Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Extracting classification rules based on a cumulative probability distribution approach

Jr-shian CHEN

(*Department of Computer Science and Information Management, Hungkuang University, Taiwan 433, Taichung*)

E-mail: jschen@sunrise.hk.edu.tw

Received June 18, 2010; Revision accepted Oct. 8, 2010; Crosschecked Mar. 31, 2011

Abstract: This paper deals with a reinforced cumulative probability distribution approach (CPDA) based method for extracting classification rules. The method includes two phases: (1) automatic generation of the membership function, and (2) use of the corresponding linguistic data to extract classification rules. The proposed method can determine suitable interval boundaries for any given dataset based on its own characteristics, and generate the fuzzy membership functions automatically. Experimental results show that the proposed method surpasses traditional methods in accuracy.

Key words: Cumulative probability distribution approach (CPDA), Classification rule, C4.5
doi:10.1631/jzus.C1000205 **Document code:** A **CLC number:** TP311

1 Introduction

Data discretization can improve predictions by reducing the search space or noise, and by pointing to important data characteristics. Data discretization is a commonly sought solution, and can be considered preprocessing of data for the classification problem.

The discretization technique treats a quantitative attribute as a qualitative attribute. There are many advantages in doing so; for example, data can be reduced and simplified. Discrete attributes are usually more compact, shorter, and more accurate than continuous ones (Liu *et al.*, 2002). However, conventional procedures have many shortcomings. For example, they need to predefine membership functions (Hong and Lee, 1996) in discretization procedures. The predefined membership functions are generated according to experiential results or a subjective decision from domain experts. Although domain experts have played, and will still play, an important role in the development of conventional studies, automatically generating membership functions from examples is very helpful when domain experts are not available, and may even provide information not previously known by experts. Previous studies partitioned the attribute interval into equal lengths and ignored the distribution characteristics of datasets.

Therefore, this paper focuses on improving the persuasiveness in determining the universe of discourse and generating the fuzzy membership functions automatically. In the empirical case study, we use an exemplary dataset as the simulation data, which contains the learning achievement data of 50 students. Experimental results show that the proposed approach is better than many other approaches.

2 Background

In this section, we briefly discuss the following research: fuzzy numbers, classification rules, and our prior research with the cumulative probability distribution approach (CPDA) (Teoh *et al*., 2008).

2.1 Cumulative probability distribution approach

CPDA partitions (Teoh *et al*., 2008) the universe of discourse and builds membership functions. It is based on the inverse of the normal cumulative distribution function. In probability theory, the inverse of the normal cumulative distribution function (CDF) is modeled by two parameters μ and σ for a given probability p. The CDF is defined as follows:

$$x = F^{-1}(p \mid \mu, \sigma) = \{x : F(x \mid \mu, \sigma) = p\}, \qquad (1)$$

where

$$p = F(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{-(t-\mu)^2}{2\sigma^2} \right) \mathrm{d}t, \qquad (2)$$

and μ and σ denote the mean and the standard deviation, respectively.

2.2 Fuzzy numbers

Zadeh (1965) introduced the concept of a fuzzy set for modeling the vagueness type of uncertainty (Ross, 2004). A fuzzy set $\tilde{A}$ defined on the universe X is characterized by a membership function $\mu_{\tilde{A}} : x \to [0, 1]$, which satisfies the following condition: (1) $\mu_{\tilde{A}}$ is interval continuous, (2) $\mu_{\tilde{A}}$ is a convex, and (3) $\mu_{\tilde{A}}$ is a normalized fuzzy set and $\mu_{\tilde{A}}(m) = 1$, where m is a real number (Chou *et al.*, 2010).

When describing imprecise numerical quantities, one should capture intuitive concepts of approximate numbers or intervals such as 'approximate *m*'. A fuzzy number must have a unique modal value *m*, convex and piecewise continuous. A common approach is to limit the shape of membership functions defined by left-right (LR) type fuzzy numbers. A special case of LR-type fuzzy numbers, the triangular fuzzy number (TFN), is defined by a triplet, denoted as $\tilde{A} \to (a, b, c)$. Fig. 1 shows the graph of a typical TFN.

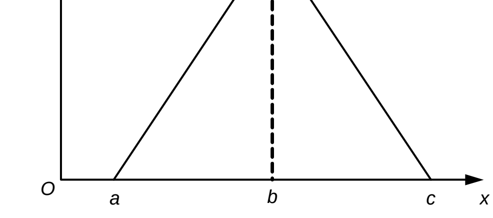

Fig. 1 Triangular fuzzy number

The membership function for this TFN (Fig. 1) is defined as

$$\mu_{\tilde{A}_i}(x) = \begin{cases} 0, & x < a, \\ (x-a)/(b-a), & a \le x \le b, \\ (c-x)/(c-b), & b < x \le c, \\ 0, & x > c, \end{cases} \qquad (3)$$

where $\mu_{\tilde{A}_i}(x)$ denotes the membership value of crisp data x belonging to fuzzy sets $\tilde{A}_i$. The lower bound, the midpoint, and the upper bound intervals of $\tilde{A}_i$ are denoted by *a*, *b*, and *c*, respectively. If the data meet two membership functions, both the maximum membership value and linguistic value are determined.

A trapezoid fuzzy number can be defined as (*a*, *b*, *c*, *d*). The membership function is defined as

$$\mu_{\tilde{A}_i}(X) = \begin{cases} 0, & x < a, \\ (x-a)/(b-a), & a \le x \le b, \\ 1, & b < x \le c, \\ (d-x)/(d-c), & c < x \le d, \\ 0, & x > d. \end{cases} \qquad (4)$$

2.3 Data discretization

Discretization produces a qualitative attribute from a quantitative attribute. Li (2001) provided a new discretization method which is based on the dual partition and the pressure by the piecewise constant. Jia *et al.* (2006) presented a discretization method which can also be used as a data pretreating step for other symbolic knowledge discovery or machine learning methods other than rough set theory. Unlike the other discretization approaches, CPDA is a more objective and reasonable approach to defining the universe of discourse, and it partitions the lengths of intervals depending on the distribution characteristics of observations, using a triangular membership function to fuzzify the observations. The interval boundaries are defined on the distribution characteristics of observations, which may increase the classification accuracy.

2.4 Classification rules

Classification is an important data mining technique. There are many classification models which have been proposed and applied in many fields (Huang and Zhu, 2010; Liu *et al.*, 2010). A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent class or class distributions (Han and Kamber, 2001). The ID3 (Quinlan, 1986) is a decision tree algorithm originating from information theory. The basic strategy used by ID3 is to choose splitting attributes with the highest information gain. The concept used to quantify the information for a corresponding attribute is called entropy. It is defined as follows.

Given a collection set S of c outcomes, the entropy is defined as

$$H(S)=\sum_i -p_i\log_2 p_i, \quad (5)$$

where p_i is the proportion of S belonging to class i.

On the other hand, Gain(S, A) is the information gain of example set S on attribute A and it is defined as

$$\text{Gain}(S,A)=H(S)-\sum_v \frac{|S_v|}{|S|}H(S_v), \quad (6)$$

where v is a value of A, S_v is a subset of S, $|S_v|$ denotes the number of elements in S_v, and $|S|$ denotes the number of elements in S.

C4.5, designed by Quinlan (1993), is an algorithm based on the ID3. C4.5 includes a number of improvements to some issues that ID3 could not overcome: avoiding overfitting the data, determining how deeply to grow a decision tree, reducing error pruning, rule post-pruning, handling continuous attributes, choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs, and improving computational efficiency.

3 The proposed approach

In this section, an enhanced model is proposed to improve the persuasiveness in determining the universe of discourse and membership functions. First, we generate the membership function based on CPDA for each conditional attribute. Second, we use the classification rule algorithm to extract rules (Fig. 2).

The proposed model is stated in more detail as follows.

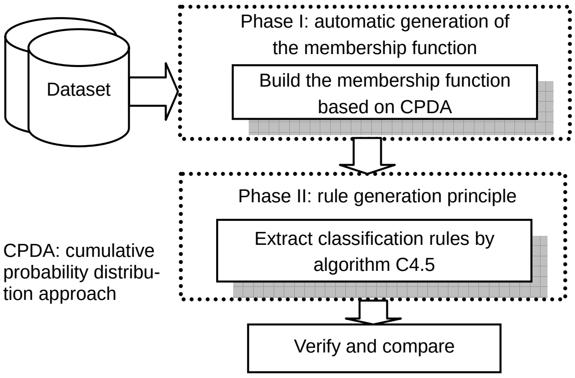
Dataset
Phase I: automatic generation of the membership function
Build the membership function based on CPDA
Phase II: rule generation principle
Extract classification rules by algorithm C4.5
CPDA: cumulative probability distribution approach
Verify and compare

Fig. 2 Framework of the proposed procedure

Step 1: define the universe of discourse.

For each attribute in a dataset, the universe is denoted by $U_i=[D_{\min}-\sigma, D_{\max}+\sigma]$, where $D_{\min}$ and $D_{\max}$ are the minimum and maximum data values in the attribute respectively, and σ is the standard deviation of all data values in the attribute. Extending both sides at the universe of discourse by an amount of σ preserves a variation space and ensures the future results falling in U_i.

Step 2: determine the length of intervals.

The universe of discourse is partitioned into several intervals based on cumulative probability distribution. The lower bound cumulative probability P_{LB_i} and the upper bound cumulative probability P_{UB_i} of each linguistic value are obtained by

$$P_{\text{LB}_i}=(2i-3)/n,\ \ 2\le i\le n, \quad (7)$$

$$P_{\text{UB}_i}=i/n,\ \ 1\le i\le n, \quad (8)$$

where i denotes the order of the linguistic values, and n denotes the number of linguistic values. The lower bound of the first linguistic value and the upper bound of the last linguistic value are directly related to the lower bound and upper bound of universe of discourse, respectively.

The lower bound LB$_i$ and upper bound UB$_i$ are defined as follows:

$$\text{LB}_i=\mu+\sigma x_{P_{\text{LB}_i}}, \quad (9)$$

$$\text{UB}_i=\mu+\sigma x_{P_{\text{UB}_i}}, \quad (10)$$

where $x_{P_{\text{LB}_i}}$ is as defined in Eq. (1).

Step 3: generate the membership functions automatically.

The CPDA is used to discretize the dataset and generate the membership functions. This algorithm partitions the universe of discourse based on cumulative probability distribution. The membership function is built by using a triangular fuzzy number.

Step 4: fuzzify the continuous data into a unique corresponding linguistic value.

According to the membership function in Step 3, the degree of membership for each data value is calculated. The maximal degree of the membership for each data value determines its unique corresponding linguistic value.

Step 5: extract classification rules by C4.5.

From the results obtained in Step 4, classification rules can be built using the C4.5 algorithm (Weka open software (Witten and Frank, 2005)).

Step 6: verify and compare.

In the last step, the classification model derived from Step 5 is evaluated. The accuracy of the proposed approach is compared with those of some existing approaches.

4 Experimental dataset

To verify the proposed approach, there are two datasets in the experiments: (1) SAP50A dataset, (2) Glass dataset.

4.1 SAP50A dataset

An exemplary dataset named SAP50A (Rasmani and Shen, 2006) was used throughout this study (Table 1). The dataset contains 50 instances, involving three conditional attributes: assignment, test, and final exam. All these attributes are numerical values. The proposed approach in Section 3 is applied to the dataset step by step.

Step 1: define the universe of discourse.

The standard deviation, minimum, and maximum data of 'assignment' are 26.59, 5, and 100, respectively. Hence, the universe of discourse, *U*, is defined as [5−26.59, 100+26.59]. The universe of discourse for each attribute is listed in Table 2.

Step 2: determine the length of intervals.

According to the inverse of normal CDF, the lower bound, midpoint, and upper bound are calculated as the triangular fuzzy numbers of each linguistic value.

Table 1 The SAP50A dataset for illustrating classification rules generation

Case	Assignment	Test	Final exam	Final mark	Grade
1	5.00	37.00	18.00	20.00	E
2	10.00	23.00	16.00	16.33	E
3	15.00	13.00	6.00	11.33	E
…	…	….	…	…	…
48	95.00	97.00	98.00	96.67	A
49	90.00	93.00	94.00	92.33	A
50	100.00	83.00	98.00	93.67	A

E: unsatisfactory; A: excellent

Table 2 Universe of discourse for the SAP50A dataset

Attribute	Min	Max	Mean	STD	Universe
Assignment	5	100	48.38	26.59	[−21.59, 126.59]
Test	10	97	51.56	25.12	[−15.12, 122.12]
Final exam	4	98	53.50	26.82	[−22.82, 124.82]

STD: standard deviation

For example, the LB_2 of 'assignment' is generated by Eq. (9):

$$\mathrm{LB}_2 = \mu + \sigma x_{P_{\mathrm{LB}_2}} = 48.38 + 26.59 \times (-1.28) = 14.34.$$

The results of Step 2 are shown in Table 3.

Step 3: generate the membership functions automatically.

The TFN defined in Eq. (3) was used to present the fuzzy sets for the linguistic variable based on the linguistic intervals from Step 2. Table 4 lists the membership functions obtained by the CPDA approach, which refer to the following linguistic values: excellent (L5), very good (L4), good (L3), satisfactory (L2), and unsatisfactory (L1). These membership functions are shown in Figs. 3a−3c.

Step 4: fuzzify the continuous data into a unique corresponding linguistic value.

According to the membership functions in Step 3, the degree of membership for each datum is calculated. For example, for the membership degree for the test score in Case 2, L1 is 0.33 and L2 is 0.28. Table 5 shows the results of this experiment.

The test score in Case 2 is 23, which falls within two scopes, and the membership degree is 0.33 for L1 and 0.28 for L2. Then the test score 23 is labeled as L1 based on the larger membership degree. The maximal degree of the membership for each datum is calculated to determine its linguistic value. Table 6 shows the results of the dataset.

Table 3 Linguistic values and intervals of the cumulative probability distribution approach for the SAP50A dataset

Attribute	Linguistic value	P_{LB_i}	P_{UB_i}	LB	Midpoint	UB	Length of interval
Assignment	L1	–	0.2	−21.60	2.20	26.00	47.60
	L2	0.1	0.4	14.30	27.97	41.64	27.34
	L3	0.3	0.6	34.43	44.78	55.12	20.69
	L4	0.5	0.8	48.38	59.57	70.76	22.38
	L5	0.7	–	62.33	94.46	126.59	64.26
Test	L1	–	0.2	−15.12	7.65	30.42	45.54
	L2	0.1	0.4	19.37	32.29	45.20	25.83
	L3	0.3	0.6	38.39	48.16	57.92	19.53
	L4	0.5	0.8	51.56	62.13	72.70	21.14
	L5	0.7	–	64.73	93.42	122.12	57.39
Final exam	L1	–	0.2	−22.82	4.05	30.93	53.75
	L2	0.1	0.4	19.13	32.92	46.71	27.58
	L3	0.3	0.6	39.44	49.87	60.29	20.85
	L4	0.5	0.8	53.50	64.79	76.07	22.57
	L5	0.7	–	67.56	96.19	124.82	57.26

L1: unsatisfactory; L2: satisfactory; L3: good; L4: very good; L5: excellent. LB: lower bound; UB: upper bound. –: The lower bound of the first linguistic value and the upper bound of the last linguistic value directly correspond to the lower bound and upper bound of universe of discourse, respectively

Table 4 The membership functions of all attributes

Linguistic value	Assignment	Test	Final exam
L1	(0.00, 2.20, 26.00)	(0.00, 7.65, 30.42)	(0.00, 4.05, 30.93)
L2	(14.30, 27.97, 41.64)	(19.37, 32.29, 45.20)	(19.13, 32.92, 46.71)
L3	(34.43, 44.78, 55.12)	(38.39, 48.16, 57.92)	(39.44, 49.87, 60.29)
L4	(48.38, 59.57, 70.76)	(51.56, 62.13, 72.70)	(53.50, 64.79, 76.07)
L5	(62.33, 94.46, 100.00)	(64.73, 93.42, 100.00)	(67.56, 96.19, 100.00)

L1: unsatisfactory; L2: satisfactory; L3: good; L4: very good; L5: excellent

Table 5 The partial degrees of the membership functions for each datum

Case	Assignment					Test					Final exam				
	L1	L2	L3	L4	L5	L1	L2	L3	L4	L5	L1	L2	L3	L4	L5
1	0.88	0.00	0.00	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.00
2	0.67	0.00	0.00	0.00	0.00	0.33	0.28	0.00	0.00	0.00	0.55	0.00	0.00	0.00	0.00
3	0.46	0.05	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00
…	…	…	…	…	...	…	...	...	...	…	…	…	...	...	...
48	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
49	0.00	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.92
50	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.00	1.00

L1: unsatisfactory; L2: satisfactory; L3: good; L4: very good; L5: excellent

Step 5: extract rules by C4.5.

We can extract the classification rules in detail as follows:

Rule 1: IF final exam is L1 THEN Grade is Unsatisfactory.

Rule 2: IF final exam is L2 THEN Grade is Satisfactory.

Rule 3: IF final exam is L3 THEN Grade is Good.

Rule 4: IF final exam is L4 and test is L1 or L2 or L3 THEN Grade is Good.

Rule 5: IF final exam is L4 and test is L4 or L5 THEN Grade is Very Good.

Rule 6: IF final exam is L5 and test is L3 or L4 THEN Grade is Very Good.

Rule 7: IF final exam is L5 and test is L1 or L2 or L5 THEN Grade is Excellent.

Step 6: verify and compare.

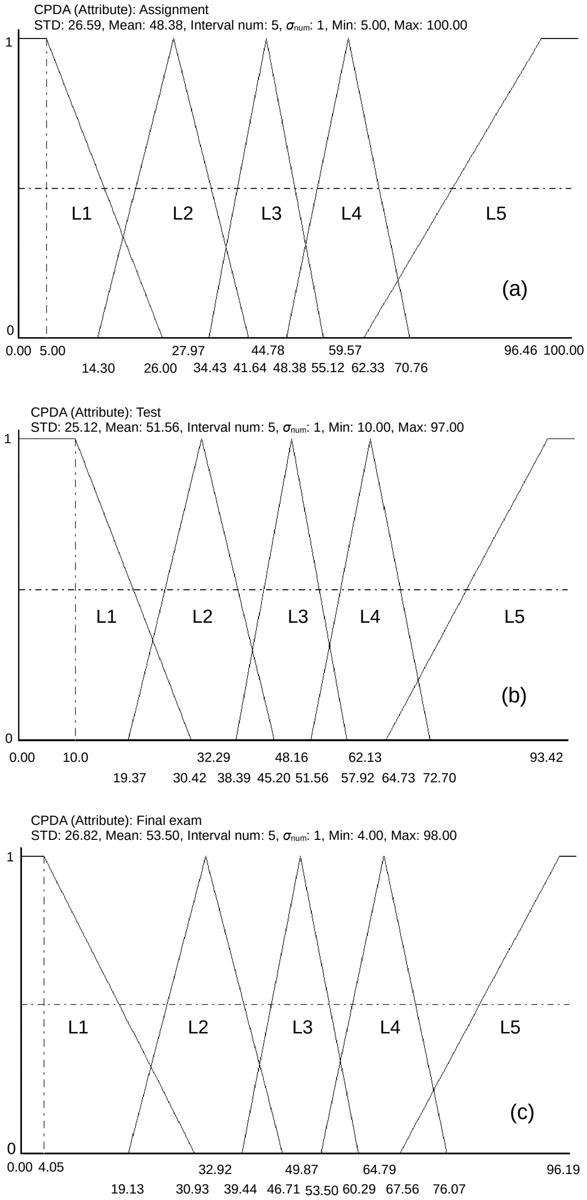

Fig. 5 Membership functions of the assignment (a), test (b), and final exam (c) attributes

L1: unsatisfactory; L2: satisfactory; L3: good; L4: very good; L5: excellent

Table 6 The partial linguistic values for each datum

Case	Assignment	Test	Final exam
1	L1	L2	L1
2	L1	L1	L1
3	L1	L1	L1
…	…	…	…
48	L5	L5	L5
49	L5	L5	L5
50	L5	L5	L5

L1: unsatisfactory; L2: satisfactory; L3: good; L4: very good; L5: excellent

To fairly evaluate our proposed method, the same dataset was used for the other five traditional methods. The details of the test dataset are shown in Table 7, cited from Rasmani and Shen (2006). In addition, there are five approaches used for comparison. To inspect the forecasting performance for our method, we use classification accuracy rate as a performance indicator in this study.

The classification accuracy rates (Table 8) show that the proposed approach is better than the listed approaches.

Table 7 The SAP50A dataset for testing the classification rules (Rasmani and Shen, 2006)

Case	Assignment	Test	Final exam	Grade
1	10.00	23.33	20.00	L1
2	5.00	16.67	12.00	L1
3	15.00	13.33	18.00	L1
4	45.00	26.67	40.00	L2
5	35.00	33.33	30.00	L2
6	35.00	50.00	38.00	L2
7	45.00	43.33	54.00	L3
8	50.00	40.00	50.00	L3
9	45.00	50.00	58.00	L3
10	50.00	70.00	62.00	L4
11	65.00	70.00	74.00	L4
12	85.00	60.00	76.00	L4
13	95.00	76.67	86.00	L5
14	85.00	83.33	96.00	L5
15	90.00	90.00	98.00	L5

L1: unsatisfactory; L2: satisfactory; L3: good; L4: very good; L5: excellent

Table 8 The SAP50A dataset comparison between the proposed approach and five other approaches

Approach	Classification accuracy rate (%)
Biswas' approach (Biswas, 1995)	60.0
Chen's approach (Chen *et al*., 2001)	66.7
Law's approach (Law, 1996)	86.7
WSBA (Rasmani and Shen, 2006)	93.3
NEFCLASS (Rasmani and Shen, 2006)	80.0
Proposed approach	100.0

WSBA: weighted subsethood-based algorithm; NEFCLASS: description of neuro-fuzzy classification

4.2 Glass dataset

The Wisconsin Glass dataset (Newman *et al*., 1998) was applied to explain our proposed approach. There are 214 instances in the dataset, characterized by the following attributes: (I) RI, (II) Na_2O, (III) MgO, (IV) Al_2O_3, (V) SiO_2, (VI) K_2O, (VII) CaO,

(VIII) BaO, and (IX) Fe_2O_3. All these attributes are real values. There are six classes in the dataset: A, building windows float processed; B, building windows non-float processed; C, vehicle windows float processed; D, vehicle windows non-float processed (none in this dataset); E, containers; F, tableware; G, headlamps. The proposed approach in Section 3 was applied to the dataset.

According to the inverse of normal CDF, the lower bound, midpoint, and upper bound, as the triangular fuzzy numbers of each linguistic value, were calculated (Table 9).

Table 9 Linguistic values and intervals of the cumulative probability distribution approach for the Glass dataset

Attribute	Linguistic value	P_{LB_i}	P_{UB_i}	LB	Midpoint	UB	Length of interval
RI	L1	–	0.2	1.5081	1.5120	1.5158	0.0077
	L2	0.1	0.4	1.5145	1.5160	1.5176	0.0031
	L3	0.3	0.6	1.5168	1.5180	1.5191	0.0023
	L4	0.5	0.8	1.5184	1.5196	1.5209	0.0025
	L5	0.7	–	1.5200	1.5285	1.5370	0.0170
Na_2O	L1	–	0.2	9.9153	11.3187	12.7222	2.8069
	L2	0.1	0.4	12.3638	12.7826	13.2015	0.8377
	L3	0.3	0.6	12.9806	13.2974	13.6143	0.6337
	L4	0.5	0.8	13.4079	13.7507	14.0935	0.6856
	L5	0.7	–	13.8351	16.0149	18.1947	4.3596
MgO	L1	–	0.2	−1.4390	0.0172	1.4734	2.9124
	L2	0.1	0.4	0.8403	1.5801	2.3200	1.4797
	L3	0.3	0.6	1.9299	2.4895	3.0491	1.1192
	L4	0.5	0.8	2.6845	3.2901	3.8957	1.2112
	L5	0.7	–	3.4392	4.6841	5.9290	2.4898
Al_2O_3	L1	–	0.2	−0.2081	0.4088	1.0257	1.2338
	L2	0.1	0.4	0.8066	1.0626	1.3187	0.5121
	L3	0.3	0.6	1.1837	1.3774	1.5711	0.3874
	L4	0.5	0.8	1.4449	1.6545	1.8641	0.4192
	L5	0.7	–	1.7061	2.8521	3.9981	2.2920
SiO_2	L1	–	0.2	69.0373	70.5189	72.0006	2.9633
	L2	0.1	0.4	71.6606	72.0579	72.4552	0.7946
	L3	0.3	0.6	72.2457	72.5462	72.8467	0.6010
	L4	0.5	0.8	72.6509	72.9761	73.3013	0.6504
	L5	0.7	–	73.0562	74.6194	76.1827	3.1265
K_2O	L1	–	0.2	−0.6507	−0.3506	−0.0506	0.6001
	L2	0.1	0.4	−0.3368	−0.0023	0.3322	0.6690
	L3	0.3	0.6	0.1558	0.4089	0.6619	0.5061
	L4	0.5	0.8	0.4971	0.7709	1.0447	0.5476
	L5	0.7	–	0.8383	3.8495	6.8607	6.0224
CaO	L1	–	0.2	4.0102	5.8861	7.7620	3.7518
	L2	0.1	0.4	7.1374	7.8673	8.5973	1.4599
	L3	0.3	0.6	8.2124	8.7645	9.3167	1.1043
	L4	0.5	0.8	8.9570	9.5544	10.1519	1.1949
	L5	0.7	–	9.7015	13.6557	17.6098	7.9083
BaO	L1	–	0.2	−0.4961	−0.3693	−0.2424	0.2537
	L2	0.1	0.4	−0.4607	−0.2057	0.0494	0.5101
	L3	0.3	0.6	−0.0851	0.1078	0.3007	0.3858
	L4	0.5	0.8	0.1750	0.3838	0.5925	0.4175
	L5	0.7	–	0.4352	2.0406	3.6461	3.2109
Fe_2O_3	L1	–	0.2	−0.0972	−0.0610	−0.0248	0.0724
	L2	0.1	0.4	−0.0676	−0.0176	0.0324	0.1000
	L3	0.3	0.6	0.0060	0.0438	0.0816	0.0756
	L4	0.5	0.8	0.0570	0.0979	0.1388	0.0818
	L5	0.7	–	0.1080	0.3576	0.6072	0.4992

L1: unsatisfactory; L2: satisfactory; L3: good; L4: very good; L5: excellent. LB: lower bound; UB: upper bound. –: The lower bound of the first linguistic value and the upper bound of the last linguistic value directly correspond to the lower bound and upper bound of universe of discourse, respectively

From the result, fifty-two rules can be extracted. The first five rules are listed as follows:

Rule 1′: IF BaO is L3 THEN Type is B.

Rule 2′: IF BaO is L4 or L5 THEN Type is G.

Rule 3′: IF BaO is L1 and MgO is L5 and RI is L5 THEN Type is A.

Rule 4′: IF BaO is L1 and MgO is L5 and RI is L3 THEN Type is B.

Rule 5′: IF BaO is L1 and MgO is L2 THEN Type is E.

The classification accuracy rates (Table 10) show that the proposed approach is better than the listed three approaches.

Table 10 The Glass dataset comparison between the proposed approach and other three approaches

Approach	Classification accuracy rate (%)
Agglomerative discretization approach (Grzymala-Busse, 2003)	69.37
Divisive discretization approach (Grzymala-Busse, 2003)	68.22
C4.5	65.89
Proposed approach	70.09

5 Conclusions

In this paper, a reinforced classification rules finding method based on CPDA has been proposed for solving classification problems. This method combines the C4.5 algorithm and our prior work CPDA. It can automatically derive the membership functions from a given training dataset and improve the persuasiveness in determining the universe of discourse and membership functions.

From the experimental results of the proposed method, three contributions are made:

1. The proposed method can determine suitable interval boundaries for any given dataset based on its own characteristics.

2. It can generate the fuzzy membership functions automatically and does not need domain experts' experiences.

3. Using the discretization method in preprocessing can improve the accuracy for classification rules without removing the noise instance.

References

Biswas, R., 1995. An application of fuzzy sets in students' evaluation. *Fuzzy Sets Syst.*, **74**(2):187-194. [doi:10.1016/0165-0114(95)00063-Q]

Chen, S.M., Lee, S.H., Lee, C.H., 2001. A new method for generating fuzzy rules from numerical data for handling classification problems. *Appl. Artif. Intell.*, **15**(7):645-664. [doi:10.1080/088395101750363984]

Chou, H.L., Chen, J.S., Cheng, C.H., Teoh, H.J., 2010. Forecasting tourism demand based on improved fuzzy time series model. *LNCS*, **5990**:399-407. [doi:10.1007/978-3-642-12145-6_41]

Grzymala-Busse, J.W., 2003. A comparison of three strategies to rule induction from data with numerical attributes. *Electron. Notes Theor. Comput. Sci.*, **82**(4):132-140. [doi:10.1016/S1571-0661(04)80712-6]

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

Hong, T.P., Lee, C.Y., 1996. Induction of fuzzy rules and membership functions from training examples. *Fuzzy Sets Syst.*, **84**(1):33-47. [doi:10.1016/0165-0114(95)00305-3]

Huang, P., Zhu, J., 2010. Multi-instance learning for software quality estimation in object-oriented systems: a case study. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **11**(2):130-138. [doi:10.1631/jzus.C0910084]

Jia, P., Dai, J.H., Chen, W.D., Pan, Y.H., Zhu, M.L., 2006. Immune algorithm for discretization of decision systems in rough set theory. *J. Zhejiang Univ.-Sci. A*, **7**(4):602-606. [doi:10.1631/jzus.2006.A0602]

Law, C.K., 1996. Using fuzzy numbers in educational grading system. *Fuzzy Sets Syst.*, **83**(3):311-323. [doi:10.1016/0165-0114(95)00298-7]

Li, D.M., 2001. Finite volume method based on the Crouzeix-Raviart element for the Stokes equation. *J. Zhejiang Univ.-Sci.*, **2**(2):165-169. [doi:10.1631/jzus.2001.0165]

Liu, H., Hussain, F., Tan, C., Dash, M., 2002. Discretization: an enabling technique. *Data Min. Knowl. Disc.*, **6**(4):393-423. [doi:10.1023/A:1016304305535]

Liu, Y.M., Ye, L.B., Zheng, P.Y., Shi, X.R., Hu, B., Liang, J., 2010. Multiscale classification and its application to process monitoring. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **11**(6):425-434. [doi:10.1631/jzus.C0910430]

Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J., 1998. UCI Repository of Machine Learning Databases. Available from http://www.ics.uci.edu/~mlearn/ [Accessed on July 25, 2007].

Quinlan, J.R., 1986. Induction of decision trees. *Mach. Learn.*, **1**(1):81-106. [doi:10.1007/BF00116251]

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Rasmani, K.A., Shen, Q., 2006. Data-driven fuzzy rule generation and its application for student academic performance evaluation. *Appl. Intell.*, **25**(3):305-319. [doi:10.1007/s10489-006-0109-9]

Ross, T.J., 2004. Fuzzy Logic with Engineering Applications. John Wiley & Sons, Ltd., USA.

Teoh, H.J., Cheng, C.H., Chu, H.H., Chen, J.S., 2008. Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets. *Data Knowl. Eng.*, **67**(1):103-117. [doi:10.1016/j.datak.2008.06.002]

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, San Francisco.

Zadeh, L.A., 1965. Fuzzy sets. *Inform. Control*, **8**(3):338-353. [doi:10.1016/S0019-9958(65)90241-X]