文章编号: 1001-0920 (2016) 09-1609-06 **DOI:** 10.13195/j.kzyjc.2015.1110

基于自适应快速决策树的不确定数据流概念漂移分类算法

刘志军[1], 张 杰[2], 许广义[1]

(1. 哈尔滨工程大学 经济管理学院，哈尔滨 150001；2. 山东科技大学 经济管理学院，山东 青岛 266590)

摘 要: 由于不确定数据流中一般隐藏着概念漂移问题, 对其进行有效分类存在着很多困难. 为此, 提出一种基于自适应快速决策树的算法. 该算法基于一般决策树算法的原理, 以自适应学习规则计算信息增益, 以无标记情景学习拆分原理检测不确定数据流中的不确定数值属性, 通过自适应快速决策树节点的拆分方法将不确定数值属性转化为不确定分类属性, 以实现对不确定数据流的有效分类, 进而有效检测到其中隐含的概念漂移现象. 仿真结果验证了所提出方法的可靠性.

关键词: 不确定数据流；自适应快速决策树；概念漂移；数值属性；分类属性

中图分类号: F830.51 **文献标志码:** A

Classifying algorithm for concept drifting of uncertain data streams based on adapting fast decision tree algorithm

LIU Zhi-jun[1], *ZHANG Jie*[2], *XU Guang-yi*[1]

(1. School of Economics and Management，Harbin Engineering University，Harbin 150001，China；2. School of Economics and Management，Shandong University of Science and Technology，Qingdao 266590，China. Correspondent：ZHANG Jie，E-mail：zhangjie0371@163.com)

Abstract: Because of the concept drift problem hidden in the uncertain data stream, it is very difficult to classify them effectively. Based on the general decision tree algorithm, the adaptive fast decision tree algorithm can count information gain based on the adaptive learning rule, and detect uncertain numerical attributes though the principle of the non-marking learning scene. The numerical attribute is transformed into a non-determined classification attribute by using splitting method, so classification of uncertain data stream is realized effectively. Then the concept drift phenomenon is effectively detected in the uncertain data stream. Simulation results show the reliability of the proposed method.

Keywords: uncertain data streams；adapting fast decision tree；concept drifting；numerical attribute；classification attribute

0 引 言

近年来, 随着信息技术和大数据的飞速发展产生了大量的不确定数据流, 不确定数据流的分类问题在工程应用领域和信息领域越来越重要. 然而, 不确定数据流因时间、空间的不断变化而使其分类面临着诸多障碍, 并因其存在着概念漂移现象, 使得对其进行有效分类变得非常困难. 不少学者对不确定数据流的分类问题进行了研究, 如 Pan 等[1]采用贝叶斯原理对不确定数据进行了简单的分类; Gao 等[2]提出了基于规则分类算法实现对不确定数据流的初步聚类; Qin 等[3]将支持向量机 (SVM) 技术推广到不确定数据流的分类研究中. 总之, 这些算法都是针对不确定数据流的某些特征或属性进行分类, 因没有抓住分类问题的本质而不能够从整体上对不确定数据流进行有效的聚类或分类.

随着信息技术的进一步发展, 一些研究开始关注数据流的在线学习、自适应学习和监督学习问题, 还有一些研究将不确定性或未标记学习规划引入数据流的分类中, 如 Aggarwal 等[4]提出的 UMicro 算法, 该算法通过添加误差维度进行自适应学习, 并通过修改 micro-cluster 结构对属性不确定的数据流进行初步聚类; Pang 等[5]提出了一类支持向量机决策树算法,

收稿日期: 2015-09-07；**修回日期:** 2015-11-15.

基金项目: 山东省自然科学基金项目(ZR2015GM013)；全国统计科研计划重点项目(2015LZ25)；中国博士后基金项目(2015M581757).

作者简介: 刘志军(1990−), 男, 博士生, 从事数据挖掘、创新管理的研究；许广义(1963−), 男, 教授, 博士生导师, 从事技术经济、数据挖掘等研究.

采用阳性未标记样本实现对不确数据流中概念漂移的检测; Shaker 等[6]提出了基于集成算法以解决数据流中不确定标签问题; 肖丹萍等[7]提出了一种基于免疫原理算法, 以实现对不确定数据流进行聚类; 邢长征等[8]基于网格特征向量存储数据的分布特征, 采用更新网格特征向量, 并将其合并成簇, 以此对不确定数据流进行分类.

总而言之, 现有的算法依据不确定数据流某些属性或特征对其进行初步聚类或分类, 大部分算法采用监督学习方法或标记样本方法, 也有一部分算法采用阳性未标记样本方法, 但这些算法未考虑不确定数据流中的概念漂移问题, 未考虑由于概念漂移问题的存在而使不确定数据流中的某些属性随着时间的变化而发生了较大的改变, 从而只能对其进行简单的分类或聚类, 不能够对其中的渐进或突变的概念漂移问题进行有效处理. 基于此, 本文提出了自适应快速决策树算法 (AFDT), 该算法能够快速地将不确定数据流的特征属性分类为不确定数值属性和不确定分类属性, 并将两种属性与概念漂移问题紧密地联系起来, 从而实现对不确定数据流的有效分类及对其中的概念漂移问题的有效检测.

1 问题描述

定义 1 设数据流 S 表示由一个相互独立的 d 维不确定元组构成的序列, 即

$$S=\{(X_1,P_1),(X_2,P_2),\cdots,(X_n,P_n)\}.$$

其中: (X_i,P_i) 为不确定数据元组, $X_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(d)})$ 为 d 维元组的属性值, $1\leqslant i\leqslant n$, $P_i\in[\theta,1]$ $(0\leqslant\theta\leqslant 1)$ 为该元组的存在概率. 则称数据流 S 为不确定数据流[8].

定义 2 设联合概率 $p(x,y)$ 在相邻时间段内发生改变, 即

$$p_{t-1}(x,y)\neq p_t(x,y).$$

其中: x 为样本, y 为类别标签. 在此基础上, 联合概率可等价转换成 $p(x,y)=p(x)\cdot p(y/x)$. 当 $p(x)$ 改变而 $p(y/x)$ 不变时, 称为特征改变, 又称缓漂移; 当 $p(y/x)$ 改变而 $p(x)$ 不变时, 称为条件改变, 又称突漂移或实漂移[9].

一般地, 数据流中两种概念漂移现象都会存在, 难以有效区分. 目前, 针对概念漂移现象的检测主要有两种思路: 一是如何发现和检测概念漂移; 二是当概念漂移发生后如何迅速地解决漂移问题[9]. 数据流中存在着不确定性属性和不确定性数值, 不确定数据的属性往往被作为分类的依据, 若将 x^u 作为数据流的不确定属性, 则用 x_t^u 表示在 x^u 中的第 i 类属性. 一类属性 x_t^u 可以分为一类不确定数值属性 (UNA) 或不确定分类属性 (UCA). 在此基础上, 可以定义如下不确定数值属性和不确定分类属性.

定义 3 设数据流 S 为 d 维不确定数据流, 令 x_t^{un} 为 S 的第 i 个不确定属性, 且 x_{it}^{un} 可以写成 x_i^{un} 的第 t 个样本 x_t^u 的值 s_t, 若该不确定属性 x_t^{un} 的值可表示为 $x_{it}^{un}=\{f_{it}(x)/x\in[a_{it},b_{it}]\}$, 概率分布函数 $f_{it}(x)$ 在滑动窗口 $[a_{it},b_{it}]$ 内, 且 $\int_{b_{it}}^{a_{it}}f_{it}(x)\,\mathrm{d}x=1$, 则称 x_{it}^{un} 为不确定数值属性.

一般而言, 若某数值属性 x_i 中的值 v_{it} 可以建模为 x_i^{un} 的一个特例, 且 $f_{it}(x)=1$, $a_{it}=b_{it}=v_{it}$, 则数据流中所有的数值属性均可以被不确定分类属性所替代.

定义 4 设数据流 S 为 d 维不确定数据流, 令 x_k^{uc} 为 S 的第 k 个不确定属性, 且 x_{it}^{uc} 可写成 x_k^{uc} 在第 m 个样本 x_m^u 的值 s_m, 若 x_{it}^{uc} 的概率分布密度 $f_{it}(x)$ 为分类域 $\phi(x_k^{uc})=\{v_1,v_2,\cdots,v_m\}$ 上的分类函数, 且由概率向量 $p=(p_{t1},p_{t2},\cdots,p_{tm})$ 组成, 则称 x_k^{uc} 为不确定分类属性.

一般而言, 若数据流中所有特征值 $p(x_{it}^{uc}=v_j)=p_{ij}$, $\sum_{j-1}^{m}p_{ij}=1$, 则分类属性 x_i 中的值 v_j 可以建模为 x_k^{uc} 的一个特例, 且对所有的 $p(x_{it}^{uc})=1$ 和 $p(x_{it}^{uc}=v)=0$, 可以得出数据流中目标概念发生了改变或存在概念漂移.

2 自适应快速决策树算法原理

2.1 不确定数值属性估计

针对不确定数据流数据量巨大的特点, 自适应快速决策树算法首先估计不确定数据流的不确定数值属性; 然后对不确定数值属性进行拆分, 将其转化为不确定分类属性, 并将不确定分类属性作为不确定数据流的分类依据对其进行分类, 进而对不确定数据流中的概念漂移进行检测. 估计不确定数据流不确定数值属性的方法是通过保持足够的数据样本传入数据流, 采用分裂和区分样本的不确定数值数据和不确定分类数据, 同时使用足够的统计数据量来评估这种分裂方式的方法. 基于此, 令集合 s_R 为自适应快速决策树根目录 R 下所观察到的样本, 设定不确定数据集 s_n 为在决策树节点 N 处观察到的样本, 假设属性 x_i^u (数值或分类) 被选作分割属性, 并且其将不确定数据集 s_n 分裂为 m 个子集 $\{s_{i1},s_{i2},\cdots,s_{im}\}$, 则不确定数据集 s_n 的信息增益为

$$\mathrm{IG}(s_n,x_i^u)=\mathrm{Entropy}(s_N)-\sum_{j=1}^{m}\frac{P(s_{ij})}{P(s_N)}\times\mathrm{Entropy}(s_{ij}),\qquad(1)$$

其中 $P(s)=\sum_{s_t\in S} w_t$ 为在 s_n 中的样本基数概率, 则相应的信息熵[10]被定义为

$$\text{Entropy}(s)=-p_k\log_2 p_k. \tag{2}$$

其中: p_k 为当前不确定数据集 s_n 的阳性目标样本的估计比 $(k=1)$ 或阴性目标样本的估计比 $(k=0)$, p_1 和 p_2 可以通过下式求得:

$$\begin{aligned} &p_1=\min\Big(\frac{p(s,l_1)}{p(s_R,l_1)}\times\tau\times\frac{p(s_R,l_0)}{p(s,l_0)},1\Big),\\ &p_0=1-p_1. \end{aligned} \tag{3}$$

其中: $p(s,l_k)=\sum_{s_t\in S} w_t\times p(l_k(s_t)=k)$ 为不确定数据集 s_n 中阳性样本的基数概率 $(k=1)$ 或未标记的目标样本 $(k=0)$ 的基数概率, $\tau\in[0,1]$ 为不确定数据集 s_n 中数值属性水平. 由此可以看出, 求解不确定数据流信息增益量的关键是确定基数概率 $P(s_N)$、$P(s_N,l_k)$、$P(s_R,l_k)$、$P(s_{ij})$ 和 $P(s_{ij},l_k)$, 其求解过程见以下部分.

现假设属性 X_i^{un} 为分割属性, 分割点分别为 z_i, 则其拆分数据集 S_N 为两个子集数的目标样本, 其基数概率计算如下:

$$\begin{aligned} &P(s_{i1},l_k)=\sum_{s_t\in s_{i1}} w_t p(l_k(s_t)=k)=\\ &\sum_{s_t\in s_N} w_t\int_{a_{it}}^{z_{i1}} f_{it}(x)\mathrm{d}x\times p(l_k(s_t)=k), \end{aligned} \tag{4}$$

$$\begin{aligned} &P(s_{i2},l_k)=\sum_{s_t\in s_{i2}} w_t p(l_k(s_t)=k)=\\ &\sum_{s_t\in s_N} w_t\int_{a_{it}}^{z_{i2}} f_{it}(x)\mathrm{d}x\times p(l_k(s_t)=k), \end{aligned} \tag{5}$$

$$P(s_{i1})=\sum_{s_t\in s_{i1}} w_t=\sum_{s_t\in s_N} w_t\times\int_{a_{it}}^{z_t} f_{it}(x)\mathrm{d}x, \tag{6}$$

$$P(s_{i2})=\sum_{s_t\in s_{i2}} w_t=\sum_{s_t\in s_N} w_t\times\int_{z_t}^{b_{it}} f_{it}(x)\mathrm{d}x. \tag{7}$$

显然, 式 (4) ∼ (7) 确定了两个子集的基数概率.

2.2　不确定数值属性拆分

为了实现对不确定数据流进行有效不确定数值属性的拆分, 现设 $x_i^{uc}\in x^u$ 为自适应快速决策树节点 N 处的分裂属性, 其将数据集 s_t 分割为 m 个样本 $\{s_{t1},s_{t2},\cdots,s_{tm}\}$. 其中: $m=|\phi(x_i^{uc})|$, s_{tj} 为 s_t 除了属性 x_i^{uc} 的属性值, 且此时所有 $v_i\neq v_j, p(x_{ij}^{uc}=v_j)=0$. 因为 s_{tj} 是属性 x_i^{uc} 在 s_t 上分裂的样本, 即被分配给决策树节点N的权重, 所以 s_{tj} 被分配到 $w_{tj}=w_t^* p(x_{it}^{uc}=v_j)$ 中, 属性 x_i^{un} 在不确定数据集 s_t 分割区间 $[a_{it},b_{it}]$ 上的 2 分间隔点为 z_i, 相应的概率 p_{i1} 和 p_{i2} 分别为

$$p_{i1}=\int_{a_{it}}^{z_i} f_{it}(x)\mathrm{d}x, \tag{8}$$

$$p_{i2}=\int_{z_i}^{b_{it}} f_{it}(x)\mathrm{d}x. \tag{9}$$

若 $z_i<a_{it}$, 则 $p_{i1}=0$; 同理, 若 $z_i>b_{it}$, 则 $p_{i2}=0$. 因此, 数据集 s_t 将被分成两个子目标样本 s_1 和 s_2, 二者将被分别拆分到决策树节点 N 的左右两个节点. 需要指出的是, 权重 w_t 可以被解释为数据集 s_t 的节点 N 的关联概率, s_1 的权重按 $w_1=w_t\times p_{i1}$ 进行分配, 则 s_1 的概率密度 $f_1(x)$ 可由下式求得:

$$f_1(x)=\begin{cases}\dfrac{f_{it}(x)}{w_1},\ x\in[a_{it},z_i];\\ 0,\ x\notin[a_{it},z_i].\end{cases} \tag{10}$$

进而有

$$x_{it_1}^{un}=\Big\{\frac{f_1(x)}{x}\in[a_{it},z_i]\Big\}, \tag{11}$$

$$s_i=\langle x^u\bigcup x_{it_1}^{un}-x_{it}^{un},l_k\rangle. \tag{12}$$

同理, s_2 也有类似的计算公式.

一般地, x_{it}^{un} 作为一个特定属性, 应具有一个指向值 v_{it}, 如果 $v_{it}\leqslant z_i$, 则 s_t 将向下传递到节点 N 的左节点, 否则进入到右节点, 从而实现对数据流 s_n 的不确定数值属性的有效拆分.

2.3　不确定分类属性与概念漂移检测

要实现对不确定数据流中的不确定数值属性进行分类并检测其中的概念漂移现象, 需要将不确定数值属性转化为不确定分类属性[11], 在此基础上才能够检测到不确定数据流中的概念漂移现象. 因此, 首先, 需要对拆分后的数值属性进行初始化, 对于一个新自适应快速决策树叶片 N, 用 l_k 初始化每个不确定属性 x_i^{un}, 且 $x_i^{un}\in x^u$, 将 $(\overline{x_{ik}^{un}},\sigma_{ik}^2,\Sigma_{ik})$ 设置为 $(0,0,0)$; 然后, 进行离散化处理, 在不确定滑动窗口 (a_{it},b_{it}) 内, 将 s_t 分成 M 个均匀单元, 并设 $\Omega=1/M$, 则 $f_{it}(x)$ 可以由 $f_{it}(n\Omega)$ 近似. 其中 $n\in I_M=\{a,1,\cdots,M\}$. 因此, 一个不确定数值属性可由一个序列权重对 (v_n,w_n) 近似替代. 其中: $w_n=w_t\Omega f_{it}(n\Omega)$, $v_n=n\Delta$. 拆分后的新样本 N' 对于每一对 (v_n,w_n), 充分统计量 $(\overline{x_{ik}^{un}},\sigma_{ik}^2,\Sigma_{ik})$ 可以由下式进行更新:

$$\begin{cases}\Sigma_{ik}=\Sigma_{ik}+wn,\\ \psi(x)=\overline{X}_{ik}^{u},\\ \overline{X}_{ik}^{un}=\overline{X}_{ik}^{un}+w_n\dfrac{v_n-\psi(x)}{\Sigma_{ik}},\\ \sigma_{ik}^2=(\Sigma_{ik}-w_n-1)\,\sigma_{ik}^2+\\ \qquad w_n\dfrac{(v_n-\psi(x))(v_n-\overline{X}_{ik}^{un})}{\Sigma_{ik}-1}.\end{cases} \tag{13}$$

其中: 若 $\Sigma_{ik}\leqslant 1$, 则 $\sigma_{ik}^2=0$. 自适应快速决策树算法保持与目标样本的滑动窗口一致, 若训练样本离开滑动窗口而变得过时, 因作为一个过时的目标样本不能代表目前的任何概念, 故删除该目标样本所对

应的每一对 (v_n, w_n), 则此时的充分统计量 $(\overline{x_{ik}^{un}}, \sigma_{ik}^2, \Sigma_{ik})$ 可通过下式进行更新:

$$\begin{cases} \Sigma_{ik} = \Sigma_{ik} - w_n, \\ \psi(x) = \overline{X}_{ik}^{un} - w_n \dfrac{v_n - \overline{X}_{ik}^{un}}{\Sigma_{ik}}, \\ \sigma_{ik}^2 = (\Sigma_{ik} + w_n - 1)\sigma_{ik}^2 - \\ \qquad w_n \dfrac{(v_n - \psi(x))(v_n - \overline{X}_{ik}^{un})}{\Sigma_{ik} - 1}, \\ \overline{X}_{ik}^{un} = \psi(x). \end{cases} \tag{14}$$

其中: 若 $\Sigma_{ik} \leqslant 1$, 则 $\sigma_{ik}^2 = 0$. 通过式 (14) 对充分统计量值的测定, 能够对不确定数据流中的概念漂移现象进行检测, 在此基础上, 还需要对概念漂移的种类进行分类. 基于此, 令

$$V_{ik}^{\min} = \overline{X}_{ik}^{un} - \mu \times \sigma_{ik}^2, \tag{15}$$

$$V_{ik}^{\max} = \overline{X}_{ik}^{un} - \mu \times \sigma_{ik}^2, \tag{16}$$

其中 μ 为指定的置信度阈值参数. 此时, 分割点 z_i 的滑动窗口被确定为 (最小值, 最大值), 即 $\min\text{value} = \min(v_{ik}^{\min})$, $\max\text{value} = \max(v_{ik}^{\max})$, 且有 $k \in \{0, 1\}$. (最小值, 最大值) 被分为 Z 的一部分, 设 $\Omega = 1/Z$, 对于不确定属性 x_i^{un}, 每个 $z_i = n\Omega$ (其中 $n \in I_Z = \{1, 2, \cdots, Z\}$) 及给定 Z, 有

$$\begin{cases} P(S_{i1}, l_k) = \Sigma_{ik} F(\overline{X}_{ik}^{u}, \sigma_{ik}^2, z_i), \\ P(S_{i2}, l_k) = \Sigma_{ik} - P(S_{i1}, l_k), \\ P(S_N, l_k) = \Sigma_{ik}, \\ P(S_N) = \displaystyle\sum_{k \in \{0,1\}} P(S_N, l_k) = \Sigma_{i0} + \Sigma_{i1}. \end{cases} \tag{17}$$

若属性 x_i^{uc} 被选作分割属性, 则其分裂集 s_N 被拆分成 $m = |\phi(x_i^{uc})|$ 个子集, 新旧样本中概念的差额由属性 x_i^{un} 超过 s_{ij} 的基数概率确定, 由下式给出:

$$P(s_{ij}) = \sum_{s_t \in s_N} w_t \times P(x_{it}^{uc} = v_j), \tag{18}$$

$$p(s_{ij}, l_k) = \sum_{s_t \in s_N} w_t \times P(x_{it}^{uc} = v_j) \times P(l_k(s_t) = k). \tag{19}$$

因此, 在自适应快速决策树的每个节点上, l_k 的每个属性 $x_i^{uc} \in x^u$ 的每个可能值 v_j 将其以足够的统计量 n_{ijk} 关联起来. 为了测评统计数据的概念漂移程度, 只需作超过 s_N 的设定, 对于每个更新后带权重 w_t 的目标样本 s_t ($s_t \in s_N$), 将其对应的统计量 n_{ijk} 按下式进行更新:

$$n_{ijk} = n_{ijk-1} + w_t \times p(l_k(s_t) = k) \times p(x_{it}^{uc} = v_j). \tag{20}$$

同样地, 若从节点 N 移除样本 s_t, 则统计数据按下式更新:

$$n_{ijk} = n_{ijk-1} - w_t \times p(l_k(s_t) = k) \times p(x_{it}^{uc} = v_j). \tag{21}$$

因此, 通过统计量 n_{ijk} 变动情况就能够检测出不确定数据流中的目标概念的漂移程度, 进而判断其属于缓漂移还是突漂移.

3 仿真算例

3.1 样本数据描述

为了测评自适应快速决策树算法对不确定数据流的分类与概念漂移的检测性能, 现对现实世界所谓“森林覆盖”型不确定数据序列进行仿真实验, 数据集来源参见文献 [12] 和 文献 [13].“森林覆盖”型数据序列是一个有着多级分类漂移问题的不确定数据流, 训练样本为按照序列生成的不确定数据流, 样本量约为 50 000 多个, 如图 1 所示, 仿真测试就是对这 50 000 多个样本评估其中的概念漂移现象.

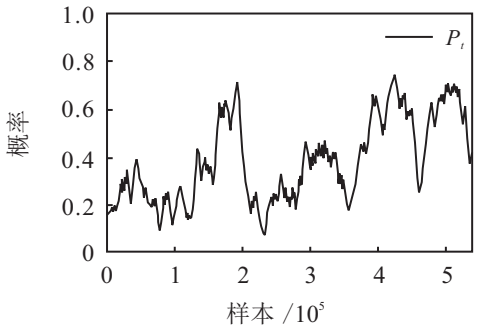
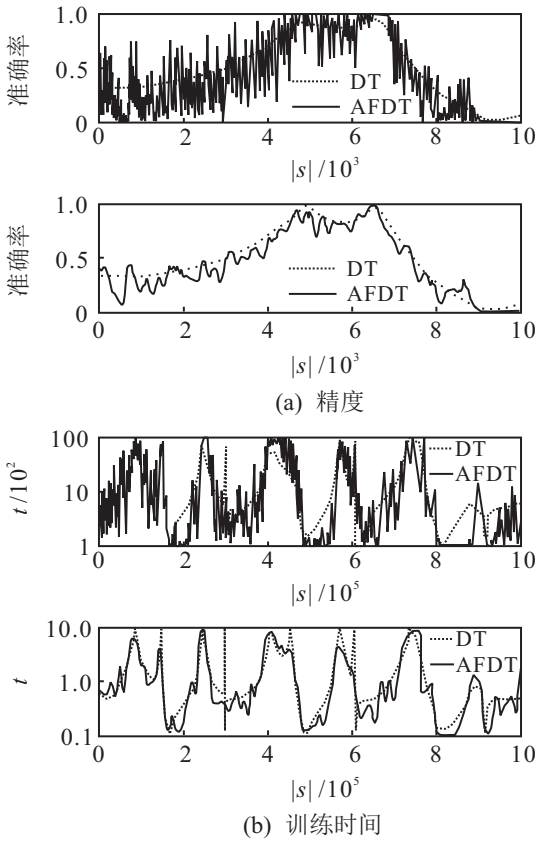


图 1 “森林覆盖”型不确定数据流

同时, “森林覆盖”型数据序列分属于 7 个等级的 581 012 个样例, 每个样例由 54 个属性组成, 包括 10 个数值变量, 4 个二进制自然保护区型变量, 40 个二进制土壤类型变量, 前 15 120 个样例被作为测试样例, 剩余的 565 892 个样例用来建立测试序列[12]. 为了测评“森林覆盖”型数据序列中样例的概念漂移类型, 采用自适应快速决策树模型进行仿真检验.

3.2 数据处理

为了验证自适应快速决策树算法 (AFDT) 的有效性, 在仿真实验中, 比较自适应快速决策树算法与一般决策树算法 (DT)[14]对不确定数据流的处理能力, 两种决策树算法均能对不确定数据流进行分类, 也可以对概念漂移现象进行检测. 根据模型参数范围和仿真需要, 现设 $\tau = 0.50$, $\omega = 50\%$. 在实验中, 为了测试 AFDT 和 DT 对不确定数值属性的拆分能力, 在每个测试节点, 将“森林覆盖”型数据序列的前 5 000 个样本用作参照样例, 其余为训练样例, 检测的准确率和训练时间如图 2 所示. 在图 2 中, 横轴表示输入不确定数据流的大小.



(a) 精度

(b) 训练时间

图 2　AFDT 与 DT 算法处理不确定数据流的能力

从图 2(a) 中可以发现: AFDT 比 DT 更早更快地实现了分类, 这是因为 AFDT 在节点上使训练样例能够快速地实现数值属性拆分, 而 DT 需要进行多次采样扫描. 从图 2(b) 中可以发现: AFDT 分类速度大大快于 DT, 究其原因, DT 采用多遍重复方法来构建节点, 以确保分类的准确性; 而 AFDT 采用将过时的样例从决策树中删除, 并同时从滑动窗口中删除. 此外, AFDT 会自动检查决策树内部节点分割的有效性, 被弃选的分裂属性将不会再次被选中, 从而使分类结果更准确.

为了进一步研究自适应快速决策树对不确定数据流的分类能力, 还需要研究置信度阈值参数 μ 如何影响其分类性能. 将置信度阈值参数 μ 的值设置为 0%, 10%, 20%, 30%, 40%, 以每 10 000 个样本组成数据流组, 近 60 000 个样本构成训练数据流, 其平均分类结果如图 3 所示.

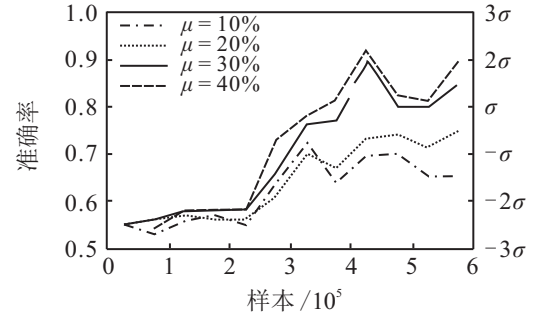
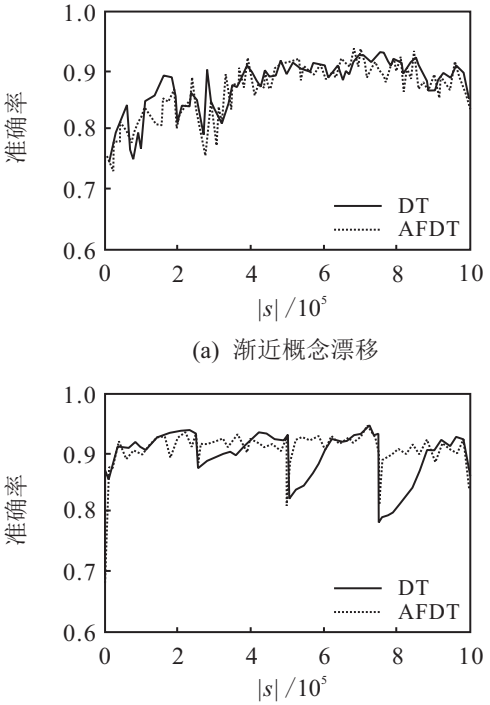


图 3　不同阈值下自适应快速决策树的分类能力

从图 3 可以看出: 整个数据流的数值属性被均匀地离散在中心 $\pm 3\sigma$ 内, 不确定数据流数值属性的平均分类精度为 50%, 其分类属性在 50%～90% 之间呈波浪式上升变化; 同时, 当对不确定数值属性的拆分能力提高时, 不确定分类属性的准确率上升较快; 当不确定数据流的不确定数值属性拆分率在 50%～60% 之间时, 相关的分类准确率在 5% 的幅度内波动. 从图 3 还可以看出, 随着参数值 μ 的取值不断增加, 自适应快速决策树的分类精度保持不断上升趋势, 即 μ 增加, 分类精度将大幅度提高.

3.3　仿真结果分析

AFDT 算法与 DT 算法对不确定数据流的分类效率不同, 其对目标概念漂移程度的检测能力也有差异. 现设定模型参数分别取 $k = 0$ 和 $k = 1$, "森林覆盖"型不确定数据流中的概念漂移检测采用 $\sigma_{ik}^2 = 5\%$ 增量方式进行滑动窗口模拟, 训练样本分类属性被均匀地离散为 5 组, 以每 10 000 个样本为训练组, 运行 $\sigma_{ik}^2 = 5\%$ 滑动窗口, 分别采用 AFDT 算法和 DT 算法来检测"森林覆盖"型不确定数据流中目标概念漂移的程度, 检测结果如图 4 所示.



(a) 渐近概念漂移

(b) 突概念漂移

图 4　不确定数据流概念漂移的分类

在图 4(a)中: 以 $\sigma_{ik}^2 = 5\%$ 滑动窗口进行检测模拟, AFDT 算法和 DT 算法检测结果显示"森林覆盖"型不确定数据流出现了缓概念漂移现象; 在图 4(b) 中, 以 $\sigma_{ik}^2 = 10\%$ 滑动窗口进行检测模拟, 其结果为"森林覆盖"型不确定数据流出现了突概念漂移现象. 尽管二者检测到的目标概念变化的轨迹大致

相同, 但DT算法明显对不确定数据流中目标概念变化的检测分类没有AFDT算法敏感, 且其收敛速度没有AFDT算法快. 因此, AFDT算法在检测不确定数据流概念漂移方面具有更好的适应能力.

4 结 论

大多数数据流分类算法需要大量精确标记属性输入的支持才能够进行分类, 然而, 不确定数据流分类和概念漂移现象检测的本身包含着对无标记属性的挖掘和收集. 由于数据流中无标记属性样本的数量很大, 使得对数据流中无标记属性样本难以挖掘和收集, 从而导致对不确定数据流概念漂移现象的检测和分类出现很多困难. 基于此, 本文采用将无标记属性拆分为不确定数值属性并将其转化为不确定分类属性的规则来构建自适应快速决策树算法, 因其基于无标记情景学习拆分原理, 具有很强的对无标记不确定数据流概念漂移的处理能力. 该算法适用于大量无标记不确定数据流概念漂移的检测和分类, 本文的实验仿真结果也验证了该方法的有效性和可靠性.

参考文献(References)

[1] Pan S, Wu K, Zhang Y, et al. Classifier ensemble for uncertain data stream classification[C]. Proc of the 14th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Boston: Harvard Business School Press, 2010: 488-495.

[2] Gao C, Wang J. Direct mining of discriminative patterns for classifying uncertain data[C]. Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Minding. New York: Free Press, 2010: 861-870.

[3] Qin B, Xia Y, Wang S, et al. A novel bayesian classification for uncertain data[J]. Knowledge-Based System, 2011, 24(7): 1151-1158.

[4] Aggarwal C, Yu P S. A framework for clustering uncertain data streams[C]. Proc of the 24th IEEE Int Conf on Data Engineering. Cancún: Cancún Press, 2008: 150-159.

[5] Pang S, Ban T, Kadobayashi Y, et al. Personalized mode transductive spanning svm classification tree[J]. Information Sciences, 2011, 181(4): 2071-2085.

[6] Shaker A, Senge R, Hllermeier E. Evolving fuzzy pattern trees for binary classification on data streams[J]. Inforamtion Sciences, 2012, 19(2): 34-51.

[7] 肖丹萍, 叶东毅. 基于免疫原理的不确定数据流聚类算法[J]. 模式识别与人工智能, 2012, 25(5): 826-834.
(Xiao D P, Ye D Y. Clustering uncertain data streams based on immune principle[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(5): 826-834.)

[8] 邢长征, 温培. 基于网格密度和引力的不确定数据流聚类算法[J]. 计算机应用研究, 2015, 32(1): 98-101.
(Xing C Z, Wen P. Uncertain data streams clustering algorithm based on grid density and force[J]. Application Research of Computers, 2015, 32(1): 98-101.)

[9] 刘三民, 孙知信. 具有概念漂移的P2P网络流量识别研究[J]. 系统工程与电子技术, 2013, 35(4): 864-869.
(Liu S M, Sun Z X. Research of traffic identification in P2P network with concept drift[J]. Systems Engineering and Electronics, 2013, 35(4): 864-869.)

[10] Salton G, Fox E A, Wu H. Extended Boolean information retrieval[J]. Commun ACM, 1983, 26(11): 1022-1036.

[11] Nandedkar A V, Biswas P K. A granular reflex fuzzy min-max neural network for classification[J]. IEEE Trans on Neural Networks, 2009, 20(7): 1117-1134.

[12] 张杰, 赵峰. 基于基序及其时序关系的耦合流数据分类算法[J]. 情报学报, 2013, 26(2): 51-56.
(Zhang J, Zhao F. Classification algorithm of coupled stream data based on motifs and their temporal relations[J]. J of the China Society for Scientific and Technical Information, 2013, 26(2): 51-56.)

[13] Leite D, Ballini R, Costa P, et al. Evolving fuzzy granular modeling from non-stationary fuzzy data streams[J]. Evolving Systems, 2012, 3(1): 65-79.

[14] 张杰, 赵峰. 流数据概念漂移的检测算法[J]. 控制与决策, 2013, 28(1): 29-35.
(Zhang J, Zhao F. Detecting algorithm of concept drift from stream data[J]. Control and Decision, 2013, 28(1): 29-35.)

(责任编辑：闫 妍)