorithms to deal with the Iris data classification problem. In [13], Hong et al. presented a method to generate fuzzy rules and membership functions from training examples to deal with the Iris data classification problem. In [15], Hong et al. presented a method to generate fuzzy rules by finding the relative attributes from training data to deal with the Iris data classification problem. In

[*] Corresponding author: e-mail: smchen@et.ntust.edu.tw

Accepted for Publication: Feb. 15, 2005

[16], Hong et al. presented a method to generate fuzzy rules by processing individual fuzzy attributes to deal with the Iris data classification problem. In [17], Hong et al. discussed the effect of merging order on performance of fuzzy rules induction for handling the Iris data classification problem. In [20], Lin et al. presented a method to generate weighted fuzzy rules from the training data based on genetic algorithms to deal with the Iris data classification problem. In [25], Wu et al. presented a method to construct membership functions and generate fuzzy rules from training instances to deal with the Iris data classification problem. In [26], Wang et al. presented a method to generate modular fuzzy rules to deal with the Iris data classification problem.

In this paper, we present a new method to deal with the Iris data classification problem based on the distribution of training instances. First, we find two useful attributes of the Iris data from the training instances that are more suitable to deal with the classification problem. It means that the distribution of the values of these two useful attributes of the three species (i.e., Setosa, Versicolor and Virginica) has less overlapping. Then, we calculate the average attribute values and the standard deviations of these two useful attributes. We also calculate the overlapping areas formed by the values of these two useful attributes between species of the training instances, the average attribute values, and the standard deviations of the values of these two useful attributes of each species. Then, we calculate the difference between the values of these two useful attributes of a testing instance to be classified and the values of these two useful attributes of each species of the training instances. We choose the species that has the smallest difference between the values of these two useful attributes of the testing instance and the values of these two useful attributes of each species of the training instances as the classification result of the testing instance. The proposed method gets a higher average classification accuracy rate than the existing methods.

The rest of this paper is organized as follows. In Section 2, we present a new method for handling the Iris data classification problem based on the distribution of training instances. In Section 3, we use an example to illustrate the proposed method. In Section 4, we compare the average classification accuracy rate of the proposed method with that of the existing methods. The conclusions are discussed in Section 5.

2. A new method for handling the Iris data classification problem based on the distribution of training instances

In this section, we present a new method to deal with the Iris data [11] classification problem based on the distribution of training instances. The Iris data has 150 instances as shown in Table 1, where there are four attributes, i.e., Sepal Length (SL), Sepal Width (SW), Petal Length (PL) and Petal Width (PW), and there are three species, i.e., Setosa, Versicolor, and Virginica, where each species has 50 training instances. In recent years, many methods have been proposed to deal with the Iris data classification problem [2, 3, 4, 7, 8, 13, 14, 15, 16, 20, 25, 26]. In the following, we present a new method to deal with the Iris data classification problem based on the distribution of training instances. Let us consider the Iris data shown in Table 1. We randomly choose n instances as the training instances and let the rest of 150 – n instances be the testing instances. The proposed algorithm is now presented as follows:

Step 1: Find the maximum attribute value and the minimum attribute value of each attribute of the n training instances.

Step 2: Based on the maximum attribute value and the minimum attribute value of any two attributes of each species of the training instances to form the boundaries of a plane. Then, calculate the overlapping area of the three planes formed by the values of any two attributes with respect to the three species of the training instances. For example, assume that a training instance belongs to the species "Setosa" and assume that the pair of the maximum attribute value and the minimum attribute value of the attributes PL and PW are ($PL_{max(Setosa)}$, $PW_{max(Setosa)}$) and ($PL_{min(Setosa)}$, $PW_{min(Setosa)}$), respectively, then a plane is formed as shown in Figure 1, and the area formed by the pairs ($PL_{max(Setosa)}$, $PW_{max(Setosa)}$) and ($PL_{min(Setosa)}$, $PW_{min(Setosa)}$) is equal to ($PL_{max(Setosa)} - PL_{min(Setosa)}$) × ($PW_{max(Setosa)} - PW_{min(Setosa)}$). The overlapping area of the three planes formed by the values of any two attributes with respect to the three species of the training instances can be calculated described as follows. Assume that A, B and C are three species of the Iris data, and assume that the overlapping area of the species A and the species B of the training instances is denoted by the area of oblique lines as shown in Figure 2. Assume that the maximum attribute values of the attribute PL and the attribute PW of the species A of the training instances are $PL_{max(A)}$ and $PW_{max(A)}$, respectively; assume that the maximum attribute values of the attribute PL and the attribute PW of the species B of the training instances are $PL_{max(B)}$ and $PW_{max(B)}$, respectively; assume that the minimum attribute values of the attribute PL and the attribute PW of the species A of the training instances are $PL_{min(A)}$ and $PW_{min(A)}$, respectively; assume that the minimum attribute values of the attribute PL and the attribute PW of the species B of the training instances are $PL_{min(B)}$ and $PW_{min(B)}$, respectively. Then, the overlapping area formed by the values of attributes PL and PW of the species A and species B is equal to $(\min(PL_{max(A)}, PL_{max(B)}) - \max(PL_{min(A)}, PL_{min(B)})) \times (\min(PW_{max(A)}, PW_{max(B)}) - \max(PW_{min(A)}, PW_{min(B)}))$. Because there are three species in the training instances, we can see that there are three planes formed by the values of the attributes PL and PW as shown in Figure 3.

Step 3: Calculate the overlapping area of the values of each pair of attributes of the training instances belonging to different species. If the attributes X and Y have the smallest total overlapping area, then these two attributes X and Y are useful attributes to be used for dealing with the Iris data classification problem, and the other attributes are useless attributes and are discarded. For example, assume that the overlapping areas formed by the values of the attribute PL and the attribute PW of the three species A, B and C of the training instances are as shown in Figure 3, then the method for calculating the overlapping area formed by the values of the attributes PL and PW of the three species A, B and C is as follows [10]. If the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species A and the species B is zero or the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species B and the species C is zero or the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species C and the species A is zero, then the total overlapping area = "the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species A and the species B" + "the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species B and the species C" + "the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species C and the species A". If the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species A and the species B is not zero and the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species B and the species C is not zero and the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species C and the species A is not zero, then the total overlapping area = "the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species A and the species B" + "the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species B and the species C" + "the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species C and the species A" – 2 × "the overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species A, B and C".

Step 4: Let the average attribute values of the attributes X and Y of the training instances belonging to the species Setosa be $\overline{x_{\text{Setosa}}}$ and $\overline{y_{\text{Setosa}}}$, respectively; let the average attribute values of the attributes X and Y of the training instances belonging to the species Versicolor be $\overline{x_{\text{Versicolor}}}$ and $\overline{y_{\text{Versicolor}}}$, respectively; let the average attribute value of the attributes X and Y of the training instances belonging to the species Virginica be $\overline{x_{\text{Virginica}}}$ and $\overline{y_{\text{Virginica}}}$, respectively, where

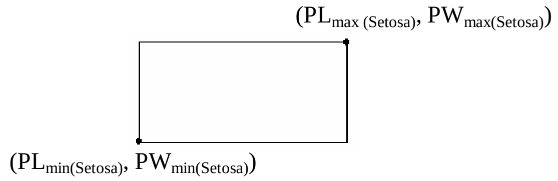
($PL_{\text{max (Setosa)}}$, $PW_{\text{max(Setosa)}}$)
($PL_{\text{min(Setosa)}}$, $PW_{\text{min(Setosa)}}$)

Figure 1. The area formed by the pairs (PL_{max}, PW_{max}) and (PL_{min}, PW_{min})

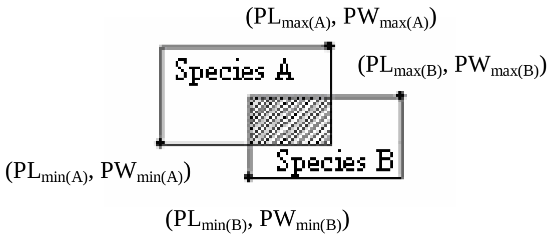
($PL_{\text{max(A)}}$, $PW_{\text{max(A)}}$)
Species A
($PL_{\text{max(B)}}$, $PW_{\text{max(B)}}$)
($PL_{\text{min(A)}}$, $PW_{\text{min(A)}}$)
Species B
($PL_{\text{min(B)}}$, $PW_{\text{min(B)}}$)

Figure 2. The overlapping area formed by the values of the attributes PL and PW of the training instances belonging to the species A and B

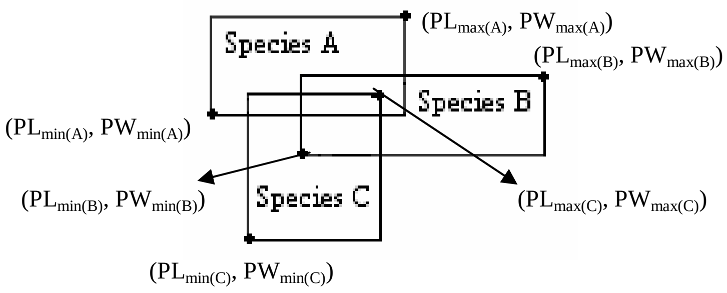
Species A
($PL_{\text{max(A)}}$, $PW_{\text{max(A)}}$)
($PL_{\text{max(B)}}$, $PW_{\text{max(B)}}$)
Species B
($PL_{\text{min(A)}}$, $PW_{\text{min(A)}}$)
($PL_{\text{min(B)}}$, $PW_{\text{min(B)}}$)
Species C
($PL_{\text{max(C)}}$, $PW_{\text{max(C)}}$)
($PL_{\text{min(C)}}$, $PW_{\text{min(C)}}$)

Figure 3. The overlapping areas formed by the values of the attributes PL and PW belonging to the training instances of the three species A, B and C, respectively

Table 1. Iris data [11]

Setosa				Versicolor				Virginica			
SL	SW	PL	PW	SL	SW	PL	PW	SL	SW	PL	PW
5.1	3.5	1.4	0.2	7	3.2	4.7	1.4	6.3	3.3	6	2.5
4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4	1.3	6.3	2.9	5.6	1.8
5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5	3.4	1.5	0.2	4.9	2.1	3.3	1	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5	2	3.5	1	6.5	3.2	5.1	2
4.8	3.4	1.6	0.2	5.9	3	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3	1.4	0.1	6	2.2	4	1	6.8	3	5.5	2.1
4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5	2
5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3	4.5	1.5	6.5	3	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6	2.2	5	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4	1.3	5.6	2.8	4.9	2
4.6	3.6	1	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5	3	1.6	0.2	6.6	3	4.4	1.4	7.2	3.2	6	1.8
5	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3	5	1.7	6.1	3	4.9	1.8
5.2	3.4	1.4	0.2	6	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1	7.2	3	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1	7.9	3.8	6.4	2
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3	4.5	1.5	6.1	2.6	5.6	1.4
5	3.2	1.2	0.2	6	3.4	4.5	1.6	7.7	3	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

$$\overline{x_{\text{Setosa}}} = \frac{1}{k}\sum_{i=1}^{k} x_{\text{Setosa}(i)}\ , \tag{1}$$

$x_{\text{Setosa}(i)}$ denotes the attribute value of the attribute X of the *i*th training instance belonging to the species Setosa, and k is the number of training instances belonging to the species Setosa;

$$\overline{x_{\text{Versicolor}}} = \frac{1}{s}\sum_{i=1}^{s} x_{\text{Versicolor}(i)}\ , \tag{2}$$

$x_{\text{Versicolor}(i)}$ denotes the attribute value of the attribute X of the *i*th training instance belonging to the species Versicolor, and s is the number of training instances belonging to the species Versicolor;

$$\overline{x_{\text{Virginica}}} = \frac{1}{t}\sum_{i=1}^{t} x_{\text{Virginica}(i)}\ , \tag{3}$$

$x_{\text{Virginica}(i)}$ denotes the attribute value of the attribute X of the *i*th training instance belonging to the species Virginica, and t is the number of training instances belonging to the species Virginica;

$$\overline{y_{\text{Setosa}}} = \frac{1}{k}\sum_{i=1}^{k} y_{\text{Setosa}(i)}\ , \tag{4}$$

$y_{\text{Setosa}(i)}$ denotes the attribute value of the attribute Y of the *i*th training instance belonging to the species Setosa, and k is the number of training instances belonging to the species Setosa;

$$\overline{y_{\text{Versicolor}}} = \frac{1}{s}\sum_{i=1}^{s} y_{\text{Versicolor}(i)}\ , \tag{5}$$

$y_{\text{Versicolor}(i)}$ denotes the attribute value of the attribute Y of the *i*th training instance belonging to the species Versicolor, and s is the number of training instances belonging to the species Versicolor;

$$\overline{y_{\text{Virginica}}} = \frac{1}{t}\sum_{i=1}^{t} y_{\text{Virginica}(i)}\ , \tag{6}$$

$y_{\text{Virginica}(i)}$ denotes the attribute value of the attribute Y of the *i*th training instance belonging to the species Virginica, and t is the number of training instances belonging to the species Virginica.

Step 5: Let the standard deviations of the values of the attributes X and Y of the training instances belonging to the species Setosa be $SD_{X(\text{Setosa})}$ and $SD_{Y(\text{Setosa})}$, respectively; let the standard deviations of the attributes X and Y of the training instances belonging to the species Versicolor be $SD_{X(\text{Versicolor})}$ and $SD_{Y(\text{Versicolor})}$, respectively; let the standard deviations of the attributes X and Y of the training instances belonging to the species Virginica be $SD_{X(\text{Virginica})}$ and $SD_{Y(\text{Virginica})}$, respectively, where

$$SD_{X(\text{Setosa})} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_{\text{Setosa}(i)} - \overline{x_{\text{Setosa}}})^{2}}\ , \tag{7}$$

k is the number of training instances belonging to the species Setosa, $x_{\text{Setosa}(i)}$ denotes the attribute value of the attribute X of the *i*th training instance belonging to the species Setosa, and $\overline{x_{\text{Setosa}}}$ denotes the average attribute value of the attribute X of the training instances belonging to the species Setosa;

$$SD_{X(\text{Versicolor})} = \sqrt{\frac{1}{s}\sum_{i=1}^{s}(x_{\text{Versicolor}(i)} - \overline{x_{\text{Versicolor}}})^{2}}\ , \tag{8}$$

where *s* is the number of training instances belonging to the species Versicolor, $x_{\text{Versicolor}(i)}$ denotes the attribute value of the attribute X of the *i*th training instance belonging to the species Versicolor, and $\overline{x_{\text{Versicolor}}}$ denotes the average attribute value of the attribute X of the training instances belonging to the species Versicolor;

$$SD_{X(\text{Virginica})} = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(x_{\text{Virginica}(i)} - \overline{x_{\text{Virginica}}})^{2}}\ , \tag{9}$$

where *t* is the number of training instances belonging to the species Virginica, $x_{\text{Virginica}(i)}$ denotes the attribute value of the attribute X of the *i*th training instance belonging to the species Virginica, and $\overline{x_{\text{Virginica}}}$ denotes the average attribute value of the attribute X of the training instances belonging to the species Virginica;

$$SD_{Y(\text{Setosa})} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_{\text{Setosa}(i)} - \overline{y_{\text{Setosa}}})^{2}}\ , \tag{10}$$

where k is the number of training instances belonging to the species Setosa, $y_{\text{Setosa}(i)}$ denotes the attribute value of the attribute Y of the ith training instance belonging to the species Setosa, and $\overline{y_{\text{Setosa}}}$ denotes the average attribute value of the attribute Y of the training instances belonging to the species Setosa;

$$\text{SD}_{\text{Y(Versicolor)}} = \sqrt{\frac{1}{s}\sum_{i=1}^{s}(y_{\text{Versicolor}(i)} - \overline{y_{\text{Versicolor}}})^2}\,, \quad (11)$$

where s is the number of training instances belonging to the species Versicolor, $y_{\text{Versicolor}(i)}$ denotes the attribute value of the attribute Y of the ith training instance belonging to the species Versicolor, and $\overline{y_{\text{Versicolor}}}$ denotes the average attribute value of the attribute Y of the training instances belonging to the species Versicolor;

$$\text{SD}_{\text{Y(Virginica)}} = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(y_{\text{Virginica}(i)} - \overline{y_{\text{Virginica}}})^2}\,, \quad (12)$$

where t is the number of training instances belonging to the species Virginica, $y_{\text{Virginica}(i)}$ denotes the attribute value of the attribute Y of the ith training instance belonging to the species Virginica and $\overline{y_{\text{Virginica}}}$ denotes the average attribute value of the attribute Y of the training instances belonging to the species Virginica.

Step 6: Calculate the area formed by the values of the attributes X and Y of the training instances belonging to the species Setosa, Versicolor, and Virginica, respectively, described as follows. Let the maximum attribute value and the minimum attribute value of the attribute X of the training instances belonging to the species Setosa be $X_{\text{max(Setosa)}}$ and $X_{\text{min(Setosa)}}$, respectively; let the maximum attribute value and the minimum attribute value of the attribute Y of the training instances belonging to the species Setosa be $Y_{\text{max(Setosa)}}$ and $Y_{\text{min(Setosa)}}$, respectively; let the maximum attribute value and the minimum attribute value of the attribute X of the training instances belonging to the species Versicolor be $X_{\text{max(Versicolor)}}$ and $X_{\text{min(Versicolor)}}$, respectively; let the maximum attribute value and the minimum attribute value of the attribute Y of the training instances belonging to the species Versicolor be $Y_{\text{max(Versicolor)}}$ and $Y_{\text{min(Versicolor)}}$, respectively; let the maximum attribute value and the minimum attribute value of the attribute X of the training instances belonging to the species Virginica be $X_{\text{max(Virginica)}}$ and $X_{\text{min(Virginica)}}$, respectively; let the maximum attribute value and the minimum attribute value of the attribute Y of the training instances belonging to the species Virginica be $Y_{\text{max(Virginica)}}$ and $Y_{\text{min(Virginica)}}$, respectively. Then, the area formed by the values of the attributes X and Y of the training instances belonging to the species Setosa, Versicolor, and Virginica are $\text{Area}_{\text{Setosa}}$, $\text{Area}_{\text{Versicolor}}$, and $\text{Area}_{\text{Virginica}}$, respectively, where

$$\text{Area}_{\text{Setosa}} = (X_{\text{max(Setosa)}} - X_{\text{min(Setosa)}}) \times (Y_{\text{max(Setosa)}} - Y_{\text{min(Setosa)}}), \quad (13)$$

$$\text{Area}_{\text{Versicolor}} = (X_{\text{max(Versicolor)}} - X_{\text{min(Versicolor)}}) \times (Y_{\text{max(Versicolor)}} - Y_{\text{min(Versicolor)}}), \quad (14)$$

$$\text{Area}_{\text{Virginica}} = (X_{\text{max(Virginica)}} - X_{\text{min(Virginica)}}) \times (Y_{\text{max(Virginica)}} - Y_{\text{min(Virginica)}}). \quad (15)$$

From the above steps of the proposed algorithm, we can obtain the needed information from the training instances. Based on this information, we can deal with the classification of testing instances. In the following, we describe how to classify a testing instance. First, we can find two useful attributes X and Y through Step 1 to Step 3 of the proposed algorithm. Assume that the values of these two useful attributes X and Y of the testing instance are x and y, respectively. Based on Step 4 to Step 6 of the proposed algorithm, we can obtain the values of $\overline{x_{\text{Setosa}}}$, $\overline{y_{\text{Setosa}}}$, $\overline{x_{\text{Versicolor}}}$, $\overline{y_{\text{Versicolor}}}$, $\overline{x_{\text{Virginica}}}$, $\overline{y_{\text{Virginica}}}$, $\text{SD}_{\text{X(Setosa)}}$, $\text{SD}_{\text{Y(Setosa)}}$, $\text{SD}_{\text{X(Versicolor)}}$, $\text{SD}_{\text{Y(Versicolor)}}$, $\text{SD}_{\text{X(Virginica)}}$, $\text{SD}_{\text{Y(Virginica)}}$, $\text{Area}_{\text{Setosa}}$, $\text{Area}_{\text{Versi-}}$

color and $Area_{Virginica}$, respectively. After we add the testing instance into the training instances, we can see that the areas formed by the values of the useful attributes X and Y of the training instances belonging to the species Setosa, Versicolor and Virginica are $NArea_{Setosa}$, $NArea_{Versicolor}$ and $Narea_{Virginica}$, respectively, where

$$NArea_{Setosa} = (\max(X_{\max(Setosa)}, x) - \min(X_{\min(Setosa)}, x)) \times (\max(Y_{\max(Setosa)}, y) - \min(Y_{\min(Setosa)}, y)), \quad (16)$$

$$NArea_{Versicolor} = (\max(X_{\max(Versicolor)}, x) - \min(X_{\min(Versicolor)}, x)) \times (\max(Y_{\max(Versicolor)}, y) - \min(Y_{\min(Versicolor)}, y)), \quad (17)$$

$$NArea_{Virginica} = (\max(X_{\max(Virginica)}, x) - \min(X_{\min(Virginica)}, x)) \times (\max(Y_{\max(Virginica)}, y) - \min(Y_{\min(Virginica)}, y)). \quad (18)$$

Then, based on these numerical data, we can see that the difference between the testing instance and the training instances belonging to the species Setosa, Versicolor and Virginica are $Diff_{Setosa}$, $Diff_{Versicolor}$, and $Diff_{Virginica}$, respectively, where

$$Diff_{Setosa} = (NArea_{Setosa} - Area_{Setosa}) + |x - \overline{x_{Setosa}}| - SD_{X(Setosa)} + |y - \overline{y_{Setosa}}| - SD_{Y(Setosa)}, \quad (19)$$

$$Diff_{Versicolor} = (NArea_{Versicolor} - Area_{Versicolor}) + |x - \overline{x_{Versicolor}}| - SD_{X(Versicolor)} + |y - \overline{y_{Versicolor}}| - SD_{Y(Versicolor)}, \quad (20)$$

$$Diff_{Virginica} = (NArea_{Virginica} - Area_{Virginica}) + |x - \overline{x_{Virginica}}| - SD_{X(Virginica)} + |y - \overline{y_{Virginica}}| - SD_{Y(Virginica)}. \quad (21)$$

If $Diff_i$ is the smallest value among the values of $Diff_{Setosa}$, $Diff_{Versicolor}$ and $Diff_{Virginica}$, where $Diff_i \in \{Diff_{Setosa}, Diff_{Versicolor}, Diff_{Virginica}\}$, then the testing instance is classified into the species i, where $i \in \{$Setosa, Versicolor, Virginica$\}$.

3. An example

In this section, we use an example to illustrate the proposed method for dealing with the Iris data classification problem. First, we randomly choose 75 instances from the Iris data as the training instances as shown in Table 2, and let the other 75 instances of the Iris data be the testing instances as shown in Table 3.

[**Step 1**] Based on Table 2, we can obtain the maximum attribute value and the minimum attribute value of each attribute of the training instances belonging to the species Setosa, Versicolor and Virginica, respectively, as shown in Table 4.

[**Step 2**] After calculating the overlapping area formed by the values of each pair of attributes of the three species of the training instances, we get the results shown in Table 5.

[Step 3] From Table 5, we can see that the total overlapping area between the attributes PL and PW is the smallest (i.e., 0.16). Thus, the attributes PL and PW are useful attributes to be used for dealing with the Iris data classification problem, and the attributes SL and SW are useless attributes and are discarded.

[Step 4] By applying formulas (1)-(6), we can obtain the average attribute values of the attributes PL and PW of each species of the training instances, respectively, where $\overline{PL_{Setosa}} = 1.456$, $\overline{PW_{Setosa}} = 0.224$, $\overline{PL_{Versicolor}} = 4.308$, $\overline{PW_{Versicolor}} = 1.352$, $\overline{PL_{Virginica}} = 5.564$, and $\overline{PW_{Virginica}} = 2.076$.

[Step 5] By applying formulas (7)-(12), we can obtain the standard deviations of the attributes PL and PW of each species of the training instances, where $SD_{PL(Setosa)} = 0.202$, $SD_{PW(Setosa)} = 0.081$, $SD_{PL(Versicolor)} = 0.470$, $SD_{PW(Versicolor)} = 0.190$, $SD_{PL(Virginica)} = 0.534$ and $SD_{PW(Virginica)} = 0.273$.

[Step 6] By applying formulas (13)-(15), we can calculate the area formed by the values of the attributes PL and PW of each species of the training instances, where $\text{Area}_{\text{Setosa}} = 0.27$, $\text{Area}_{\text{Versicolor}} = 1.52$ and $\text{Area}_{\text{Virginica}} = 2.64$.

In the following, we choose a testing instance from Table 3 to illustrate how to deal with the classification of the testing instances. For example, we use the first testing instance (4.9, 3, 1.4, 0.2) shown in Table 3, where the value of the attribute PL = 1.4 cm and the value of the attribute PW = 0.2 cm, which belongs to the species Setosa. Based on formulas (16)-(18), we can calculate the new area $\text{NArea}_{\text{Setosa}}$, $\text{NArea}_{\text{Versicolor}}$ and $\text{NArea}_{\text{Virginica}}$ after adding the testing instance into the training instances, where $\text{NArea}_{\text{Setosa}} = 0.27$, $\text{NArea}_{\text{Versicolor}} = 5.60$ and $\text{NArea}_{\text{Virginica}} = 12.65$. Based on formulas (19)-(21), we can get $\text{Diff}_{\text{Setosa}} = 0.21$, $\text{Diff}_{\text{Versicolor}} = 7.48$ and $\text{Diff}_{\text{Virginica}} = 15.23$. Because the value of $\text{Diff}_{\text{Setosa}}$ is the smallest among the values of $\text{Diff}_{\text{Setosa}}$, $\text{Diff}_{\text{Versicolor}}$ and $\text{Diff}_{\text{Virginica}}$, the testing instance (4.9, 3, 1.4, 0.2) is classified into the species Setosa. From Table 3, we can see that it is a correct classification result.

Table 2. Training instances

Setosa				Versicolor				Virginica			
SL	SW	PL	PW	SL	SW	PL	PW	SL	SW	PL	PW
5.1	3.5	1.4	0.2	7	3.2	4.7	1.4	6.3	3.3	6	2.5
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3	5.9	2.1
5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3	5.8	2.2
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
5.4	3.7	1.5	0.2	5	2	3.5	1	6.5	3.2	5.1	2
4.8	3	1.4	0.1	6	2.2	4	1	6.8	3	5.5	2.1
5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.4	3.9	1.3	0.4	5.6	3	4.5	1.5	6.5	3	5.5	1.8
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
4.6	3.6	1	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.4	1.4	0.2	6	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
4.9	3.1	1.5	0.2	5.4	3	4.5	1.5	6.1	2.6	5.6	1.4
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3

Table 3. Testing instances

Setosa				Versicolor				Virginica			
SL	SW	PL	PW	SL	SW	PL	PW	SL	SW	PL	PW
4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.6	3.1	1.5	0.2	5.5	2.3	4	1.3	6.3	2.9	5.6	1.8
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3	6.6	2.1
5	3.4	1.5	0.2	4.9	2.1	3.3	1	7.3	2.9	6.3	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
4.8	3.4	1.6	0.2	5.9	3	4.2	1.5	6.4	2.7	5.3	1.9
4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5	2
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1	7.7	3.8	6.7	2.2
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6	2.2	5	1.5
5.1	3.7	1.5	0.4	6.1	2.8	4	1.3	5.6	2.8	4.9	2
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
5	3	1.6	0.2	6.6	3	4.4	1.4	7.2	3.2	6	1.8
5.2	3.5	1.5	0.2	6.7	3	5	1.7	6.1	3	4.9	1.8
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1	7.2	3	5.8	1.6
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1	7.9	3.8	6.4	2
5.5	4.2	1.4	0.2	6	2.7	5.1	1.6	6.3	2.8	5.1	1.5
5	3.2	1.2	0.2	6	3.4	4.5	1.6	7.7	3	6.1	2.3
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

Table 4. The maximum attribute value and the minimum attribute value of each attribute of each species of the training instances

Attributes / Species	SL		SW		PL		PW	
	Max	Min	Max	Min	Max	Min	Max	Min
Setosa	5.8	4.4	4.1	2.9	1.9	1.0	0.4	0.1
Versicolor	7.0	5.0	3.3	2.0	4.9	3.0	1.8	1.0
Virginica	7.7	4.9	3.4	2.5	6.9	4.5	2.5	1.4

4. Experimental results

Based on the proposed method, we have implemented a program on a Pentium 4 PC by using Jbuilder version 5.0 for dealing with the Iris data classification problem. A comparison of the average classification accuracy rate of the proposed method with that of the existing methods is shown in Table 6. From Table 6, we can see that the proposed method gets a higher average classification accuracy rate than the existing methods.

Table 5. The overlapping areas formed by the values of the attributes between species of the training instances

Overlapping between species Overlapping area Pair of attributes	Setosa and Versicolor	Versicolor and Virginica	Virginica and Setosa	Setosa, Versicolor, and Virginica	Total overlapping area
Attribute SL and attribute SW	0.32	1.60	0.45	0.32	1.72
Attribute SL and attribute PL	0.0	0.80	0.0	0.80	0.80
Attribute SL and attribute PW	0.0	0.80	0.0	0.80	0.80
Attribute SW and attribute PL	0.0	0.32	0.0	0.32	0.32
Attribute SW and attribute PW	0.0	0.32	0.0	0.32	0.32
Attribute PL and attribute PW	0.0	0.16	0.0	0.16	0.16

Table 6. A comparison of the average classification accuracy rates for different methods

Methods	Average classification accuracy rate
Hong-and-Lee’s method [13] (training data set: 75 instances; testing data set: 75 instances; executing 200 runs)	95.57%
Hong-and-Lee’s method [14] (training data set: 75 instances; testing data set: 75 instances; executing 200 runs)	95.57%
Chang-and-Chen’s method [3] (training data set: 75 instances; testing data set: 75 instances; executing 200 runs)	96.07%
Wu-and-Chen’s method [25] (training data set: 75 instances; testing data set: 75 instances; after executing 200 runs)	96.21%
The proposed method (training data set: 75 instances; testing data set: 75 instances; executing 200 times)	96.28%
Castro’s method [2] (training data set: 120 instances; testing data set: 30 instances; executing 200 runs)	96.60%
The proposed method (training data set: 120 instances; testing data set: 30 instances; executing 200 runs)	96.72%
Dasarathy’s method [9] (training data set: 150 instances; testing data set: 150 instances)	94.67%
Hong-and-Chen’s method [15] (training data set: 150 instances; testing data set: 150 instances)	96.67%
The proposed method (training data set: 150 instances; testing data set: 150 instances)	97.33%

5. Conclusions

In this paper, we have presented a new method for handling the Iris data classification problem based on the distribution of training instances. First, we find two useful attributes of the Iris data from the training instances that are more suitable to deal with the classification problem. It means that the distribution of the values of these two useful attributes of the three species (i.e., Setosa, Versicolor and Virginica) has less overlapping. Then, we calculate the average attribute values and the standard deviations of these two useful attributes. We also calculate the overlapping areas formed by the values of these two useful attributes between species of the training instances, the average attribute values, and the standard deviations of the values of these two useful attributes of each species. Then, we calculate the difference between the values of these two useful attributes of each species of the training instances. We choose the species that has the smallest difference between the values of these two useful attributes of the testing instance and the values of these two useful attributes of each species of the training instances as the classification result of the testing instance. We also have implemented a program on a Pentium 4 PC by using Jbuilder version 5.0 for dealing with the Iris data classification problem. The experimental results show that the proposed method gets a higher average classification accuracy rate than the existing methods.

Acknowledgements

This work was supported in part by the National Science Council, Republic of China, under Grant NSC 92-2213-E-011-074.

References

[1] Burkhardt, D. G. and Bonissone, P. P. 1992. Automated fuzzy knowledge base generation and tuning. *Proceedings of the 1992 IEEE International Conference on Fuzzy Systems*, San Diego, California: 179-188, 1992.

[2] Castro, J. L., Castro-Schez, J. J. and Zurita, J. M. 1999. Learning maximal structure rules in fuzzy logic for knowledge acquisition in expert systems. *Fuzzy Sets and Systems*, 101: 331-342.

[3] Chang, C. H. and Chen, S. M. 2001. Constructing membership functions and generating weighted fuzzy rules from training data. *Proceedings of the 2001 Ninth National Conference on Fuzzy Theory and Its Applications*, Chungli, Taoyuan, Taiwan, Republic of China: 708-713.

[4] Chang, C. H. and Chen, S. M. 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 Artificial Intelligence*, Chiayi, Taiwan, Republic of China: 57-64.

[5] Chen, S. M. and Yeh, M. S. 1999. Generating fuzzy rules from relational database systems for estimating null values. *Cybernetics and Systems: An International Journal*, 28: 695-723.

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

[7] Chen, Y. C. and Chen, S. M. 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*, Taipei, Taiwan, Republic of China.

[8] Chen, Y. C. and Chen, S. M. 2001. Constructing membership functions and generating fuzzy rules using genetic algorithms. *Proceedings of the 2001 Ninth National Conference on Fuzzy Theory and Its Applications*, Chungli, Taoyuan, Taiwan, Republic of China: 195-200.

[9] Dasarathy, B. V. 1980. Noise around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2: 67-71.

[10] Fang, Y. D. and Chen, S. M. 2002. A new

method for handling classification problems based on the distribution of training instances. *Proceedings of the Seventh Conference on Artificial Intelligence and Applications,* Taichung, Taiwan, Republic of China: 101-106.

[11] Fisher, R. 1936. The use of multiple measurements in taxonomic problem. *Ann. Eugenics*, 7: 179-188.

[12] Hirsh, H. 1994. Generalizing version space. *Machine Learning*, 17: 5-46.

[13] Hong, T. P. and Lee, C. Y. 1996. Induction of fuzzy rules and membership functions from training examples. *Fuzzy Sets and Systems*, 84: 33-47.

[14] Hong, T. P. and Lee, C. Y. 1999. Effect of merging order on performance of fuzzy induction. *Intelligent Data Analysis,* 3: 139-151.

[15] Hong, T. P. and Chen, J. B. 1999. Finding relevant attributes and membership functions. *Fuzzy Sets and Systems,* 103: 389-404.

[16] Hong, T. P. and Chen, J. B. 2000. Processing individual fuzzy attributes for fuzzy rule induction. *Fuzzy Sets and Systems*, 112: 127-140.

[17] Kibler, D. and Langley, P. 1988. Machine learning as an experimental science. *Proceedings of the European Working Session of Learning*: 87-92.

[18] Ishibuchi, H., Nozaki, K., and Tanaka, H. 1992. Distributed representation of fuzzy rules and it's application to pattern classification. *Fuzzy Sets and Systems,* 52: 21-32.

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

[20] Lin, H. L. and Chen, S. M. 2001. A new method for generating weighted fuzzy rules from training instances using genetic algorithms. *Proceedings of the 6th Conference on Artificial Intelligence and Applications*, Kaohsiung, Taiwan, Republic of China: 628-633.

[21] Nomura, H., Hayashi, I., and Wakami, N. 1992. A learning method of fuzzy inference rules by descent method. *Proceedings of the 1992 IEEE International Conference on Fuzzy Systems*, San Diego, California: 203-210.

[22] Quinlan, J. R. 1987. Simplifying decision trees. *International Journal of Man-Machine Studies,* 27: 221-234.

[23] Quinlan, J. R. 1993. *"C4.5 Programs for Machine Learning*". Morgan Kaufmann, California.

[24] Sudkamp, T. and Hammell II, R. J. 1994. Interpolation, completion, and learning fuzzy rules. *IEEE Transactions on Systems, Man, and Cybernetics,* 24: 332-342.

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

[26] Wang, C. H., Liu, J. F., Hong, T. P., and Tseng, S. S. 1999. A fuzzy inductive learning strategy for modular rules. *Fuzzy Sets and Systems*, 103: 91-105.

[27] Wang, L. X. and Mendel, J. M. 1992. Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics,* 22: 1414-1427.

[28] Zadeh, L. A. 1965. Fuzzy sets. *Information and Control*, 8: 338-353.*Computer Symposium: Workshop on Artificial Intelligence*, Chiayi, Taiwan, Republic of China: 11-18.