International Journal of Database Theory and Application
Vol.7, No.5 (2014), pp.201-210
http://dx.doi.org/10.14257/ijdta.2014.7.5.15

Research on Dynamic Cost-sensitive Decision Tree for Mining Uncertain Data Based on the Genetic Algorithm

Yuwen Huang [1,2]

[1]*Department of Computer and Information Engineering, Heze University, Heze 274015, Shandong, China*
[2] *Key Laboratory of computer Information Processing, Heze University, Heze 274015, Shandong, China*
hzxy_hyw@163.com

Abstract

The existing classifiers for uncertain data don't consider the dynamic cost, so this paper proposes the classification approach of the dynamic cost-sensitive decision tree for uncertain data based on the genetic algorithm (GDCDTU) , which overcomes the limitations of the stationary cost, and searches automatically the suitable cost space of every sub datasets. Firstly, this paper gives the dynamic cost-sensitive learning thought, and disposes the continuous and discrete attributes for uncertain data by the probabilistic cardinality. Secondly, we give the selection methods for the splitting attributes and the construction process for cost-sensitive decision tree, and the interval number for describing dynamic cost is coded by its centre and radius. At last, the dynamic cost-sensitive decision tree for uncertain data is structured, which uses the genetic algorithm as the optimal misclassification cost searching way, and the optimum cost is got by the hybridization, the mutation, the selection. The experiments using both artificial and real data sets show that, compared to the other decision tree classification algorithms for uncertain data, GDCDTU has higher classification accuracy and performance, and the total expenditure is lower.

Keywords: *Dynamic cost-sensitive; Decision tree; Uncertain data; Genetic algorithm*

1. Introduction

There are more and more uncertain data in the wireless sensor networks, the radio frequency identification, the astronomy and earth science, etc. The traditional methods for data mining are only suitable for the accurate data, and they don't consider the uncertain data. With uncertain data increasing, we need data mining that takes into account of uncertainty urgently. The research of uncertain data mining is a forthcoming area in data mining. At present, the main classifiers for uncertain data have the support vector machine (SVM), the extend Bayesian classification, the decision tree method. The paper [1] introduced a novel classification algorithm for uncertain data based on Bayesian, which adopted probabilistic and statistical theory for uncertain data. Liu proposed a classification rule for uncertain data based on belief functions, which the meta-classes is introduced to structure the uncertain data [2]. The paper proposed a naive classifier for uncertain data [3]. Sun proposed a learning machine classification algorithm for uncertain data [4]. The paper [5] introduced a frequent pattern mining for uncertain data. Liu, *etc.*, proposed a classification approach for uncertain data by

ISSN: 2005-4270 IJDTA

the evidence theory [6]. Qin introduced a decision tree classification algorithm for uncertain data (DTU) [7]. The paper [8] introduced an approach for decision tree for uncertain data (UDT). The paper [9] proposed a rule-based classification algorithm for uncertain data. Liang, etc proposed a very fast decision tree for uncertain data streams with positive and unlabeled samples [10]. The current researches for uncertain data mining are in pursuit of high accuracy and low error, and the cost factors are less considered. If data mining tasks involve the different cost, the existing mining approaches are inefficient. The mainstream cost-sensitive data mining must exactly know the real data in the mining process, and can't effectively deal with the uncertain data. At present, the classification researches for uncertain data don't connect with the cost-sensitive learning. The cost-sensitive learning is a very important direction in data mining, and more and more research on classification connects with the cost-sensitive learning. This paper proposes the classification method of the dynamic cost-sensitive decision tree for uncertain data, which overcomes the limitations of the stationary cost, and searches automatically suitable misclassification costs for every sub datasets, and combines with the genetic algorithm to optimize the dynamic cost.

2. Dynamic Cost-sensitive Learning

In recent years, cost-sensitive learning is a hot research in data mining, and it minimizes the total cost [11]. The misclassification and test cost influence most on the classification results, and the current researches for the cost-sensitive learning focus on them [12]. Misclassification Cost is caused by the classification error, and test cost is the expense of testing attribute value [13]. At present, the research of cost-sensitive machine learning concerns mainly on the minimization for classification cost, and the formula based on Bayes risk is as follows.

$$\mathrm{R}\left(\alpha_i \mid x\right) = \min \arg \sum_{j=1}^{c} \lambda\left(\alpha_i \mid w_j\right) P\left(w_j \mid x\right)$$

$\mathrm{R}\left(\alpha_i \mid x\right)$ is conditional risk for taking the observation sample x . If w_j is the real class of x , the risk cost is $\lambda\left(\alpha_i \mid w_j\right)$. $P\left(w_j \mid x\right)$ is the prior probability.

In practical application, the addition of test attributes number leads to the increase of test cost, and the total costs including test cost and misclassification also rise. The function of test costs is made by each attribute $c_i \in A$.

$$Test\cos t(A) = \sum_{i=1}^{|A|} T\left(c_i\right).$$

$T\left(c_i\right)$ is cost of each test attribute. The total cost $sum\_\cos t(x, A)$ is constituted by $test\_\cos t(A)$ and $mis\_\cos t(x, A)$.

$$sum\cos t(x, A) = mis\_\cos t(x, A) + test\_\cos t(A).$$

The important research for cost-sensitive classification is how to obtain the minimum classification cost, and to balance the different classification cost. At present, the cost-sensitive learning adopts the stationary static cost, but the static mechanism has some deficiencies in the imbalanced sets, and the performance of the classifier falls fast. The dynamic cost-sensitive learning can solve effectively imbalanced classification, and it isn't sensitive to the bias classifications.

In practical applications, the misclassification and test cost is different under different environments, so the cost is dynamic and unfixable, and we can sort the different misclassification and test cost. At last, the dynamic cost space is structured by the experience and knowledge of the experts in application field. For example, the blood tests cost is not the same in different hospitals. If $[tc_1, tc_2, ..., tc_n]$ is the sequence from small to large, the closed interval $[tc_1, tc_n]$ is test cost space. At the same way, misclassification cost space is structured. Therefore, it is reason to believe that the cost in the closed interval is feasible.

3. Description for Uncertain Data

If $[a,b]$ is the interval for uncertain attribute, its probability distribution function is $f(x)$, and the probability of the real value in interval $[a,b]$ is $\int_a^b f(x)dx$. The probabilistic cardinality of each example is the probability sum that belongs to the interval$[a,b]$.

3.1 Uncertain Discrete Attribute

A_i is the attribute, and its range is $\{v_{i1}, v_{i2}, ..., v_{il}\}$. The probability vector $P = \{p_{i1}, p_{i2}, ..., p_{il}\}$. $P(A_i = v_{ik}) = p_{ik}$, $\sum_{k=1}^{l} p_{ik} = 1$. There is a special situation for the fixed value, $P(A_i = v_{ik}) = 1$, the other probability is zero.

In training sets, the probabilistic cardinality of v_{ik} is defined as follows.

$$PC(v_{ik}) = \sum_{j=1}^{n} p(A_{ij} = v_{ik}) \ .$$

When A_{ij} is v_{ik} in training sets, the probabilistic cardinality of classification C_j is defined as follows.

$$PC(v_{ik}, C_j) = \sum_{j=1}^{n} p\left(A_{ij} = v_{ik} \wedge C_{T_j} = C_j\right).$$

3.2. Uncertain Continuous Attribute

$A_{ij} \in [a,b]$ is continuous attribute, and its value is a range or interval. $f(x)$ is the density function of A_{ij}, and the probabilistic cardinality of the whole training sets in $[a,b]$ is as follows.

$$PC(P_{ab}) = \sum_{j=1}^{n} p_{ab}\left(A_{ij} \in [a,b]\right) = \sum_{j=1}^{n} \int_a^b f(x)dx$$

C_{T_j} is the classification of example T_j, and the probabilistic cardinality of classification C_j in the interval $[a,b]$ is defined as follows.

$$PC\left(p_a, C_j\right) = \sum_{j=1}^{n} p\left(A_{ij} \in [a, b] \wedge C_{T_j} = C_j\right).$$

4. Dynamic Cost-sensitive Decision Tree for Mining Uncertain Data Based on the Genetic Algorithm

4.1. Chromosome Coding

$A = \{A_1, A_2, ..., A_n\}$. The test cost of A_i is the interval $[tc_{il}, tc_{ih}]$. The matrix C is misclassification cost. $\mathrm{C} = \{\mathrm{C}_1, \; C_2, ..., C_m\}$, $\mathrm{C}_{\mathrm{i}} = (c_{i1}, c_{i2}, c_{i3}, ..., c_{im})$, $c_{ij} \in \left[c_{ijl}, c_{ijh}\right]$. If coding directly the upper and lower bound for the interval, the requirement that the lower bound is less than the upper isn't met, so the centre and radius of the interval are coded. ac_i is the centre of test cost, and ar_i is its radius. cc_{ij} is the centre of misclassification cost, and cr_{ij} is its radius. This paper adopts the real code, and each chromosome code is as follows.

$$\underbrace{ac_1, ar_1, ac_2, ar_2, ..., ac_n, ar_n}_{test\ cost},$$

$$\underbrace{cc_{11}, cr_{11}, cc_{12}, cr_{12}, ..., \mathrm{cc}_{1m}, cr_{1m}}_{mistake\ test},$$

$$\underbrace{cc_{21}, cr_{21}, cc_{22}, cr_{22}, ..., \mathrm{cc}_{2m}, \mathrm{cr}_{2m}}_{mistake\ test},$$

....

$$\underbrace{cc_{m1}, cr_{m1}, cc_{m2}, cr_{m2}, ..., \mathrm{cc}_{mm}, cr_{mm}}_{mistake\ test}$$

When decoding the chromosome, the lower bound $c_{ijl} = cc_{ij} - cr_{ij}$, the upper bound $c_{ij\mathrm{h}} = cc_{ij} + cr_{ij}$, the lower bound $tc_{il} = ac_{ij} - ar_{ij}$, and the upper bound $c_{ij\mathrm{h}} = cc_{ij} + cr_{ij}$.

4.2. Selection of Splitting Attribute

Because the structure of decision tree needs to select the splitting attribute, it is the key how to select it in decision tree algorithm. Ling selected the splitting attribute by the reduction of test and misclassification and replaced the information gain in C4.5 [14]. The paper [15] proposed a cost-sensitive attribute selection approach, which used a new cost-sensitive fitness function based on histogram comparison. The paper [16] proposed a cost-sensitive decision tree approach.

In the interest of the total cost, this paper selects the splitting attribute by the cost decrement for misclassification and total costs before and after splitting, the cost formula is as follows.

$$\cos t\_ruduce(A_i) = Mis\cos t\_befor(D) - total\_\cos t(D)$$

$$= Mis\cos t\_before(D) - \left(Mis\cos t\_after(D, \; A_i) + Test\_\cos t(D, A_i)\right)$$

$Mis\cos t\_befor(D)$ is the misclassification cost before splitting, and $Mis\cos t\_after(D)$ is all splitting subclass cost according to A_i after splitting. $Test\_\cos t(D, A_i)$ is test cost of attribute A_i in all examples. $Mis\cos t\_before(D)$ is the same when selecting the splitting attribute. $Mis\cos t\_after(D)$ is as follows.

$$Mis\cos t\_after(D, A_i) = \sum_{j=1}^{|DOM(A_i)|} Mis\cos t\_befor(S_i)$$

$$= \sum_{j=1}^{|DOM(A_i)|} \sum_{k}^{m} \left(PC\left(S_{jk}\right) \times Matrix\_\cos t_{jk}\right)$$

S_i is the subclass with number i by attribute A_i after splitting. $PC\left(S_{jk}\right)$ is the probabilistic cardinality of subclass S_j. $Matrix\_\cos t_{jk}$ is the misclassification cost when class j mistakes class k.

If $\cos t\_ruduce(A_i) \succ 0$, the total cost reduces after attribute selection. Otherwise, the attribute A_i can't be selected. If more attributes are $\cos t\_ruduce(A_i) \succ 0$, select the maximum as the splitting attribute. When all attributes meet $\cos t\_ruduce(A_i) \prec 0$, the attribute test can’t reduce the total cost, and stop the splitting, at the same time, the attribute A_i is labeled as the leave node. The leaves are not marked by the classification, but are saved the total probabilistic cardinality ratio of the all classes. In other word, we can know the proportion of each category form the leaves nodes.

4.3. Structure of Cost-sensitive Decision Tree for Uncertain Data

There are many approaches of structuring the cost-sensitive decision tree. The paper [17] proposed a cost-sensitive decision tree approach, and misclassification costs are taken as varying. Zhang proposed a sensitive tree decision tree classifier [18]. Sheng proposed a cost-sensitive decision tree based on simple test [19]. DUT and UDT are good approaches for uncertain data, and this paper uses the thought of probabilistic cardinality in DTU and UDT to structure the decision tree for uncertain data. Therefore, this paper searches dynamically the cost space, and the steps of structuring the dynamical cost-sensitive tree are as follows.

```
Import
    $D$ : Initial data set.
    $TC$ : Dynamic list space for test cost, for example, an element is $[tc_{il}, tc_{ih}]$ .
    $Matrix$ ：Dynamic matrix space for misclassification cost, for example, an element is $\left[c_{ijl}, c_{ijh}\right]$ .
    Export: Cost-sensitive Decision Tree $M$ .
    CSDTUD ( $D$ , $TC$ , $Matrix$ , $M$ )
    Begin
    1. Create the node $N$ .
```

```
  2. If (all elements of $D$ are the same class $C$ )
    The $N$ is the leaf node, and its class is $C$ .
      Else if (the test list is empty, and the all elements of $D$ have $k$ types )
      Then the $N$ is the leaf node, and save $pc_1(D_1), pc_2(D_2), ..., pc_k(D_k)$ in $N$ .
  3. Calculate $\cos t\_ruduce$ of each attribute.
  4. If (the maximum $\cos t\_ruduce$ is less than zero, and the all elements of $D$ have $k$ types)
    Then $N$ is the leaf node, and save the $pc_1(D_1), pc_2(D_2), ..., pc_k(D_k)$ in $N$ .
  Else if (the attribute of the maximum $\cos t\_ruduce$ is uncertain)
  Then split the uncertain attribute $a_i$ for the interval $(a_{i1}, a_{i2}, ..., \mathrm{a}_{\mathrm{in}})$ .
    For (k=1; k<=n; k++)
      Build the branches $D_{ik}$ for $a_{ik}$ .
  For((k=1;k<=n; k++)
    For(each element $R_j$ in $D$ )
      Calculate the probabilistic Cardinality $R_j.a_{ik}.p$ .
Label $D_i$ as $\sum_{R_j \in D} R_j.a_{ik}.p$ .
  for(each $D_i$ )
CSDTUD ( $D_i$ , $TC$ , $Matrix$ , $M$ )
End.
```

An example may pass through multiple paths in calculating its classification. If the example T passes paths numbers m , the probability of belonging to classification c_i is as follows.

$$P_{C_i} = \sum_{i=1}^{m} P_{c_i}^{i} .$$

The example T is forecasted as classification c_i of the maximum P_{C_i} .

4.4. Fitness Function

In feasible dynamic cost space, finds the optimal misclassification and test cost, and the total cost is reduced as much as possible at the same time. In order to validate the practicability, Response ration (Re) and precision ration (Pr) of the optimal cost should been considered.

$$\mathrm{Re} = \frac{TP}{TP + FN} \times 100\%$$

$$\Pr = \frac{TP}{TP + FP} \times 100\%$$

Expand the G-mean algorithm, and the fitness function is as follows.

$$f(C_x) = \alpha \times \arg\max_{c_x \in C}\left(\sqrt{\sum_{k=1}^{m} \mathrm{Re}_k(C_x) \times \mathrm{Pr}_k(C_x)}\right) + \beta \times \arg\min_{c_x \in C}\left(\sqrt{text\_\cos t(C_x) \times mis\_\cos t(C_x)}\right)$$

C_x is a point in the cost space, and m is the total classification number. $\mathrm{Re}_k(C_x)$ is the response rate, and $Pr_k(C_x)$ is the precision rate after applying C_x.

4.5 Classification Algorithm of Cost-sensitive Decision Tree for Uncertain Data

For the sake of balancing both the majority and minority class, the paper proposes a dynamic cost approach based on the genetic algorithm, which combines the practical application, searches automatically the cost space made by every sub datasets, and then structures a classifier for cost-sensitive decision tree. The steps of classification algorithm for cost-sensitive decision tree are as follows.

Import：

D ：Initial data set.

TC : Matrix for test cost, and an element is $[tc_{il}, tc_{ih}]$.

$Matrix$: Matrix for misclassification cost, and an element is $\left[c_{ijl}, c_{ijh}\right]$.

P_c , P_m : Probability of crossover and mutation.

T : Termination condition of Genetic algorithm.

Export: Decision tree M for uncertain data, the optimal test cost tc , the optimal misclassification cost mc .

GDCDTU$(D, TC, Matrix, P_c, P_m, \varepsilon, tc, mc, M, \ T)$

1. Initialize: Generate randomly an initial population with chromosome number N , T, P_c , P_m , TC , $Matrix$.

2. Take chromosomes number N in population. N is taken as data set D, and use decision tree algorithm to structure classifier M . Pass 10-fold cross validation, and obtain Re, Pr. Utilize the fitness function to calculate the value of chromosomes with number N .

3. Selecting operation: calculate selective probability for each chromosome. $p_i = \dfrac{fitness(x_i)}{\sum_{j=1}^{m} fitness(x_j)}$, $i = 1, 2, \ldots, m$. Select randomly some chromosome by roulette to structure a new population New_Pop .

4. Crossover operation：Select the crossover algorithm, and hybridize every two chromosomes by probability Pc in New_Pop . Get the hybrid populations $cross\_pop$ including chromosome numbers M .

5. Mutation operation：Mutate some chromosomes by the probability Pm in $cross\_pop$, and form the new population mut_pop after mutation.

6. If the error is less than ε or the reproduction exceed 600 generation, import the optimal solution with the maximum fitness, and turn to step 7. Otherwise, turn to step 2.

7. Assign the optimal test cost to tc, and the optimal misclassification to mc. Apply tc and tc to data set D, and structure the classifier M by the decision tree algorithm.

8. The algorithm ends.

5. Simulation Experiment

At present, there aren't the standard uncertain data sets, so we reform the certain data sets to the uncertain data in the experiment, and the certain sets come from the standard UCI. The whole value of attributes has the rational distribution, for example, the normal distribution. For obtaining the uncertain data sets, the attributes value is shown by the probability of the whole domain. We assume that all attributes cost are unknown, and must test for getting values, at the same time, the corresponding test cost is paid.

Adopt Weka 3.6.4 as the experiment platform, and the program is optimized by Linux 3.1.2 and GCC4.1.2. The initial m=200, crossover probability Pc=0.6, mutation probability Pm=0.2, the maximum iteration MAX=600. Choose Automobile, Ecolin, Sponge and Yeast in UCI as test data, and each data set is chosen 10% as uncertain data. The decision tree classifiers of DTU [7], NS-PDT [20], FDTU [10] are often used for uncertain data. The experimental results of this paper proposes the classification approach (GDCDTU) and other algorithms in classification accuracy are as follows.

Table 1. Classification Precision for Uncertain Data

Classification algorithm	Classification precision			
	Automobile	Ecolin	Sponge	Yeast
DTU	68%	72%	74%	79%
NS-PDT	73%	78%	81%	82%
FDTU	79%	86%	78%	86%
GDCDTU	87%	92%	89%	90%

From Table 1, GDCDTU achieves higher accuracy than the other algorithms. In order to verify the effectiveness of dynamic cost, which is replaced by stationary cost, and get stationary cost-sensitive classification algorithm based on the genetic algorithm for uncertain data (CSCDTU). Choose the equal fixed value for test and misclassification in DTU and CSCDTU, and GDCDTU, and use the interval number near the fixed. Choose Automobile (AU), ISOLET(IS), Liver Disorders (LD), Nursery(NU), Wine(WN) and Yeast(YE) in UCI as the uncertain data, and the missing data ratio is 10%. Choose Area Under the ROC Curve (AUC) to measure the classifiers, and the area of under ROC curve can judge the performance. Experiment 50 times in each data set, and the results in Figure 1 are as follows.

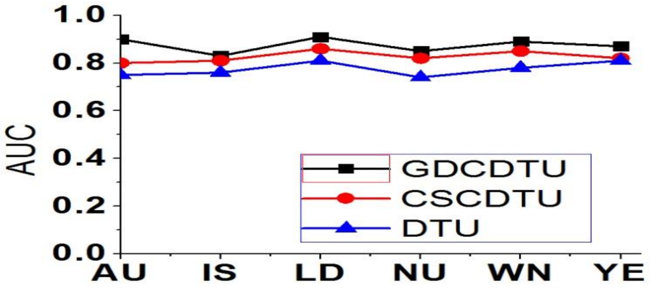

Figure 1. AUC of GDCDTU, CSCDTU and DTU

AUC is larger, and the classification accuracy is higher, so it can be seen that GDCDTU has higher classification accuracy than the other algorithms from figure 1. GDCDTU adopts dynamic cost combining the genetic algorithm, and its performance is superior to CSCDTU and DTU. Use the parameter, and select the same data sets with above experiments. The total cost of DTU, CSCDTU and GDCDTU is as follows in Figure 2.

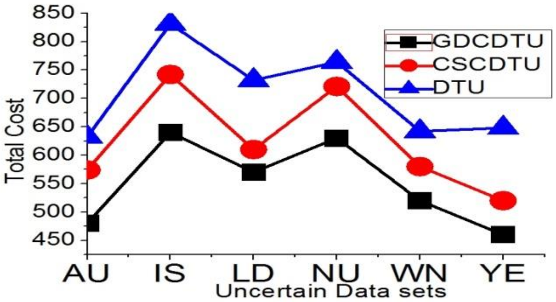

Figure 2. Total cost of GDCDTU,CSCDTU and DTU

The total cost of GDCDTU is lower than CSCDTU and DTU form the Figure 2. GDCDTU can effectively reduce the total costs by building the dynamic cost, save the total expenses.

6. Conclusion

Data uncertainty and classification cost are prevalent in many real-world applications, so this paper proposes the cost-sensitive classification approach for uncertain data, which overcomes the limitations of the stationary cost, and incorporates the expert's experience and knowledge, searches automatically suitable misclassification costs for every sub datasets. We dispose the continuous and discrete attribute for uncertain data by the probabilistic cardinality. The dynamic cost-sensitive decision tree based on the genetic algorithm for uncertain data is structured, and the optimum cost is produced by the hybridization, the mutation and selection. Experimental result shows, the decision tree we proposed has higher classification accuracy and performance than the other comparative similar classification algorithms, which can save especially the cost, and is very suitable for extremely imbalanced datasets with a high stability.

Acknowledgements

This work has been partially supported by grants from the research fund of heze university (No. XY12KJ03). This work was supported by the science and technology projects of the Shandong province universities (No. J12LN55, No. J13LN53).

References

[1] B. Qin, Y. N. Xia, S. Wang and X. Y Du, "A novel Bayesian classification for uncertain data", "Knowledge-Based Systems", vol. 24, Issue 8, **(2011)**, pp. 1151-1158.

[2] Z. G. Liu, Q. Pan, J. Dezert and G. Mercier, "Credal classification rule for uncertain data based on belief functions", Pattern Recognition, vol. 47, Issue 7, **(2014)**, pp. 2532-2541.

[3] M. Bounhas, M. G. Hamed, H. Prade, M. Serrurier and K. Mellouli, “Naive possibilistic classifiers for imprecise or uncertain numerical data”, Fuzzy Sets and Systems, vol. 239, **(2014)**, pp. 137-156.

[4] Y. J. Sun, Y. Yuan and G. R. Wang, “Extreme learning machine for classification over uncertain data”, “Neurocomputing”, vol. 128, **(2014)**, pp. 500-506.

[5] Y. H. Liu and C. S. Wang, “Constrained frequent pattern mining on univariate uncertain data”, Journal of Systems and Software, vol. 86, iss. 3, **(2013)**, pp. 759-778.

[6] Z. G. Liu, Q. Pan and J. Dezert, “Classification of uncertain and imprecise data based on evidence theory”, Neurocomputing, vol. 133, **(2014)**, pp. 459-470.

[7] B. Qin, Y. N. Xia and F. Li, “DTU: A Decision Tree for Uncertain Data. In Proc. of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining”, **(2009)**, pp.4-15.

[8] S. Tsang, B. Kao, K. Y. Yip and W. S. Ho *et.al.*, “Decision Trees for Uncertain Data”, In Proc. Of the 25th International Conference on Data Engineering (ICDE’09), **(2009)**, pp. 441-444.

[9] B. Qin, Y. Xia, S. Prabhakar and Y. Tu, “2009b A Rule-Based Classification Algorithm for Uncertain Data”, In Proc. of the 1st IEEE workshop on Management and Mining of Uncertain Data (MOUND’09), in conjunction with ICDE, **(2009),** pp. 1633-1640.

[10] C. Q. Liang, Y. Zhang, P. Shi and Z. G. Hu, “Learning very fast decision tree from uncertain data streams with positive and unlabeled samples”, Information Sciences, vol. 213, **(2012),** pp. 50-67.

[11] A. Ibáñez, C. Bielza and P. Larrañaga, “Cost-sensitive selective naive Bayes classifiers for predicting the increase of the h-index for scientific journals”, Neurocomputing, vol. 135, **(2014)** , pp. 42-52.

[12] X. B. Yang, Y. S. Qi, X. N. Song and J. Y. Yang, “Test cost sensitive multigranulation rough set: Model and minimal cost selection”, Information Sciences, vol. 250, **(2013)**, pp. 184-199.

[13] P. C. Pendharkar, “A maximum-margin genetic algorithm for misclassification cost minimizing feature selection problem”, Expert Systems with Applications, vol. 40, iss. 10, **(2013)**, pp. 3918-3925.

[14] C. X. Ling, Q. Yang, J. N. Wang and S. C. Zhang, “Decision Trees with Minimal Costs”, In the proceedings of the twenty-first international conference on Machine learning (ICML), **(2004)**, Banff, Canada.

[15] Y. Weiss, Y. Elovici and L. Rokach, “The CASH algorithm-cost-sensitive using histograms”, Information Sciences, vol. 222, **(2013)** February 10, pp. 247-268.

[16] Y. Sahin, S. Bulkan and E. Duman, “A cost-sensitive decision tree approach for fraud detection”, Expert Systems with Applications, vol. 40, iss. 15, **(2013)** November 1, pp. 5916-5923.

[17] B. Krawczyk, M. Woźniak and G. Schaefer, “Cost-sensitive decision tree ensembles for effective imbalanced classification”, Applied Soft Computing, Part C, vol. 14, **(2014)** January**,** pp. 554-562.

[18] S. C. Zhang, “Decision tree classifiers sensitive to heterogeneous costs”, Journal of Systems and Software, vol. 85, iss. 4, **(2012)** April, pp. 771-779.

[19] S. Sheng, C. X . Ling and Q. Yang, “Simple Test Strategies for Cost-Sensitive Decision Trees”, Lecture Notes in Artificial Intelligence, vol. 3720, (**2005)**, pp. 365-376.

[20] I. Jenhanni, N. B. Amor and Z. Eluouedi, “Decision trees as possibilistic classifiers, International Journal of Approximate Reasoning, vol. 48, iss. 3, pp. 784-807.

Author

Yuwen Huang, was born in 1978 at Shanxian, and received the Master of Engineering in Computer Science from the “Guangxi Normal university” in 2009. She is now a lecturer at the Department of Computer and Information Engineering, Heze University. His research interests include the data-mining, intelligence Calculation.