International Journal of Database Theory and Application
Vol.8, No.1 (2015), pp.271-280
http://dx.doi.org/10.14257/ijdta.2015.8.1.26

Dynamic Cost-sensitive Naive Bayes Classification for Uncertain Data

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 uncertain data as an important aspect of data mining, has received considerable attention, due to its importance in many applications, but little study has been paid to the cost-sensitive classification on uncertain data, so this paper proposes the dynamic cost-sensitive Naive Bayes classification for mining uncertain data (DCSUNB). Firstly, we apply the probability density to dispose uncertain discrete and continuous attributes, and give the cost-sensitive Naive Bayes classifier. Secondly, we propose the construction process of dynamic cost, and give the evaluation method for finding the optimal cost and the cost-sensitive classification with sequential test strategy. At last, the dynamic cost-sensitive Naive Bayes algorithm for uncertain data is structured, which searches the misclassification and test cost spaces to find the optimal cost. By comparing to the other cost-sensitive classification algorithms for uncertain data, the experiments on UCI Datasets show that DCSUNB can improve the classification performance, and reduce effectively the total cost.

Keywords: *Dynamic cost-sensitive; Misclassification cost; Naive Bayes; Test cost; Uncertain data*

1. Introduction

In recent years, there are more and more uncertain data in the real time monitoring system, the communication network, the remote sensor and the sensor network, etc. The traditional data mining techniques are designed for the certain data, and they don't manage effectively the uncertain data, so the researches on uncertain data mining are significance. Lei proposed a supervised UK-means for uncertain data to classify the same class into subclasses [1]. Qin gave the Novel Bayesian classification algorithm for uncertain data by probabilistic and statistical theory [2]. Sun proposed classification algorithms for uncertain data by the improved extreme learning machine algorithms [3]. Xu gave a heuristic clustering approach for uncertain data by the effective approximate UK-means [4].

The current classification algorithms on certain data consider mainly the classification precision, and their aims are to minimize the classification errors. Little study has been paid to the cost-sensitive classification on uncertain data, but the mining tasks in many applications involve different costs, so the existing data mining approaches can't meet the requirements. The cost-sensitive learning considers different cost, and its aim is to produce the optimal decision by the minimum cost. There are some works on cost-sensitive mining for uncertain data. Liu proposed a decision tree for uncertain data by extending the traditional cost-sensitive decision tree [5]. Zhang gave a new cost-sensitive Naive Bayes classification for uncertain data [6]. Huang proposed classification algorithm of the dynamic cost-sensitive decision tree for uncertain data based on the genetic algorithm [7]. Liu proposed a cost-sensitive clustering algorithm for uncertain data [8]. The current classification algorithms for uncertain data consider mainly the misclassification and test costs as the static costs. However, in real world, the misclassification and test costs change dynamically, and the current mining algorithms for uncertain data rarely think of the dynamic cost. This paper proposes the dynamic cost-sensitive Naive Bayes classification for mining uncertain data, which searches the misclassification and test cost spaces to find the optimal cost.

2. Related Work

2.1. Probability Density of Uncertain Attribute

2.1.1. Probability Density of Uncertain Discrete Attribute: Training set $\{X_1, X_2, ..., X_n\}$, $X_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$, x_{ij} is uncertain discrete attribute, $x_{ij} = \left[x_{ij}^1, x_{ij}^2, ..., x_{ij}^u\right]$. p_{ij} is the probability of each discrete attribute, $p_{ij} = \left\{p_{ij}^1, p_{ij}^2, ..., p_{ij}^u\right\}$, $\sum_{x=1}^{u} p_{ij}^x = 1$.

According to the maximum likelihood method and Laplace correction, $P\left(x_{ij}^a \mid c_i\right)$ is as follows.

$$P\left(x_{ij}^a \mid c_i\right) = \frac{1 + N\left(x_{ij}^a, c_i\right)}{\left|V\left(x_{ij}\right)\right| + N\left(C_i\right)}$$

$N\left(x_{ij}^a, c_i\right)$ is the numbers of class c_i with attribute x_{ij}^a , and $N\left(C_i\right)$ is numbers of class c_i . $\left|V\left(x_{ij}\right)\right|$ is value numbers of attribute X_{ij} .

The probability density of $P\left(x_{ij} \mid c_i\right)$ is as follows.

$$P\left(x_{ij} \mid c_i\right) = \sum_{a=1}^{u} P\left(x_{ij}^a \mid c_i\right) p\left(x_{ij}^a\right) .$$

2.1.2. Probability Density of Uncertain Continuous Attribute: Uncertain continuous attributes distribute by the interval number, so this paper uses the segmentation strategy to divide the interval number into many blocks by the histogram section, where the interval number is located into each histogram by a certain probability. A histogram set $\{I_1, I_2, ..., I_L\}$, $I_i = [a_i, a_{i+1}]$, and the probability of the interval number $u = \left[u^-, u^+\right]$ in each histogram is as follows.

$$p_i = \begin{cases} \dfrac{u^+ - a_i}{u^+ - u^-} & u^- \prec a_i \prec u^+ \prec a_{i+1} \\ 1 & a_i \prec u^- \prec u^+ \prec a_{i+1} \\ \dfrac{a_{i+1} - u^-}{u^+ - u^-} & a_i \prec u^- \prec a_{i+1} \prec u^+ \\ 0 & others \end{cases}$$

Input: Training set $\{x_1, x_2, ..., x_n\}$, $x_i \in \left[x_i^-, x_i^+\right]$, the sample $u_p = \left[u_p^-, u_p^+\right]$, the segment number L .

```
Output： Probability density of $u_p = \left[u_p^-, u_p^+\right]$ .
 Step 1: $x_{\min}^- = x_1^-$ ; $x_{\max}^+ = x_1^+$ ;
for(i=2;i<=n; i++)
{
 if( $x_{\min}^- \succ x_i^-$ ) $x_{\min}^- = x_i^-$ ;
if( $x_{\max}^+ \prec x_i^+$ ) $x_{\max}^+ = x_i^+$ ;
}
```

Step 2: $a_1 = x_{\min}^-$, $a_{L+1} = x_{\max}^+$, $h = {(a_{L+1} - a_1)}/{L}$. $[a_1, a_{L+1}]$ is divided as the subintervals with the width $h = {(a_{L+1} - a_1)}/{L}$, $[a_1, a_{L+1}] = \{[a_1, a_2], [a_2, a_3], \ldots, [a_L, a_{L+1}]\}$. $a_i = a_1 + (i-1) \times h, i = 1, 2, \ldots, L$.

Step 3: Calculate the frequency number P_i of $x_i \in \{[a_1, a_2], [a_2, a_3], \ldots, [a_L, a_{L+1}]\}$, $j = 1, 2, \ldots, \mathrm{L}$. $P_i = \sum_{j=1}^{L} p_j$, p_j is the probability that the interval number x_i is located into the histogram $\mathrm{I}_j = \left[a_j, a_{j+1}\right]$.

Step 4: Calculate the frequency of I_i .

$$f_i = \frac{P_i}{n}(i = 1, 2, \ldots, L)$$

Step 5: Calculate the probability density function: $f(x) = \begin{cases} \dfrac{f_i}{h} & x \in I_i \,(i = 1, 2, \ldots, L) \\ 0 & others \end{cases}$

Step 6: The probability density of $u_p = \left[u_p^-, u_p^+\right]$ is :

$$P\left(u_p = \left[u_p^-, u_p^+\right]\right) = \sum_{i=1}^{L} P_i \times \frac{f_i}{h} .$$

2.2. Cost-sensitive Naive Bayes Classifier

There are many kinds of costs, and Tunney proposed nine costs in real life. In particular, the misclassification and test costs are often considered. Misclassification costs are the expenditure when one class is mistaken as the other class, and test costs are the expenditure when the attributes values are acquired by test methods. When misclassification cost is a fixed value, and it is rational to decide if the further test is done.

$mis\_Cost_{M \times M}$ is the misclassification cost matrix, $mis\_Cost_{M \times M} = \begin{bmatrix} c_{11}, c_{12}, \ldots, c_{1\mathrm{M}} \\ c_{21}, c_{22}, \ldots, c_{2\mathrm{M}} \\ \cdots \; \cdots \; \cdots \; \cdots \\ c_{\mathrm{M}1}, c_{\mathrm{M}2}, \ldots, c_{\mathrm{MM}} \end{bmatrix}$.

C_{ij} is the cost of classifying a sample with class C_i as class C_j . If $i = j$, $C_{ij} = 0$. $test\_\cos t_N$ is a vector, $test\_\cos t_N = [\cos t_1, \cos t_2, \ldots, \cos t_{\mathrm{n}}]$, $\cos t_i$ is the test cost for getting the attribute value.

Data sets $S = (U, A, V, f)$, the attributes $A = \{a_1, a_2, a_3, ...a_m, C\}$, the classification $C = \{C_1, C_2, C_3.......C_k\}$. $X_i \in U$, $X_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$, $x_{ij} \in A_{ij}$. The posterior probability of sample X is as follows.

$$P(c_i \mid X) = \frac{P(X \mid c_i) P(c_i)}{P(X)} = \frac{\prod_{i=1}^{m} P(a_i \mid c_i) P(c_i)}{\sum_{i=1}^{k} P(X \mid c_i) P(C_i)}$$

$P(X) = \sum_{i=1}^{k} P(X \mid C_i) P(C_i)$ is the constant, so the formula that sample X is classified as class C_j is as follows by Naive Bayes classification .

$$\begin{aligned} C_j &= \arg\max_{C_j} \left\{ P\left(C_j \mid X\right) \right\} = \arg\max_{C_j} \left\{ P(X \mid c_i) P(c_i) \right\} \\ &= \arg\max_{C_j} \left\{ \prod_{i=1}^{m} P(a_i \mid c_i) P(c_i) \right\} \end{aligned} .$$

$F\left(C_j, C_i\right)$ is the risk that class C_j is classified as belonging to class C_i , and $F\left(C_j, C_i\right) = mis\_cost + test\_cost$. The structure of risk function is as follows.

$$R(C_i \mid X) = \min \arg_{C_i \in C} \sum_{j=1}^{k} F\left(C_j, C_i\right) P\left(C_j \mid X\right) .$$

The sample X is classified as class C_i with the minimum risk.

3. The Proposed Scheme

3.1. Dynamic Test Cost

The current researches on test costs focus on the static costs only for experiments and prospective study, and it don’t suit the imbalanced uncertain data. When facing great imbalance classification, the performance of classifiers with static cost is greatly degraded, and the dynamic cost-sensitive learning can solve the problem. We structure test and misclassification cost space that are collected from different application domain experts, and each cost from space is representative, which can truly reflect the distribution characteristics of data sets. The optimization algorithms can search the optimal appropriate cost in different application, so the dynamic costs don’t depend on the application area. The construction process for dynamitic cost is as follows.

Step 1: Give the problem needed to get costs. For example, a patient needs a chest X-ray test.

Step 2: Choose N methods for getting differential costs. For example, choose N hospitals for doing the chest X-ray test.

Step3: In combination with the experience and knowledge background of the application experts, collect the N costs of N methods. For example, the levels of medical professionals for the chest X-ray test in difference hospitals are different, so the test costs are different.

Step 4: N costs are sorted in an ascending order, and gets the minimum $cost_{\min}$ and maximum $cost_{\max}$ of N costs. $[cost_{\min}, cost_{\max}]$ is the cost space of the problem needed to get costs. For example, the minimum cost of chest X-ray test is \$20, and the maximum is \$50, so its cost space is the interval [20, 50].

3.2. Evaluation Method of Optimal Cost

There are many evaluation methods for classifiers, and the precision rate (PR) and response rate (RE) are usually used.

$$PR = \frac{TP}{TP + FP} \ .$$

$$RE = \frac{TP}{TP + FN}$$

The response rate is the ratio of the prediction correct and the sum in all positive samples, and the precision rate is the ratio of the prediction correct positive samples and the sum of the prediction positive samples. A classifier with poor performance may be very high response rate, but the precision rate is very low. Facing with the imbalanced data sets, the traditional classifiers classify usually a sample as the majority class, but in fact, the minority class is important. Therefore, we must increase the accuracy of minority class, and assign the high misclassification cost for minority class. At the same time, the accuracy of majority class should also be considered, so this paper compromises the response and precision ratio. The cost function is described as follows by the extended geometric average formula $G-mean$.

$$f(\cos t) = G\left(PR(cost), RE(\cos t)\right) = \sqrt{PR(\cos t) \times RE(\cos t)}$$

The cost in the above formula is a point of cost space, and G is $G-mean$ function. When the point is applied in data sets, PR and RE are respectively the precision and response rate of the structured classifier after cross validation.

The function $f\left(\cos t_{opt}\right)$ is the optimal cost, and it is defined as follows when the geometric average of $PR(\cos t)$ and $RE(\cos t)$ is the maximum.

$$f\left(\cos t_{opt}\right) = \arg\max_{\cos t \in C} f\left(\cos t_{opt}\right) = \arg\max_{\cos t \in C}\left(\sqrt{PR(\cos t) \times RE(\cos t)}\right)$$

It is difficult for getting a real cost of different data sets in many applications, and the dynamic cost can approach the real cost, so it is feasible to get the dynamic cost.

3.3. Cost-sensitive Classification with Sequential Test Strategy

In the process of classification for uncertain data, if the attributes are missing, the selection for further attribute tests depends on not only the reduction of misclassifications cost, but also the cost of testing the missing attributes.

Training set $\{X_1, X_2, ..., X_n\}$, $X_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$, $x_{ij} \in \left[x_{ij}^-, x_{ij}^+\right]$. $X_i = X_i^{known} + X_i^{unknown}$, X_i^{know} is a set of known attributes, and $X_i^{unknown}$ is a set of unknown attributes.

The structure of risk function based on the misclassification cost is:

$$R(C_i \mid X_i) = \min\arg_{C_i \in C} \sum_{j=1}^{k} F\left(C_j, C_i\right) P\left(C_j \mid X_i^{known}\right)$$

$P\left(C_j \mid X_i^{known}\right) = \dfrac{P\left(X_i^{known} \mid C_j\right)}{P\left(X_i^{known}\right)}$ is the posterior probability, and $F\left(C_j, C_i\right)$ is the cost that classifies class C_j as class C_i . If the class C_j is predicted by the known attributes, $F\left(C_j, C_i\right)$ is made completely by the misclassification costs matrix $mis\_Cost_{M \times M}$, and test costs are zero, where no any tests are performed. However, if a test can reduce the misclassification cost, and the test should be done. In order to decide whether an attribute value is acquired by test, the formula of cost change is:

$$\cos t\_change\left(x_{ij}^{unknown}\right) = \cos t\_chang_{misclass}\left(X_i^{known}, x_{ij}^{unknown}\right) - C_{test}\left(x_{ij}^{unknown}\right)$$

$x_{ij}^{unknown} \in X_i^{unknown}$, $C_{test}\left(x_{ij}^{unknown}\right)$ is the test cost for $x_{ij}^{unknown}$, and $\cos t\_chang_{misclass}\left(X_i^{known}, x_{ij}^{unknown}\right)$ is the reduction of misclassification when $x_{ij}^{unknown}$ is tested.

$$\cos t\_chang_{misclass}\left(X_i^{known}, x_{ij}^{unknown}\right) = \cos t_{misclass}\left(X_i^{known}\right) - \cos t_{misclass}\left(X_i^{known} \cup x_{ij}^{unknown}\right)$$

$\cos t_{misclass}\left(X_i^{known}\right) = \min_j R\left(C_i \mid X_i\right)$ is misclassification cost that is classified only by the known attributes X_i^{known} . When $x_{ij}^{unknown}$ is tested, $\cos t_{misclass}\left(X_i^{known} \cup x_{ij}^{unknown}\right)$ is misclassification cost that is classified by the known attributes X_i^{known} and $x_{ij}^{unknown}$.

$\cos t_{misclass}\left(X_i^{known} \cup x_{ij}^{unknown}\right)$ is:

$$\begin{aligned}
&\cos t_{misclass}\left(X_i^{known} \cup x_{ij}^{unknown}\right) \\
&= E_{x_{ij}^{unknown}}\left[\min_k R\left(C_k \mid X_i^{known} \cup x_{ij}^{unknown}\right)\right] \\
&= \sum_{j=1}^{\left\|x_i^{unknown}\right\|} P\left(x_{ij} \mid X_i^{known}\right) \times \min_k R\left(C_k \mid X_i^{known} \cup x_{ij}\right)
\end{aligned}$$

Where $x_{ij} \in x_i^{unknown}$ is chosen as test attribute, $P\left(x_{ij} \mid X_i^{known}\right)$ is the conditional probability. $\min_k R\left(C_k \mid X_i^{known} \cup x_{ij}\right)$ is minimum misclassification cost that is classified by the attributes $X_i^{known} \cup x_{ij}$.

An unknown attribute $x_{ij}^{unknown}$ is worth testing if the reduction of misclassification cost is more than test cost. In other word, the change of $\cos t\_change\left(x_{ij}^{unknown}\right)$ is more than zero, and it is beneficial for the further test. If the attribute $x_{ij}^{unknown}$ is tested, and the known attributes X_i^{known} are expanded to $X_i^{known} \cup x_{ij}^{unknown}$. At the same time, the unknown $X_i^{unknown}$ is reduced to $X_i^{unknown} \Big/ x_{ij}^{unknown}$. The total cost $F\left(C_j, C_i\right)$ comprises misclassification and test cost, $F\left(C_j, C_i\right) = mis\_\mathrm{cost} + \mathrm{test\_cost}$. The steps of $X_i = \left\{x_{i1}, x_{i2}, \ldots, x_{im}\right\}$ are classified as follows.

Input: The sample $X_i = \left\{x_{i1}, x_{i2}, \ldots, x_{im}\right\}$, $x_{ij} \in \left[x_{ij}^-, x_{ij}^+\right]$.

Output: The classification of the sample X_i .

Step 1: Divide $X_i = \left\{x_{i1}, x_{i2}, \ldots, x_{im}\right\}$ into the known attributes set X_i^{known} and unknown attributes set $X_i^{unknown}$, $X_i = X_i^{known} + X_i^{unknown}$, $test\_\mathrm{cost} = 0$.

Step 2: While ($X_i^{unknown}$ is not empty)

{

for each $x_{ij}^{unknown} \in X_i^{unknown}$ do

Calculate the cost change reduction $\cos t\_change\left(x_{ij}^{unknown}\right)$;

```
End for
If $not\ \exists\ \cos t\_change\left(x_{ij}^{unknown}\right) \succ 0$ , then break;
$x_{ij} = \max_{j}\left(\cos t\_change\left(x_{ij}^{unknown}\right)\right)$.
Test the attribute $x_{ij}$ to get its value.
$x_i^{unknown} = x_i^{unknown} - x_{ij}$;
$X_i^{known} = X_i^{known} + x_{ij}$;
$test\_\mathrm{cost} = test\_\mathrm{cost} + test\_\mathrm{cost}\left(x_{ij}\right)$;
}//end while
```

Step 3: Calculate the risk function $R\left(C_i \mid X_i\right) = \min \arg_{C_i \in C} \sum_{j=1}^{k} F\left(C_j, C_i\right) P\left(C_j \mid X_i^{known}\right)$

Step 4: The sample class $X_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$ is the minimum of $\min_{C_i \in C}\left(R\left(C_i \mid X_i\right)\right)$

All missing attributes aren't been tested if each cost change $\cos t\_change\left(x_{ij}^{unknown}\right)$ is less zero. In other word, if the test costs are more than the change of misclassification cost, the further test isn't necessary. If all attributes don't miss, the test costs are zero.

3.4. Dynamic Cost-Sensitive Naive Bayes Algorithm for Mining Uncertain Data

When facing the different data sets, it is key to structure the algorithm for searching the optimal cost. Dynamic cost-sensitive Naive Bayes algorithm for mining uncertain data is follows.

Input: Training set D , Dynamic misclassification cost space $mis\_Cost_{M \times M}$, $mis\_\cos t_{ij} \in mis\_Cost_{M \times M}$, $mis\_\cos t_{ij} = \left[mis\_\cos t_{ij}^{-}, mis\_\cos t_{ij}^{+}\right]$. Dynamic test cost space $test\_\mathrm{cost}_N$, $test_i \in test\_\mathrm{cost}_N$, $test_i = \left[test_i^{-}, test_i^{+}\right]$. The current misclassification cost $mis\_\mathrm{cost}_{M \times M}^{cur}$, the current test cost $test\_\mathrm{cost}_N^{cur}$, the optimal misclassification cost $mis\_\mathrm{cost}_{M \times M}^{opt}$, the optimal test cost space $test\_\mathrm{cost}_N^{opt}$, the increment □$mis\_\mathrm{cost}_{M \times M}^{incr}$ for misclassification cost, □$test\_\mathrm{cost}_N^{incr}$ for test cost.

Output: Classification of uncertain data set D .

Step 1: Initialize the current and optimal matrixes. $mis\_\cos t_{ij}^{-} \in mis\_Cost_{M \times M}^{-}$, $mis\_\mathrm{cost}_{M \times M}^{cur} = mis\_\mathrm{cost}_{M \times M}^{opt} = mis\_\cos t_{M \times M}^{-}$. $test_i^{-} \in test\_\mathrm{cost}_N^{-}$, $test\_\mathrm{cost}_N^{cur} = test\_\mathrm{cost}_N^{opt} = test\_\mathrm{cost}_N^{-}$.

Step 2: Apply $mis\_\mathrm{cost}_{M \times M}^{cur}$ and $test\_\mathrm{cost}_N^{cur}$ to uncertain data set D , and use the cost-sensitive Bayes classifier to structure the basic classifier M_{opt} .

Step 3: Use 10-fold cross-validation for the basic classifier M_{opt} in data set D . Calculate the response and precision rate of minority class $f\left(\cos t_{cur}\right) == \sqrt{PR\left(\cos t\right) \times RE\left(\cos t\right)}$, $f_{\max} = f\left(\cos t_{cur}\right)$. Choose the next cost point, $mis\_\mathrm{cost}_{M \times M}^{cur} = mis\_\mathrm{cost}_{M \times M}^{cur} +$□$mis\_\mathrm{cost}_{M \times M}^{incr}$, $test\_\mathrm{cost}_N^{cur} = test\_\mathrm{cost}_N^{cur} +$□$test\_\mathrm{cost}_N^{incr}$.

Step 5: If $mis\_cost_{M\times M}^{cur}$ and $test\_cost_{N}^{cur}$ don't belong to the cost space, turn to step 8. Otherwise, apply $mis\_cost_{M\times M}^{cur}$ and $test\_cost_{N}^{cur}$ to uncertain data set D , and use the cost-sensitive Bayes classifier to structure the basic classifier M_{cur} .

Step 6: Use 10-fold cross-validation to calculate PR and RE in classifier M_{cur} , $f(cost_{cur}) = \sqrt{PR(cost)\times RE(cost)}$.

Step 7: If $f(cost_{cur}) > f(cost_{\max})$, $f(cost_{\max}) = f(cost_{cur})$, $test\_cost_{N}^{opt} = test\_cost_{N}^{cur}$. Choose the next cost point by the searching algorithm, $mis\_cost_{M\times M}^{cur} = mis\_cost_{M\times M}^{cur} + \Delta mis\_cost_{M\times M}^{incr}$, $test\_cost_{N}^{cur} = test\_cost_{N}^{cur} + \Delta test\_cost_{N}^{incr}$, $M_{opt} = M_{cur}$,

$mis\_cost_{M\times M}^{cur} = mis\_cost_{M\times M}^{cur} + \Delta mis\_cost_{M\times M}^{incr}$, $test\_cost_{N}^{cur} = test\_cost_{N}^{cur} + \Delta test\_cost_{N}^{incr}$, turn to step 5. If $f(cost_{cur}) \prec f(cost_{\max})$, $mis\_cost_{M\times M}^{cur} = mis\_cost_{M\times M}^{cur} + \Delta mis\_cost_{M\times M}^{incr}$, $test\_cost_{N}^{cur} = test\_cost_{N}^{cur} + \Delta test\_cost_{N}^{incr}$, turn to step 5.

Step 8: For each sample $X_i \in D$, use $mis\_cost_{M\times M}^{opt}$, $test\_cost_{N}^{opt}$ and M_{opt} to calculate the risk function $R(C_i \mid X_i) = \min_{C_i\in C} \arg \sum_{j=1}^{k} F(C_j, C_i) P(C_j \mid X_i^{known})$. The predicted class of sample X_i is the minimum of $\min_{C_i\in C}(R(C_i \mid X_i))$.

4. Simulation Experiment

4.1. Data Preprocessing

At present, there aren't the open standard uncertain data sets for the experiments, so this paper introduces the uncertain information from the uncertain standard UCI dataset. The continuous attribute x_{ij} is added the uncertain information as the interval number $\left[\bar{x}_{ij} - r_j, \bar{x}_{ij} + r_j\right]$ in definite proportions, $r_j = \frac{\max_i\left(\bar{x}_{ij}\right) - \min_i\left(\bar{x}_{ij}\right)}{10}$. For discrete attribute x_{ij} , $x_{ij} = \left[x_{ij}^1, x_{ij}^2, ..., x_{ij}^u\right]$, p_{ij}^u is the probability that x_{ij} has value x_{ij}^u , $p_{ij} = \left\{p_{ij}^1, p_{ij}^2, ..., p_{ij}^u\right\}$. $\sum_{x=1}^{u} p_{ij}^x = 1$. Test cost of attributes are set as the interval number $[1,100]$. FP is the cost that the positive sample is misclassified as the negative sample, and FP is set as the interval number $[1,40]$. TP is the cost that the negative sample is misclassified as the positive sample, and TP is set as the interval number $[60,100]$. Choose Windows7 with Intel E5800 (3.2GHz) + 4.0GB main memory and Weka3.6.4 as development platform.

4.2. Result of Experiments

Dynamic cost-sensitive naive Bayes classification for Mining Uncertain Data is abbreviated to DCSUNB, and CSDTU [5], CS-UNB [6] are other cost-sensitive classification algorithms for uncertain data. CS-UNB and CSDTU use the fixed test and misclassification cost, and DCSUNB choose the interval value near the fixed number in experiments. Choose Ecolin, Vote, Vowel, Wine and Segment in UCI dataset as test data, and the ratio of uncertain data is set as 10%. Use Area under the ROC Curve (AUC) as the performance criteria of the cost-sensitive classifiers, the results are as follows after 50 times experiments in each data set.

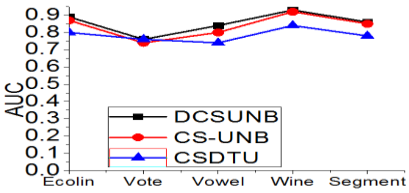

Figure 1. AUC of DCSUNB, CS-UNB and CSDTU

As we can see from the Figure 1, AUC of DCSUNB is larger than CS-UNB and CSDTU, so DCSUNB has better classification accuracy than the others. DCSUNB adopts dynamic cost, and its performance is superior to CS-UNB and CSDTU.

In order to verify the average total cost of DCSUNB, CS-UNB and CSDTU, the ratio of uncertain data is set as 10% in Ecolin, Vote, Vowel, Wine and Segment. Average total costs of DCSUNB, CS-UNB and CSDTU are as follows.

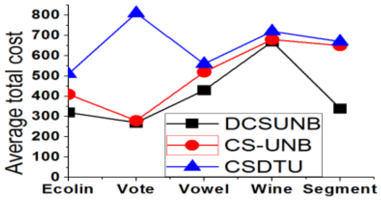

Figure 2. Average Cost of DCSUNB, CS-UNB and CSDTU

From Figure 2, it can be seen that average total cost of DCSUNB based on dynamic cost-sensitive Naive Bayes classification is smaller than CS-UNB and CSDTU, so DCSUNB has better performance than the others.

In order to verify the influence of the uncertain level, the ratios of uncertain levels are set form 0 to 50%. The average total costs of DCSUNB, CS-UNB and CSDTU in Vote and Segment data sets are as follows.

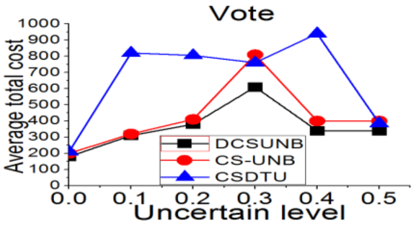

Figure 3. Average Total Cost with Varying Uncertain Level in Vote

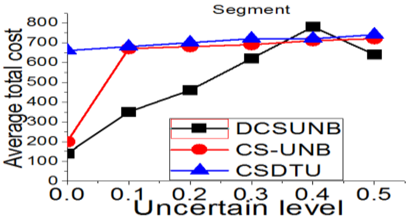

Figure 4. Average Total Cost with Varying Uncertain Level in Segment

From Figure 3 and Figure 4, we can see that the average total cost of DCSUNB is smaller than CS-UNB and CSDTU, so this strongly suggests that DCSUNB has a better performance for uncertain data. The total costs of DCSUNB and CS-UNB drop after the uncertain level with 40% from figure 3, and we can conclude that the added test cost can decrease the misclassification cost, so the total cost would drop.

5. Conclusion

There are more and more uncertain data in many real-world applications including communication network, the remote sensor and the sensor network, and so forth, and this paper proposes the dynamic cost-sensitive Naive Bayes classification for mining uncertain data, which overcomes the limitations of the stationary cost. We dispose the continuous and discrete attribute for uncertain data by the probability density, and give the cost-sensitive Naive Bayes classifier. In this paper, we give the evaluation method for dynamitic cost, and propose the test strategy for further attribute selection. The dynamic cost-sensitive Naive Bayes Algorithm for uncertain data can find the optimal cost for the real-world applications. Our experimental results demonstrate that DCSUNB has higher performance than the other cost-sensitive classification algorithms for uncertain data, which can save especially the cost.

Acknowledgements

This work was supported by the science and technology projects of the Shandong province universities (No. J13LN53, No. J14LN65). This work has been partially supported by grants from the research fund of heze university (No. XY14KJ06).

References

[1] L. Xu and E. Hung. Improving classification accuracy on uncertain data by considering multiple subclasses. Neurocomputing 2014; 145(5): 98-107.
[2] B. Qin , Y. Xia , S. Wang and X. Y. Du . A novel Bayesian classification for uncertain data. Knowledge-Based Systems 2011, 24(8): 1151-1158.
[3] Y. J. Sun , Y. Yuan and G. Wang . Extreme learning machine for classification over uncertain data. Neurocomputing 2014, 128(1): 500-506.
[4] L. Xu , Q. H. Hu , E. Huang and C. C. Szeto. A heuristic approach to effective and efficient clustering on uncertain objects. Knowledge-Based Systems 2014, 66: 112-125.
[5] M. J.Liu , Y. Zhang , X. Zhang and Y. Wang . Cost-Sensitive Decision Tree for Uncertain Data. Advanced Data Mining and Applications 2011, Lecture Notes in Computer Science, 7120: 243-255.
[6] X. Zhang , M. Li, Y. Zhang and J.F. Ning . Cost-sensitive Naïve Bayes Classification of Uncertain Data. JOURNAL OF COMPUTERS 2014, 9(8):1897-1903.
[7] Y. W. Huang . Research on Dynamic Cost-sensitive Decision Tree for Mining Uncertain Data Based on the Genetic Algorithm. International Journal of Database Theory and Application 2014, 7(5): 201-210.
[8] C. Y. Liu. Cost-Sensitive Clustering for Uncertain Data Based on Genetic Algorithm. International Journal of Applied Mathematics and Statistics 2013. 40(10): 161-169.

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. He is now a lecturer at the Department of Computer and Information Engineering, Heze University. His research interests include the data mining, intelligence Calculation.