

大数据环境下的不确定数据流在线分类算法

吕艳霞,王翠荣,王 聪,于长永

(东北大学 信息科学与工程学院,辽宁 沈阳 110819)

摘 要:在大数据环境下,由于隐私保护、数据丢失等原因,数据普遍存在不确定性,数据流系统中数据不断地到达系统,只扫描一遍且不能一次性全部获得,所以要构建一个增量分类模型来处理不确定数据流分类。本文基于 VFDT 算法提出了 WBVFDTu 算法,该算法在学习和分类阶段都可快速而有效地分析不确定信息。在学习期间,采用 Hoeffding 分解定理构造决策树模型;在分类期间,在决策树的叶子节点利用加权贝叶斯分类算法提高模型的分类准确率和算法的执行效率。最终证明该算法能够非常快速地学习不确定数据流,提高分类的准确率。

关 键 词:不确定数据流;加权贝叶斯;VFDT;分类算法;大数据

中图分类号: TP 311 文献标志码: A 文章编号: 1005-3026(2016)09-1245-05

Online Classification Algorithm for Uncertain Data Stream in Big Data

LYU Yan-xia, WANG Cui-rong, WANG Cong, YU Chang-yong

(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: LYU Yan-xia, E-mail: shaoqilyx@163.com)

Abstract: Under the background of big data, there exist data uncertainties due to privacy protection, data loss and so on. In data stream system, data arrive at continuously and cannot be obtained all. In addition, all the information cannot be acquired with only one scan. Therefore, an incremental classification model is constructed to deal with uncertain data stream classification. The weighted Bayes based on VFDT (very fast decision tree) for uncertain data stream—WBVFDTu on the basis of VFDT algorithm is presented in the paper. The uncertain information can be analysed quickly and effectively in both the learning stage and classification stage. In the learning stage, a decision tree model for uncertain data stream is quickly constructed by using Hoeffding bound theory. In the classification stage, the weighted Bayes classifier in the tree leaves is used to improve the performance of the classification. Experimental results show that the proposed algorithm can very quickly learn uncertain data stream and improve the classification performance of the model.

Key words: uncertain data stream; weighted Bayes; VFDT (very fast decision tree); classification algorithm; big data

数据流模型在各领域都有着广泛的应用,如物联网、金融、互联网等。随着技术的进步,人们发现这些领域的数据由于重复测量、隐私的保护以及数据丢失等原因,数据普遍存在不确定性。数据的不确定性导致数据项的值不能使用单值表示,而用多值以及相对应的概率分布表示^[1]。传

统的数据流分类算法假设流数据的值是精确和确定的^[2],而数据的不确定性对分类研究有着重要的作用,合理地利用概率分布所包含的不确定信息,并不是简单地使用概率分布期望值,而是显著地提高分类准确率^[1]。

VFDT(very fast decision tree)^[2]是数据流分

类的经典算法之一,本文受 VFDT 和 UDT (uncertain decision tree)^[1]的启发,针对大数据环境下不确定数据流在线分类问题,研究如何利用数据的不确定信息在学习和分类阶段对流数据进行在线分类学习.

1 相关工作

1.1 不确定分类

数据挖掘领域有一个重要的研究方向是关于不确定数据的.在经典模型的基础上提出了几个不确定分类算法,文献[1]采用小数采样的技术来处理不确定性并提出 UDT 算法;文献[3-4]提出适用于不确定数据的基于规则的分类算法;文献[5]采用极限学习机对不确定数据进行分类,提出 UU-ELM 和 NU-ELM 算法来分别处理均匀分布和不均匀分布的不确定数据集.以上这些算法都是用来学习静态数据集的.在文献[6]中,提出了一个积极的没有标签的学习环境下的不确定性 CVFDT 算法 puuCVFDT,该算法侧重于对具有积极和未标记的样本的不确定数据流进行二分类学习.本文研究了同时具有类别属性和数值属性的数据流的分类问题,为学习一个准确的不确定快速决策树,提出了一个新的模型,该模型在学习和分类过程中可以同时利用不确定信息.

1.2 数据流分类

数据流分类有两种主要的方法,基于单一分类器的方法和集成分类器的方法.基于单一分类器的方法中最著名的就是 VFDT 和 CVFDT, CVFDT 提高了 VFDT 处理概念漂移的能力.文献[7]采用支持向量机作为训练器,基于样本不确定性值更新分类器,先测试后训练使得算法开始时候的准确率较低且只能处理确定的数据.文献[8]提出了两种对不确定数据进行挖掘的集成分类器算法,静态集成分类器(SCE)和动态集成分类器(DCE).然而,文中假设样本的类别是不确定的,而属性值是精确的.本文算法是在类别精确的前提下,认为属性值是不确定的.

1.3 Hoeffding 树

针对流数据建立决策树,Hoeffding 算法从数据流的最先到达的一部分采样数据开始选择一个属性作为决策树的根节点,然后从根节点开始分裂.一旦作为根节点的属性确定,之后到来的数据根据根节点的属性值分别落到对应的叶子节点,然后被用来作为叶子节点继续分裂的依据,该算

法对后续到来的数据递归执行这一过程.树的叶子节点只需存放关于这些样本属性值的充分统计量,充分统计量包含足以计算启发式度量值的统计信息.为选择最佳分裂属性,在这些统计信息上执行如下基于 Hoeffding 边界理论^[9]的 Hoeffding 测试.

假设实值 a 是一个随机变量,如果给出 m 个 a 的值,则 Hoeffding 边界理论会以 $1-\eta$ 的概率保证实际均值与观察均值之差的绝对值小于 ε ,其中 $\varepsilon = \sqrt{(R^2 \ln(1/\eta))/2m}$,然后计算充分统计量,用来确定叶子节点是否继续分裂以及分裂的属性.给出启发函数 $H(X_i)$,如果用 $H(X_i)$ 来计算信息增益,则 $H(X_i)$ 的范围为 $R = \ln C$,其中 C 表示分类数.当数据流中的样本有 m 个落到决策树的某个叶子节点时,本文假设 X_1 和 X_2 分别为样本所有属性中具有最高和次高启发式度量值的属性,令 $\Delta \bar{H} = H(X_1) - H(X_2)$ 为一个新的随机变量.给定一个期望值 η ,如果满足 $\Delta \bar{H} > \varepsilon$,那么 Hoeffding 边界理论以概率 $1-\eta$ 确保 X_1 是最佳分裂属性.

2 不确定数据流环境下的分类算法

2.1 问题定义

不确定性既可以出现在数值型属性值上,也可以出现在名词性属性值上. $A^u = \{A_1^u, A_2^u, \dots, A_k^u\}$ 代表不确定属性集,其中 A_i^u 代表 A^u 的第 i 个不确定属性, A_{it}^u 代表第 t 个采样上不确定属性 A_i^u 的属性值, k 表示不确定属性的个数, $i \in [1, k]$.同文献[1],不确定属性值 A_{it}^u 包含一个取值范围以及该范围上的概率分布.如果 A_i^u 是数值型属性,其取值范围用 $[a_{it}, b_{it}]$ 来表示,其中 $a_{it}, b_{it} \in \mathbf{R}$,概率分布则由一个概率密度函数 $g_{it}(x)$ 来表示,且满足 $\int_{b_{it}}^{a_{it}} g_{it}(x) dx = 1$;如果 A_i^u 是名词性属性,其取值范围定义为 $\text{Ran}(A_i^u) = \{r_1, \dots, r_m\}$,概率分布则由一个向量 $P = \{p_{i1}, \dots, p_{im}\}$ 来表示,其中 $\text{Pr}(A_i^u = r_i) = p_{ij}$ 且有 $\sum_{j=1}^m p_{ij} = 1$.

在大数据环境下,不确定数据流就是一系列不断到达的不确定数据样本序列,用 $D^u = \{D_1^u, D_2^u, \dots, D_t^u, \dots\}$ 表示,其中 D_t^u 表示一个不确定数据样本,每个不确定数据样本包含一个属性向量 A^u 和一个类别 y^u ,即 $D_t^u = (A^u, y^u)$,其中 $y^u \in C^u = \{C_1^u, C_2^u, \dots, C_{|C|}^u\}$ 表示样本 D_t^u 所属的类别.本文旨在针对不确定数据流 D^u 构造分类器,对

后续到来的样本 $D_i^u = (A^u, y^u = ?)$ 给出正确分类. 传统的针对静态不确定数据集的分类算法, 是一次获得全部训练数据, 学习出一个分类模型, 根据该模型对后续的未知数据进行测试. 在大数据环境下, 不确定数据流系统中数据源源不断地到达系统, 数据不可能一次性全部获得, 而且只允许扫描一遍, 所以本文要构建一个增量分类模型即增量决策树模型, 并且使用该模型将不确定属性 A^u 转换为一个类别概率分布 $\{\text{Pr}(C_1^u), \dots, \text{Pr}(C_{|C|}^u)\}$. 这样无论何时, 根据模型预测后续的样本 D_i^u 所属类别为

$$y^u = \arg \max_{c=1, \dots, |C|} \{\text{Pr}(C_c^u)\}.$$

2.2 WBVFDTu 算法

在 VFDT 算法的基础上构造了 WBVFDTu 算法 (weighted Bayes based very fast decision tree for uncertain data stream). 该决策树的生长过程如算法 1 所示.

算法 1 不确定非常快速决策树算法

WBVFDTu Stream(UT, $G, D_i^u, \eta, \sigma, n_{\min}$)

输入:

D_i^u 不确定数据流到达的一个样本

Gain(\cdot) 启发式度量函数, 计算节点属性分裂

η 1 减去该节点最佳分裂属性的期望概率值

σ 自定义的一个阈值

n_{\min} 每个 Hoeffding 测试需要的最小样本数

输出: UT 用于不确定数据分类的决策树模型

1 Let R be the root of UT.

2 If R is not a leaf, then

3 : Split D_i^u into m fractional items $\{D_{i1}^u, D_{i2}^u, \dots, D_{im}^u\}$

4 : For each item $D_{ij}^u \in \{D_{i1}^u, D_{i2}^u, \dots, D_{im}^u\}$

5 : Let R_i be the j -th branch child for D_{ij}^u .

6 : WBVFDTuStream(HT, $G, D_i^u, \eta, \sigma, n_{\min}$).

7 Else

8 Collect sufficient statistics from item D_i^u

9 Let n_1, n_2 be the expected count of items last seen and current seen at R

10 If items seen so far at R are not all of the same class and $n_2 - n_1 > n_{\min}$, then

11 Estimate Gain(A_i^u) on sufficient statistics

12 Let A_a^u and A_b^u be the attribute with first and second highest G , respectively

13 Let $\varepsilon = \sqrt{R^2 \ln(1/\eta) / (2\text{PC}(D_N^u))}$

14 Let $\Delta \bar{G} = \text{Gain}(A_a^u) - \text{Gain}(A_b^u)$

15 If $\Delta \bar{G} > \varepsilon$ or $\Delta \bar{G} \leq \varepsilon < \sigma$, then

16 : Replace R by an internal node that splits on A_a^u

17 : For each branch of the split

18 : Add a new leaf L_j

19 : Initiate sufficient statistics of leaf L_j

20 return UT.

算法中 D_i^u 代表新到达的样本, 函数 Gain(\cdot) 用来确定哪一个属性作为节点的分裂属性, 本文选用不确定信息增益 (uncertain information gain, UIG)^[3] 作为在 Hoeffding 测试时所使用的启发式度量函数. 本文用 D_N^u 表示在叶节点 N 观察到的样本集, 假设属性 A_i^u 是被选出的分裂属性, 它将样本集 D_N^u 分裂成 m 个子样本 $\{D_{i1}^u, D_{i2}^u, \dots, D_{im}^u\}$, 则 UIG 可通过如下公式计算得到:

$$\text{UIG}(D_N^u, A_i^u) = \text{uEntropy}(D_N^u) -$$

$$\sum_{i=1}^m \frac{\text{PC}(D_{ij}^u)}{\text{PC}(D_N^u)} \times \text{uEntropy}(D_{ij}^u).$$

其中: PC(D) 表示样本集 D 的概率基数; uEntropy(D) 表示期望信息熵. η 的值设置为 1 减去任一节点上最佳分裂属性的期望概率值; σ 为用户自定义的一个阈值参数, 用来决定当前节点的分裂与否; n_{\min} 表示在一个叶节点上进行基于 Hoeffding 测试的分类的最小样本数.

对后续到来的不确定数据样本, 模型从根节点开始最终传递到叶子节点. 通常情况下, 分类算法会将该样本的类别标记为叶子节点上先验概率最大的类别. 而在叶节点上采用不同的分类器将会提高分类的性能, 比如使用朴素贝叶斯分类器, 朴素贝叶斯分类器具有很好的增量式学习能力, 然而它只能学习确定数据, 其他的改进针对不确定数据的贝叶斯分类模型^[10], 只是针对批处理数据, 不具有增量学习能力, 因此不适合数据流环境. 使用 WBVFDTu 决策树模型, 在叶节点采用不确定加权贝叶斯分类策略来更新决策树. 同时, 本文也给出多数分类 (majority class) 的分类策略用来进行比较^[11-12].

多数分类策略使用最大概率值对不确定样本的属性进行分类, 即返回概率最大的分类:

$$\arg \max (\text{Pr}(C_1^u), \text{Pr}(C_2^u), \dots, \text{Pr}(C_{|C|}^u)) = \arg \max (f(D_i^u, R)).$$

其中: R 表示决策树的根, $f(D_i^u, R)$ 表示不确定属性集映射为类概率分布的映射函数. 针对本文的决策树模型, 递归地定义映射函数, 增量地构造 WBVFDTu 决策树的过程中, 新到达的样本从决策树根节点开始, 根据其属性的取值不断被分割, 到达各个内部节点 N , 最终到达相应的叶子节点. 在内部节点 N 上, 函数可定义为

$$f(D_i^u, N) = \sum_{j=1}^m f(D_{ij}^u, NT_j).$$

其中 D_{ij}^u 代表第 j 个分割后的样本, NT_j 代表内部节点 N 的第 j 棵子树. 到叶节点 L 上函数定义为

$$f(D_i^u, L) = \{Pr(C_1^u), \dots, Pr(C_{|C|}^u)\}.$$

这样, 后续样本 D_i^u 所对应的概率分布可以通过映射函数 $f(D_i^u, R)$ 计算得到.

本文算法包含两个阶段, 首先是 WBVFDTu 决策树的构建, 此时不确定数据样本信息增量存储到叶节点所维护的充分统计量中, 只有在满足 Hoeffding 测试结果的情况下最终分裂. 其次是确定样本所属的分类, 与多数分类策略相比, 返回概率最大的分类的公式是相同的, 不同的是该分类策略在叶节点 L 上所定义的映射函数:

$$\begin{aligned} f(D_i^u, L) &= \frac{w_i^u Pr(C_k^u)}{Pr(A^u)} \left\{ \prod_{i=1}^k Pr(A_{it}^u | C_1^u), \dots, \prod_{i=1}^k Pr(A_{it}^u | C_{|C|}^u) \right\} \\ &\propto w_i^u Pr(C_k^u) \left\{ \prod_{i=1}^k Pr(A_{it}^u | C_1^u), \dots, \prod_{i=1}^k Pr(A_{it}^u | C_{|C|}^u) \right\}. \end{aligned}$$

w_i^u 为所有样本的权值, 该权值根据系统的应用由专家给出, 令 d_i^u 为叶节点所维护样本集, $Pr(C_k^u)$ 由 $Pr(C_k^u) = PC(d_i^u, C_k^u) / PC(d_i^u)$ 得到, 对于 $Pr(A_{it}^u | C_k^u)$ 要根据属性为连续和离散的情况分别计算得到.

2.3 充分统计量

在决策树的构建过程中, 由于时空效率的限制, 叶节点只存储每一个子样本的 pdf 值的充分统计量. 对名词性属性, 在每个叶节点 L , 让每个属性 $A_i^u \in A^u$ 的可能值 A_{it}^u 和每个 $C_i^u \in C^u$ 都对应一个充分统计量 s_{ijk} , 初值为 0. WBVFDTu 算法只扫描一遍到达叶子节点 L 上的子样本集, 来维护 s_{ijk} 的值. 即一旦叶子节点 L 有新样本 D_i^u 到达, 充分统计量 s_{ijk} 就更新为

$$\begin{aligned} s_{ijk} &= s_{ijk} + w_i^u \times Pr(\mathcal{C}(D_i^u) = C_k^u) \times \\ &\quad Pr(A_{it}^u = A_{it}^u). \end{aligned}$$

其中, $\mathcal{C}(D_i^u)$ 为样本 D_i^u 的类别. 所有到达叶子节点 L 的样本集在该节点上的数值型属性 A_i^u 的所有取值形成一个总体的 pdf $g_{ik}^u(a)$, 通过对不确定数据流进行单遍扫描, 计算 $g_{ik}^u(a)$. 给出参数 \sum_{ik} 为节点 L 上已经到达的样本权值之和, 这样 \sum_{ik} 和 $g_{ik}^u(a)$ 就是叶子节点所需要维护的充分统计量. 针对数值型属性, 这两个值的计算要求获得所有该类型属性的全局概率分布函数. 在大多数实际应用中, 不确定信息一般采用高斯分布 pdf 来描述^[13]. 本文采用高斯逼近 (Gaussian

approximation) 来描述不确定数据流中数值型属性值的分布情况.

3 实验分析

算法使用 MOA 平台和 Weka 软件包, Java 语言编程实现. 实验数据集来自于 UCI 数据库. 本文采用的数据集为 Waveform(版本 1 和版本 2)和 LED. Waveform1 包含 21 个数值型属性, Waveform2 包含 40 个数值型属性, 二者都有 3 个类别选项代表三种类型的波. LED 数据集包含 7 个布尔属性和 10 个类别选项, 该数据集的最优贝叶斯分类准确率为 74%.

实验对比了算法在使用多数分类策略 (WBVFDT - MC) 和加权贝叶斯分类策略 (WBVFDT) 时在不同数据集上的分类准确性, 并将其与 VFDT 和 UDT 进行比较, 以此证明本文所提出的 WBVFDTu 在处理不确定数据流的在线分类问题时具有更高的准确率.

在数据集 Waveform1, Waveform2 和 LED 上算法执行结果分别如图 1 ~ 图 3 所示. 可以看出, 整个学习过程中, WBVFDT - MC 的准确率接近于 VFDT, 而且在样本输入数目不多的情况下, WBVFDTu 的准确率仍然与 UDT 相差无几, 需要强调的是 UDT 是针对不确定静态数据的批处理学习模型, 而本文设计的是针对不确定动态流数据的学习模型. 随着样本数量增加, WBVFDT - MC 的准确率逐渐接近 UDT. 然而, WBVFDTu 的准确率始终超越 WBVFDT - MC 和 VFDT, 并且与 UDT 接近且与目前为止公开确认的最优贝叶斯分类准确率相等, 准确率为 74%. 实验观察到无论输入的不确定数据样本数量如何变化, WBVFDTu 算法的分类性能在整个数据流样本上几乎都没有变化, 也就是说在任何时刻, WBVFDTu 都可以获得较好的分类准确率.

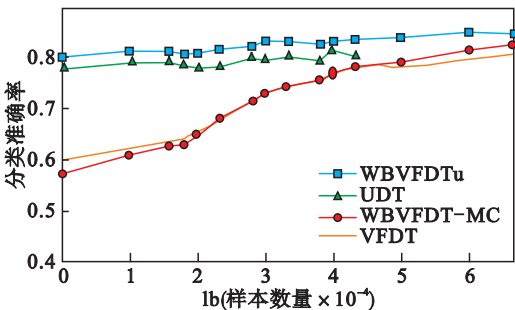


图 1 Waveform1 分类准确率的比较
Fig. 1 Classification accuracy comparisons in waveform 1

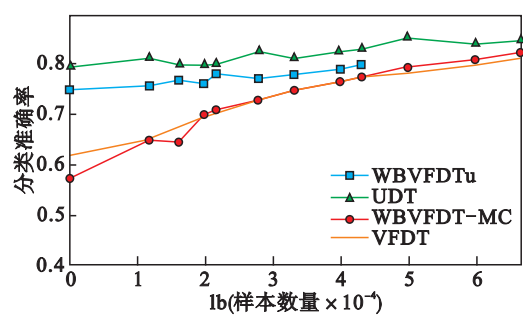


图 2 Waveform2 分类准确率的比较
Fig. 2 Classification accuracy comparisons in waveform 2

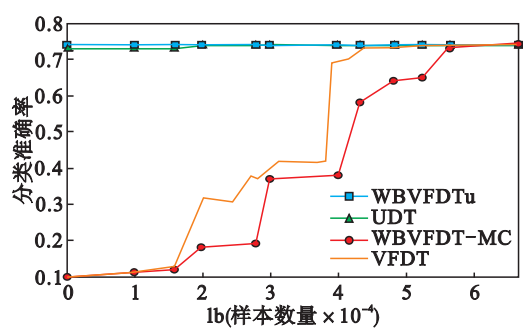


图 3 LED 分类准确率的比较
Fig. 3 Classification accuracy comparisons in LED

4 结 论

本文提出 WBVFDTu 算法,通过采用决策树的学习方法对不确定数据流进行分类.该算法通过计算不确定数据流中样本的充分统计量,并根据它来快速构造决策树模型.同时扩展了朴素贝叶斯分类模型,创建了处理不确定数据的不确定加权贝叶斯分类模型,在 WBVFDTu 算法创建的决策树叶子节点的数据集上采用该分类模型,作为叶子节点属性分裂策略.实验结果表明, WBVFDTu 在处理不确定数据流在线分类问题时能够非常快速地学习出决策树模型,并且在叶子节点上采用不确定加权贝叶斯分类方法,使 WBVFDTu 算法在处理不确定数据流分类时准确率有所提高.

参考文献：

[1] Tsang S ,Kao B ,Yip K Y ,et al. Decision trees for uncertain data[J]. *Knowledge &Data Engineering IEEE Transactions* , 2009 23(1) :64 – 78.

[2] Hulten G ,Spencer L ,Domingos P. Mining time changing data streams[C]// *Process of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S. l.] :ACM 2001 97 – 106.

[3] Qin B ,Xia Y ,Li F. DTU :a decision tree for uncertain data [J]. *Advances in Knowledge Discovery and Data Mining* , 2009 5476 :4 – 15.

[4] Gao C ,Wang J. Direct mining of discriminative patterns for classifying uncertain data[C]// *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S. l.] :ACM 2010 861 – 870.

[5] Cao K Y ,Wang G ,Han D. An algorithm for classification over uncertain data based on extreme learning machine[J]. *Neurocomputing* 2016 174 :194 – 202.

[6] Liang C ,Zhang Y ,Shi P ,et al. Learning very fast decision tree from uncertain data streams with positive and unlabeled samples [J]. *Information Sciences* , 2012 , 213 (23) : 50 – 67.

[7] 刘三民 ,孙知信 ,刘涛. 基于样本不确定性的增量式数据流分类研究 [J]. *小型微型计算机系统* 2015(2) :193 – 196.
(Liu San-min ,Sun Zhi-xin ,Liu Tao. Research of incremental data stream classification based on sample uncertainty[J]. *Journal of Chinese Computer Systems* 2015(2) :193 – 196.)

[8] Pan S ,Wu K ,Zhang Y ,et al. Classifier ensemble for uncertain data stream classification[J]. *Lecture Notes in Computer Science* 2010 6118(1) :488 – 495.

[9] Hoeffding W. Probability inequalities for sums of bounded random variables[J]. *Journal of the American Statistical Association* ,1962 58(301) :13 – 30.

[10] He J ,Zhang Y ,Shi X L P. Learning naive Bayes classifiers from positive and unlabelled examples with uncertainty[J]. *International Journal of Systems Science* ,2012 ,43(10) : 1805 – 1825.

[11] Liang C ,Zhang Y ,Hu P S Z. Learning accurate very fast decision trees from uncertain data streams[J]. *International Journal of Systems Science* 2015 46(16) :3032 – 3050.

[12] 卢惠林. 基于加权 Bayes 分类器的流数据在线分类算法研究 [J]. *计算机科学* 2014 41(5) :227 – 229.
(Lu Hui-lin. Weighted Bayes based data streaming online classification algorithm[J]. *Computer Science* 2014 41(5) : 227 – 229.)

[13] Aggarwal C C ,Yu P S. A survey of uncertain data algorithms and applications[J]. *IEEE Transactions on Knowledge & Data Engineering* 2009 21(5) :609 – 623.