代价敏感不确定贝叶斯分类器的单批测试算法

张 星[1],李 梅[1],张 阳[1],宁纪峰[2]

(1. 西北农林科技大学 机械与电子工程学院,陕西 杨凌 712100;2. 西北农林科技大学 信息工程学院,陕西 杨凌 712100)

摘 要:提出了一种针对不确定数据的贝叶斯代价敏感分类器算法SBT-CSUNB用来进行单批测试。SBT-CSUNB算法在代价敏感贝叶斯分类器的框架上定义了不确定数据属性对总代价的影响,提出了单批算法的最优属性集合的选择方式。在UCI数据集上的实验表明:SBT-CSUNB有效地降低了总代价,并且在不同的参数设定下表现平稳,甚至在高不确定率的情况下算法仍旧表现良好。

关键词:人工智能;不确定单批测试;不确定数据;代价敏感

中图分类号:TP391 **文献标志码**:A **文章编号**:1671-5497(2015)02-0583-06

DOI:10.13229/j.cnki.jdxbgxb201502036

Single batch test algorithm on cost-sensitive uncertain Naïve Bayes for uncertain data

ZHANG Xing[1], LI Mei[1], ZHANG Yang[1], NING Ji-feng[2]

(1. *College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling* 712100, *China*; 2. *College of Information Engineering, Northwest A&F University, Yangling* 712100, *China*)

Abstract: In this paper, we propose a single batch test algorithm on Cost-sensitive Uncertain Naïve Bayes for Uncertain Data (SBT-CSUNB). We define the influence of an uncertain attribute on the total cost in cost-sensitive Naïve Bayes Classifier, and put forward a method to fine an optimal batch test strategy. Experiment results on UCI Database demonstrate that the proposed algorithm can effectively reduce the total cost, and the performance is stable with different parameters and under high uncertain rate.

Key words: artificial intelligence; uncertain single batch; uncertain data; cost sensitive

0 引 言

目前,针对不确定数据的分类技术受到了研究者的广泛关注,普遍的做法是将传统的处理确定数据的算法进行扩展,使其能处理不确定数据。常见算法有:基于规则的uRule算法[1],基于决策树的DTU算法[2]、UDT算法[3],基于支持向量机的TSVC[4]、USVC算法[5]、AUSVC算法[6]、基于贝叶斯的NBU算法[7]和FBC算法[8]等。上述不确定数据的分类器都是以分类准确率来评判分类器的优劣,其目的都是使分类错误最小化。当不同的分类错误具有相同的误分类代价时,分类错误最小化能导致误分类的代价最小,然而在现实世界中,不同的错误所付出的代价是不同的,

收稿日期:2014-07-22.

基金项目:国家自然科学基金项目(61003151).

作者简介:张星(1986-),男,博士研究生.研究方向:数据挖掘. E-mail:zhangxing@nwsuaf.edu.cn

对此,研究者们提出了一种以最小化代价为目标的代价敏感学习[9-13]。

文献[14]中提出一种针对不确定数据的代价敏感决策树算法CSDTU。CSDTU是一种进行连续测试的算法,刘明建等[15]在CSDTU的基础上又提出了一种进行单批测试的针对不确定数据的代价敏感决策树算法。但是,文献[12]指出,使用决策树进行单批测试得到的是一个属性的序列,有着严格的先后次序。单批测试仅需要求解测试属性集,基于决策树的算法在求解过程中额外要求测试属性间存在某种次序关系,会影响求解结果的正确性,而使用贝叶斯方法则能有效地避免这种情况。本文提出了一种进行单批测试的针对不确定数据的贝叶斯代价敏感学习算法SBT-CSUNB,该算法在代价敏感贝叶斯分类器的框架上定义了不确定数据属性对总代价的影响,提出了单批算法的最优属性集合的选择方式。

1 问题定义

数据集D由N个样本组成,$D=\{T_1,T_2,\cdots,T_N\}$,每一个样本T_i包括M个属性和1个类别标签。类别有$|c|$种可能,$c=\{c_1,c_2,\cdots,c_{|c|}\}$,属性集为:$A=\{A_1,A_2,\cdots,A_M\}$;其中每一个属性可能为确定的属性,也可能为不确定属性。

为简化问题,本文仅讨论数据集中只含有离散属性的情况,关于连续属性的研究将在以后工作中进行。若属性A_i是不确定属性,用A_i^u来表示。对一个确定的属性A_i,其取值就是值域$Dom(A_i)=\{v_1,v_2,\cdots,v_n\}$中的某个取值$v_k$。而对于不确定离散属性$A_i^u$,其取值用值域$Dom(A_i^u)$上所有可能取值的概率表示,概率可以表示为:$P=\{P_{i1},P_{i2},\cdots,P_{in}\}$,用$A_{ij}^u$表示属性$A_i^u$的概率中的第$j$个值。

在代价敏感学习中,有些属性的属性值已知,测试代价为0,而有些属性的属性值未知,需要花费一定的测试代价才能得到。对数据集D中的一个样本T_j而言,用A表示属性集合;用集合$\tilde{A}$表示测试代价为0的属性集合,即属性值已知的集合;用$\bar{A}$表示测试代价大于0的属性集合,即属性值未知的集合。

对属性值未知的属性进行测试就可以获得其确切的属性值,但这个过程会付出一定的代价,称之为测试代价,记为$cost_{test}$。

在分类器构建好后,对每一个进来的样本,分类器都会给它分类,记为c_j,而该样本本来的类别是c_i,这样就会产生一个误分类代价,记为$cost_{ij}$。误分类代价就是将原本类别为c_i的样本错误地分类为c_j所产生的代价,很显然,$cost_{ii}$为0,并且$cost_{ij}$不等于$cost_{ji}$。

总代价为测试代价和误分类代价的总和,本文的目的就是构造一个能让总代价最小化的分类器。该分类器能实现单批测试,即在不进行任何测试的情况下,一次选择出一批需要检测的属性。这就好比医生不是在你检测完了一项指标后,再根据该指标的情况给你安排下一步检查内容,而是直接给你一系列需要检查的项目。并且该分类器的输入数据中还包含有不确定数据。

2 单批算法

2.1 属性选择方法

在单批算法中,必须在所有未知属性都没有做过测试的情况下给出一系列需要做测试的属性值。为了找到最佳的单批测试属性集合,一种最为简单直接的方式就是找出未知属性集合$\bar{A}$的所有子集,计算每一个子集对总代价的影响,选取总代价减少最多的子集作为需要选取的单批测试属性集合。

在这里,用$Util(\bar{A}_i)$来表示属性$\bar{A}_i$对总代价的影响,其中$\bar{A}_i\in\bar{A}$。

$$Util(\bar{A}_i)=Gain(\tilde{A},\bar{A}_i)-c_{test}(\bar{A}_i) \quad (1)$$

式中:$c_{test}(\bar{A}_i)$为属性$\bar{A}_i$的测试代价;$Gain(\tilde{A},\bar{A}_i)$为添加了属性$\bar{A}_i$后误分类代价减少的值。

$$Gain(\tilde{A},\bar{A}_i)=c_{mc}(\tilde{A})-c_{mc}(\tilde{A}\cup\{\bar{A}\}_i) \quad (2)$$

式中:$c_{mc}(\tilde{A})$为属性值已知集合$\tilde{A}$的误分类代价期望,$c_{mc}(\tilde{A})=\min R(c_j|\tilde{A})$;$c_{mc}(\tilde{A}\cup\bar{A}_i)$为未知属性加入到已知属性集后期望的误分类代价。

$$c_{mc}(\tilde{A}\cup\bar{A}_i)=\sum_{k=1}^{\|\bar{A}_i\|}P(\bar{A}_i=v_{i,k}|\tilde{A})\times \min_{c_j\in c}R(c_j|\tilde{A},\bar{A}_i=v_{i,k}) \quad (3)$$

式中:$P(\bar{A}_i=v_{i,k}|\tilde{A})$表示以已知属性集$\tilde{A}$为前提的情况下,未知属性$\bar{A}_i$的属性值为$v_{i,k}$的概率;$R(c_j|\tilde{A},\bar{A}_i=v_{i,k})$表示在已知属性集为$\tilde{A}$,并且

未知属性的属性值为 $v_{i,k}$ 的前提下类别为 c_j 的概率。

$R(c_j \mid \tilde{A})$ 的计算式为:

$$R(c_j \mid \tilde{A}) = \sum_{k=1}^{|c|} cost_{kj} \times P(c_k \mid \tilde{A}) \quad (4)$$

$$1 \leqslant k \leqslant |c|$$

式中:$cost_{kj}$ 为将原本类别为 c_k 的样本错误地分类为 c_j 所产生的代价;$R(c_j \mid \tilde{A})$ 为已知属性集合 $\tilde{A}$ 的情况下类别为 c_j 的概率。

根据贝叶斯定理有:

$$P(c_k \mid \tilde{A}) = \frac{P(\tilde{A} \mid c_k) P(c_k)}{P(\tilde{A})} \quad (5)$$

式中:$P(c_k)$ 和 $P(\tilde{A})$ 都是常量;$P(\tilde{A} \mid c_k) = \prod_{A_i^u \in \tilde{A}} P(A_i^u \mid c_k)$。

对其中的一个属性 $A_i{}^u$ 进行计算,则有:

$$P(A_i^u \mid c_k) = p_{i1} P(A_{i1}^u \mid c_k) + p_{i2} P(A_{i2}^u \mid c_k) + \cdots + p_{ij} P(A_{ij}^u \mid c_k) \quad (6)$$

$P(A_{ij}{}^u \mid c_k)$ 的值可以很容易计算出来,假设 $A_{ij}{}^u$ 的属性值是 v_m,则有:

$$P(A_{ij}^u = v_m \mid c_k) = \frac{PC(v_m, c_k)}{PC(c_k)} \quad (7)$$

式中:$PC(c_k)$ 为该数据集中所有类别为 c_k 的概率势,对于一个样本 T_j,用 c_{T_j} 表示其类别,那么 $PC(c_k)$ 可以表示为:

$$PC(c_k) = \sum_{j=1}^{|D|} P(c_{T_j} = c_k) \quad (8)$$

同样地,$PC(v_m, c_k)$ 表示类别为 c_k、属性值为 v_m 的概率势,是数据集中所有样本的类别为 c_k 且属性值为 v_m 的概率之和,表示为:

$$PC(v_m, c_k) = \sum_{j=1}^{|D|} P(v_m \in T_j \wedge c_{T_j} = c_k) \quad (9)$$

为了找到最佳的单批测试算法的未知属性集合,用 $\bar{A}'$ 表示未知属性集合 $\bar{A}$ 的子集,那么有($\bar{A}' \subseteq \bar{A}$)。对每一个子集 $\bar{A}'$,都计算出它的 $Util(\bar{A}')$,从中选择 $Util(\bar{A}')$ 值最大的未知属性集合,这样就可以找出最佳的单批测试算法的未知属性集合。但是,这种方式的计算效率非常低,当 $\bar{A}'$ 组合方式有很多种时,计算所有可能的 $Util(\bar{A}')$ 值会耗费大量资源,花费很多时间,本文采用贪心算法的思路来解决这种问题。

对任意一个未知属性 $\bar{A}_i$,计算其 $Util(\bar{A}_i)$ 值,如果 $Util(\bar{A}_i) > 0$,则该未知属性对减少总代价是有意义的,将其加入到子集 $\bar{A}'$ 中去,这样可以得到一个需要进行测试的未知属性集合。用 $\bar{A}'$ 表示单批测试被选出需要进行测试的未知属性集合,那么可以将其表示为:

$$\bar{A}' = \{\bar{A}_i \mid Util(\bar{A}_i) > 0, \bar{A}_i \in \bar{A}\}$$

2.2 SBT-CSUNB 算法

上面介绍了进行单批测试时未知属性的具体选取方法,下面将详细介绍该单批测试算法的具体步骤。

该算法基于贝叶斯定理,由训练和预测两部分组成。首先是训练过程,从训练数据集 D 中学习一个分类器的本质是根据训练数据对相关的概率值进行统计。本文分类器的构建过程就是统计数据集 D 中 $PC(c_k)$ 和 $PC(v_k, c_l)$ 的值,在此不做赘述。

对一个测试用例,用 2.1 节中介绍的属性选择方式选取出需要进行测试的未知属性,预测该测试用例的类别,使得该预测总代价值最小。下面将详细介绍单批测试算法的预测过程。其算法具体步骤如下:

Algorithm: Single Batch Test of Cost-sensitive Uncertain Naïve Bayes

Input: T_j (atext example), B (an cost-sensitive uncertain naïve Bayes Classier)

Output: classlabel

Begin:

1. let $\tilde{A}$= the set of known attributes
   let $\bar{A}$= the set of unknown attributes.
2. set $cost_{test}=0$
3. while ($\bar{A} \neq \varnothing$) do
4. for(each $\bar{A}_i \in \bar{A}$)
5. calculate Util($\bar{A}_i$)
6. if (Util($\bar{A}_i$)>0) then
7. $\tilde{A} \cup \{\bar{A}_i\} \rightarrow \tilde{A}$
8. $\bar{A} - \{\bar{A}_i\} \rightarrow \bar{A}$
9. do $c_{test} = c_{test} + c_{\bar{A}_i}$
10. end if
11. end for

12. end while

13. return arg $\min_j$ ($R(c_j \mid \tilde{A}) + c_{test}$)

(1)在上述算法中用 $\tilde{A}$ 代表已知属性值的属性集合,用 $\bar{A}$ 代表未知属性值的属性集合,初始时的 $cost_{test}$ 值记为0(见算法步骤1和2)。

(2)当 $\bar{A}$ 不是空集时,对 $\bar{A}$ 中的每个属性 $\bar{A}_i$,计算 $Util(\bar{A}_i)$ 的值(见算法步骤4~步骤6)。

(3)对于每一个 $Util(\bar{A}_i)$ 大于0的属性,将该属性从未知属性集合 $\bar{A}$ 移除,并将其加入到集合 $\tilde{A}$ 中(见算法步骤7和8)。

(4)将测试 $\bar{A}_i$ 所花费的测试代价加入到总的测试代价中(见算法步骤9)。

(5)得到预测的类别,该类别为总的代价最小的类别(见算法步骤13)。

3 实 验

为了验证SBT-CSUNB算法的性能,本文在UCI数据集上进行了实验,基于WEKA软件实现了SBT-CSUNB算法。实验环境为Intel Core2 Duo 2.53 GHz CPU和2.0 GB主存,在UCI数据库中选择了8组真实的二类别数据集。实验中所使用的数据集在表1中详细列出。

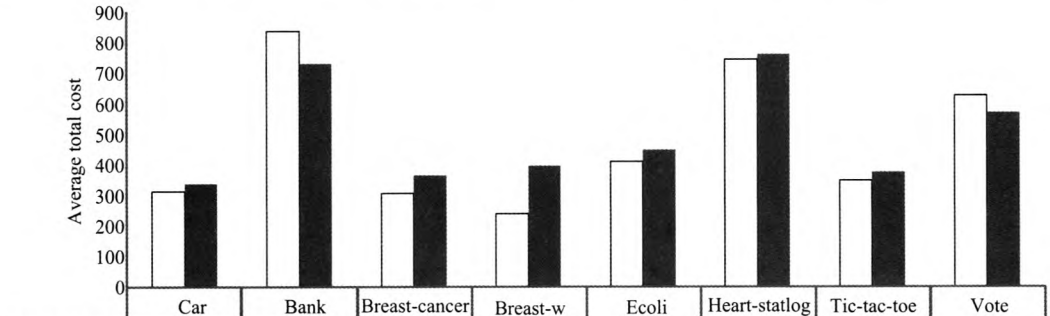
表1 实验中使用的数据集

Table 1 Dataset used in the experiment

数据集	属性数	样本数	类别分布(正/负)
Breast-w	10	699	458/241
Vote	17	435	267/168
Car	7	1733	1211/522
Bank	11	600	274/326
Breast-cancer	10	286	201/85
Ecoli	8	336	220/116
Heart-statlog	14	270	150/120
Tic-tac-toe	10	985	322/626

由于缺少真正的不确定数据集,本文采用文献[7]中引入不确定性的方法。向选取的数据集中引入不确定性,将现实中存在的确定数据转化为不确定数据。当引入10%的不确定性时,将原属性值的概率记为0.9,将剩下的0.1随机分配到剩下的属性值中去。即对原始的确定数据集 $E.A_i^{v_i} = v_k$,记 $P_{ik} = 0.9$,那么对于其他的属性值的概率 $p_{ik}(2 \leqslant k \leqslant n)$,它们的和 $\sum_{k=2}^{n} p_{ik} = 0.1$。下文中,用 $U10$ 表示不确定率为10%的情况。

图1给出了 $U10$、$cost_1/cost_2 = 1000/1000$ 时,SBT-CSUNB算法与算法SingleB[15]在每个数据集上总代价的比较结果。



	Car	Bank	Breast-cancer	Breast-w	Ecoli	Heart-statlog	Tic-tac-toe	Vote
□SBT-CSUNB	313.2092	837.0923	307.0443	237.3224	411.7625	743.2554	346.5570	627.6959
■SingleB	336.9284	730.7590	364.6673	396.0867	451.0723	762.6115	374.5733	571.7302

图1 两种方法的平均总代价比较

Fig. 1 Average total cost comparisons of two methods

从图1可以看出:在Breast-cancer、Breast-w和Tic-tac-toe数据集上,SBT-CSUNB较SingleB在总代价上有大幅度的降低,说明SBT-CSUNB在这3个数据集上优势明显;在Car、Ecoli和Heart-statlog数据集上,SBT-CSUNB较SingleB在总代价上分别有7.03%、8.71%、2.54%的降低,说明SBT_CSUB在这3个数据集上表现略优于SingleB;在Bank和Vote数据集中,SBT-CSUNB较SingleB在总代价上分别有14.55%和9.78%的提升,说明在这两个数据集中SBT-CSUNB的表现要略逊于SingleB。

由表1可知:在进行实验的8个数据集上,有6个数据集SBT-CSUNB算法的表现优于SingleB。这是因为SingleB算法基于决策树,决策树算法有着严格层次结构,前面节点的选择对后面节点的选择有较大影响,这种结构在进行一

次选择多个属性的单批算法时不够灵活,而基于贝叶斯的 SBT-CSUNB 算法则没有这方面的限制,所以其性能要更优。

为进一步比较两种算法,本文采用文献[14]和文献[15]中衡量代价敏感算法优劣的宏平均方法,即将所有进行实验的数据集的总代价平均值进行比较。图 2 给出了 $cost_1/cost_2=1000/1000$、U 以 10 为步长从 0 到 50 变化时 8 个数据集在 SBT-CSUNB 算法和 SingleB 算法上总代价的比较结果。

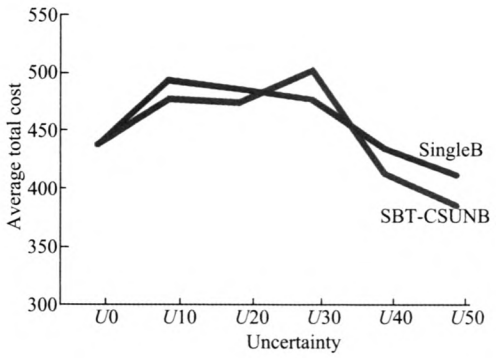


图 2　两种方法的宏平均代价比较

Fig. 2　Macro average cost comparisons of two methods

从图 2 可以看出:在比较所有进行实验的数据集的总代价平均值时,除了 $U30$,SBT-CSUNB 的总代价平均值都小于 SingleB。这进一步说明了 SBT-CSUNB 算法的表现优于 SingleB。

SBT-CSUNB 算法在处理 Vote 和 Bank 数据集的表现并不理想。这可能是因为 Vote 和 Bank 数据集中属性之间存在联系,而 SBT-CSUNB 算法是基于贝叶斯算法的,需要服从属性间相互独立的贝叶斯假设,故 SBT-CSUNB 算法在这些数据集上性能有所下降。

图 3 给出了 $cost_1/cost_2=1000/1000$、U 以 10 为步长从 0 到 50 变化时数据集 Ecoli 和 Heart 在 SBT-CSUNB 算法和 SingleB 算法上总代价的比较结果。

从图 3 可以看出:随着不确定性的变化,总花费会有起伏,但相比 SingleB 算法,SBT-CSUNB 算法要更稳定一点。值得注意的是,在不确定性增大到一定程度时,总代价会有一定的下降。这是因为当不确定程度较大时,其测试过程会减少,会导致测试代价的减少,这就使得总代价会有一定的下降。另外,可以看出,SBT-CSUNB 算法在多数情况下总代价都小于 SingleB 算法。为了观察 $cost_1/cost_2$ 比值变化对 SBT-CSUNB 算法性能的影响,图 4 给出了 $cost_1/cost_2=1000/1000$,$cost_1/cost_2=600/1000$,$cost_1/cost_2=1000/2000$

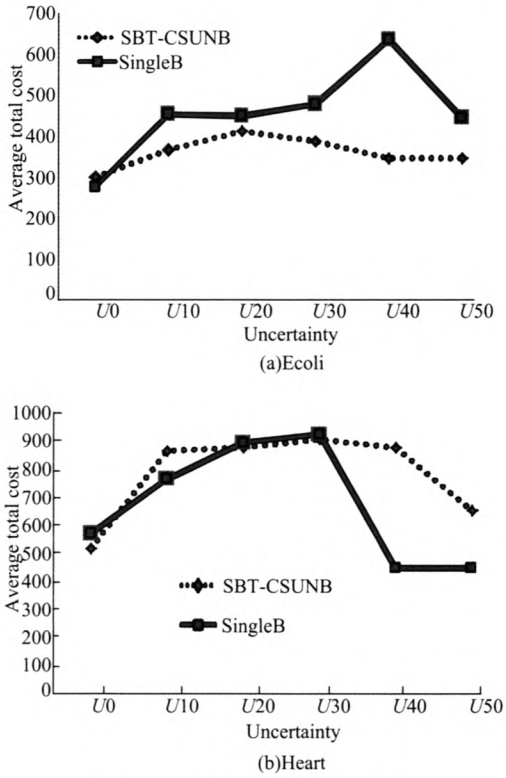


图 3　不确定率变化时的比较

Fig. 3　Comparisons with varying uncertainty

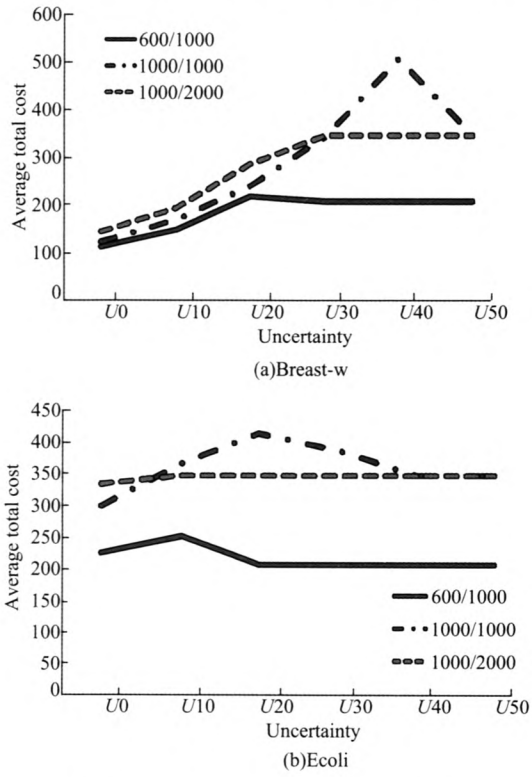


图 4　$cost_1/cost_2$ 比值变化时的比较

Fig. 4　Comparisons with different $cost_1/cost_2$

时数据集 Breat-w 和 Ecoli 在 SBT-CSUNB 算法上的结果。

从图 4 可以看出:当 $cost_1/cost_2$ 比值不同时,总代价的值是不相同的。这是由于 $cost_1/cost_2$ 比值不同时,会导致误分类需要付出的代价不同,故其值会变化。但不同线条的大致走势是相同的,这说明 $cost_1/cost_2$ 比值不同时,算法的性能稳定。

4 结束语

提出了一种进行单批测试的针对不确定数据的贝叶斯代价敏感学习算法 SBT-CSUNB。相比已有的基于决策树的该类算法 SingleB,基于贝叶斯的 SBT-CSUNB 避免了层次结构僵硬,不适用于单批测试的缺点。实验结果表明:SBT-CSUNB 有效地降低了总代价,并且在不同的参数设定下表现平稳,甚至在高不确定率的情况下,算法仍旧表现良好。

在未来的工作中,会将上述工作扩展到具有连续属性的数据集中,并进一步细化代价的种类,将测试代价和误分类代价做出更精确的划分。

参考文献:

[1] Qin B, Xia Y, Prabhakar S, et al. A rule-based classification algorithm for uncertain data[C]// 25th International Conference on Data Engineering. Shanghai: IEEE, 2009: 1633-1640.

[2] Qi B, Xia Y, Li F. DTU: a Decision Tree for Uncertain Data [M]. Berlin Heidelberg: Springer, 2009: 4-15.

[3] Tsang S, Ben K, Yip K Y, et al. Decision trees for uncertain data[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(1): 64-78.

[4] Bi J, Zhang T. Support vector classification with input data uncertainty[J]. Advances in Neural Information Processing Systems, 2004, 17: 161-169.

[5] Yang J, Gunn S. Exploiting uncertain data in support vector classification[C]// Knowledge Based Intelligent Information and Engineering Systems. Heidelberg. Berlin: Springer, 2007: 148-155.

[6] Yang J, Gunn S. Iterative constraints in support vector classification with uncertain information[J]. Constraint-based Mining and Learning, 2007, 1: 49-60.

[7] Qin B, Xia Y, Li F. A Bayesian classifier for uncertain data[C]// Proceedings of the 2010 ACM Symposium on Applied Computing. New York: ACM, 2010: 1010-1014.

[8] Ren J, Lee S D, Chen X, et al. Naive Bayes classification of uncertain data[C]// Ninth IEEE International Conference on Data Mining. Miami, FL: IEEE, 2009: 944-949.

[9] Ling C X, Yang Q, Wang J, et al. Decision trees with minimal costs[C]// Proceedings of the Twenty-first International Conference on Machine Learning. New York: ACM, 2004: 69.

[10] Zubek V B, Dietterich T G. Pruning improves heuristic search for cost-sensitive learning[R]. Corvallis, OR: Oregon State University, Dept of Computer Science, 2004.

[11] Turney P. Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm[J]. Journal of Artificial Intelligence Research (JAIR), 1995, 2: 369-409.

[12] Chai X, Deng L, Yang Q, et al. Test-cost sensitive naive bayes classification[C]// Fourth IEEE International Conference on Data Mining, IEEE, 2004: 51-58.

[13] Turney P. Types of cost in inductive concept learning[C]// In Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning, Stanford 2000: 1-7.

[14] Liu M, Zhang Y, Zhang X, et al. Cost-sensitive Decision Tree for Uncertain Data [M]. Heidelberg, Berlin: Springer, 2011: 243-255.

[15] 刘明建,张阳,王勇. 代价敏感不确定决策树的不确定单批测试算法研究[J]. 工程数学学报, 2012, 29(4): 559-566.
Liu Ming-jian, Zhang Yang, Wang Yong. Uncertain single batch test algorithm on cost-sensitive decision tree for uncertain data[J]. Chinese Journal of Engineering Mathematics, 2012, 29(4): 559-566.