基于C4.5和NB混合模型的数据流分类算法

李 燕 张玉红 胡学钢
(合肥工业大学计算机与信息学院 合肥 230009)

摘 要 具有概念漂移的含噪数据流的分类问题成为数据流挖掘领域研究的热点之一。提出了一种基于C4.5和Naïve Bayes混合模型的数据流分类算法CDSMM。它以C4.5作为基分类器,采用朴素贝叶斯分类器过滤噪音,同时引入假设检验中的μ检验方法检测概念漂移,动态更新模型。实验结果表明,CDSMM算法在处理带有噪音的概念漂移数据流时具有比同类算法更好的分类正确率。

关键词 数据流,概念漂移,分类,噪音
中图法分类号 TP181 **文献标识码** A

Classification Algorithm for Data Stream Based on Mixture Models of C4.5 and NB

LI Yan ZHANG Yu-hong HU Xue-gang
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract Classification on the noisy data stream with concept drifts has recently become one of the most popular topics in streaming data mining. Classification algorithm for mining Data Streams based on Mixture Models of C4.5 and NB was proposed, called CDSMM, in which decision trees based on C4.5 are selected as the basic classifiers and the classifier of Naïve Bayes is adopted to filter noise data. Meanwhile, it introduces the μ-hypothesis testing method to detect concept drifts. Extensive studies demonstrate that CDSMM is superior to several existing algorithms in the predictive accuracy when handling noisy data streams with concept drifts.

Keywords Data streams, Concept drifts, Classification, Noise

现实世界的许多应用领域,如电信、网络、超市交易等正在以惊人的速度产生大量的数据流,其中蕴含着丰富的信息。由于数据流具有快速性、无限性和实时性的特点[1,2],使得传统的方法难以有效地进行挖掘;且数据流中隐含的知识或概念可能会随着时间的推移而发生变化,即概念漂移。因此,如何正确有效地检测概念漂移并适应漂移变化已成为近年来数据挖掘领域中的一个热点和难点。

在数据流分类方面,目前已有不少学者提出了一些针对概念漂移的挖掘模型和算法,其中包括基于规则的算法[3,6]、决策树算法[7]、基于实例的方法[8,9]和集成学习[10,13,14]方法。然而,这些方法在处理概念漂移数据流时,难以适应数据流环境,尤其是含噪音较多的情况。事实上,一个好的算法不仅需要满足耗时少、准确度高,又要在噪音环境中对概念漂移具有敏感性、迅速收敛到当前概念的特点。因此,如何在含有噪音的数据流中处理概念漂移,仍是一个亟待解决的问题。

针对上述问题,本文提出了一种基于C4.5决策树和Naïve Bayes混合模型的数据流分类算法CDSMM(A Classification algorithm for mining Data Streams based on Mixture Models of C4.5 and NB)。它以C4.5为基分类器建立集成模型,利用朴素贝叶斯分类器降低噪音影响,同时依据假设检验中μ检验的方法,判定概念漂移,并动态地调整集成分类器。在若干概念漂移基准数据集上的实验结果表明:与同类算法相比,CDSMM算法具有较强的抗噪性,即使在含噪率为30%的数据集上仍具有较高的分类正确率。

本文第2节介绍相关工作;第3节详细介绍所提出的混合模型的数据流分类算法及理论分析;第4节分析实验结果;最后是小结与展望。

1 相关工作

针对数据流的分类问题,目前已提出很多模型和算法。如Domingos等于2000年提出了基于Hoeffding不等式决策树的数据流分类算法VFDT[3],通过不断地将叶节点替换为分支节点而生成。然而它仅能处理离散属性。Oza等于2001年提出了Boosting和Bagging的在线版本[4,5],使得Boosting和Bagging算法可以处理流数据。Gama等于2003年提出VFDT的改进算法VFDTc[6],将算法扩展到可以处理连续属性,同时在叶节点引入了贝叶斯分类器。

以上算法均假设数据流是稳定分布的。而在现实情况中,数据流往往存在概念漂移的现象。因此,研究者开始关注数据流概念漂移的问题。如2001年Hulten等提出了基于单

到稿日期:2010-01-29 返修日期:2010-04-12 本文受国家973重点基础研究发展计划(2009CB326203),国家自然科学基金课题(60975034),安徽省自然科学基金课题(090412044)资助。

李 燕(1985-),女,硕士生,主要研究方向为数据挖掘、数据流,E-mail:liyan0918@gmail.com;张玉红(1979-),女,博士生,讲师,主要研究方向为数据挖掘、数据流等;胡学钢(1961-),教授,博士生导师,主要研究方向为人工智能、数据挖掘。

决策树模型的概念漂移发现算法 CVFDT[7]，采用窗口机制为树中的每个节点创建替换子树，周期性地检测节点对应的概念是否发生漂移，用替换子树取代漂移节点以适应当前数据分布。同年，Street 等提出了 SEA 算法[8]，采用集成决策树方法来处理数据流中的概念变化。2003 年，Wang 等提出了 Weighted-bagging 算法[9]，讨论了多分类器中的权值变化和裁减问题，同时根据分类器的分类错误率动态改变权值。Kolter 和 Maloof 等于 2007 年提出了一种集成方法检测概念漂移[10]，随着分类器性能的变化创建新的分类器和删除过时的分类器模型从而进行更新。2008 年，Zhang 等人提出了一个新的最优权值调整方法 OWA[11]，利用最近的数据块检测分类器最佳的权值，并使用了一个核心均值匹配方法，以尽量减少数据块之间在内核空间的差异。2009 年，Bifet 等人提出了两种改进的在线 Bagging 算法[12]：ADWIN Bagging 和 ADHT Bagging 算法。ADWIN Bagging 是在 Oza 和 Rusell 提出的在线 bagging 方法上增加了 ADWIN(变窗口)机制，以检测变化；而 ASHT 使用可调整规模的 Hoeffding 树，并根据每棵树分类错误率平方的反比为分类器设定权值。

近年来，概念漂移数据流中的噪音问题越来越受到广泛的关注。Chu 等于 2004 年提出了一种基于 EM 框架的判别模型[13]处理噪音数据流，它建立集成加权分类器来最大程度地适应流数据。Wang 等于 2006 年提出了一种基于聚类的方法[14]，过滤数据流中的困难数据和噪音数据。同年，Gama 等将 VFDTc 扩展到处理适当带有噪音的概念漂移数据流[15]。Li 等于 2008 年提出了一种增量式的 MSRT 算法[16]，根据 Hoeffding 边界不等式建立集成半随机决策树，处理噪音数据流。

然而，以上提及的算法均采用单一类型的分类器，并且难以有效降低噪音对带有概念漂移的数据流分类的影响。因此，针对噪音环境中的概念漂移数据流分类问题，本文提出了基于 C4.5 分类器和 NB 分类器组成的混合模型。实验表明，所提出的方法比单一分类器构成的集成模型具有更好的分类准确率。

2 CDSMM：基于 C4.5 和 NB 混合模型的数据流分类算法

2.1 算法描述

本文所提出的数据流分类算法 CDSMM，由多个 C4.5 分类器和一个 Naïve Bayes 分类器组成集成模型。它以 C4.5 基分类器在当前数据块上的分类正确率为权值，以错误率作为衡量指标，采用假设检验-μ 方法检测漂移，动态地调整基分类器，以适应概念变化。同时，利用基于统计的 Naïve Bayes 分类器过滤噪音，利用被误分类实例为主体构建新的基分类器，替代正确率较低的分类器。本文方法能有效增强发生漂移的实例对新分类器的影响，使其快速适应变化后的概念。

首先对 CDSMM 算法中涉及的符号进行说明：S 表示当前数据块，E_i 表示 S 中的第 i 条记录，EC 表示集成分类器，K 描述集成分类器的容量，num 表示 EC 中当前的分类器个数，C_i 表示 EC 中的第 i 个分类器，d 表示基分类器对应的训练集的大小。

算法流程如下。1)构建集成分类器：当 S 中的记录个数达到 d 后，若 $num<K$，则用 S 构建一个新的分类器 C_{num}，以 C_{num} 在 S 上的正确率为权值，将 C_{num} 加入 EC 中；否则，用窗口中最近的 K 个数据块，共 $K*d$ 个数据构建一个朴素贝叶斯分类器 C'；2)噪音过滤：用 EC 对 S 中的每一个实例进行分类，若实例同时被 EC 和 C' 误分类，则将其加入误分类缓冲区中；3)漂移检测：计算 EC 在 S 上分类错误率的均值 err，并利用此值检测概念漂移；4)分类器调整更新：一旦检测出漂移，且集成模型中分类器个数达到 K，将误分类缓冲区的数据读入 S，构建一个新的分类器 C_{new}，以其在 S 上的正确率为权值(weight_new)，同时更新 EC 中所有分类器的权值。若新分类器 C_{new} 的权值 weight_new 大于 EC 中权值最小的分类器 C_k 的权值，则丢弃 C_k 分类器，将 C_{new} 加入 EC 中组成新的集成模型。

需要说明的是，集成分类器 EC 采用投票机制进行分类预测，即若一个实例被 EC 中半数以上的分类器误分类，则认为其为误分类实例。

Input：集成分类器 EC=Null；数据流 DS；分类器容量 K；当前数据块的大小 d；

Output：训练好的集成分类器 EC

Procedure of CDSMM：

While(新数据到来){
  读入 d 条数据形成当前数据块 S；
  if(num $<$ K)
    用 S 训练一个分类器 C_{num} 并加入 EC 中，num++；
  else
    用当前的 $K*d$ 个实例训练一个朴素贝叶斯分类器 C'；
  for($E_j \in S$){
    if(E_j 被 EC 误分类 && E_j 被 C' 误分类)
      将 E_i 放入临时误分类缓冲区 $ErrInst$ 中；
  }
  用 S 更新 EC 中分类器的权值，并计算分类器分类错误率的均值，记为 err；
  令 $U=\dfrac{(err-\mu_0)}{\sqrt{err(1-err)/n}}$；
  if($U>\mu_a$ && $num == K$){ //进行漂移检测
    用 $ErrInst$ 和部分新的实例训练一个分类器 C_{new}，并计算其权值 $weight\_new$；
    用 $ErrInst$ 更新 EC 中所有分类器的权值，找出其中权值最小的分类器，记 C_k；
    if($weight\_new > weight_k$)
      用 C_{new} 替换 C_k；
  }
}

2.2 处理机制

在上述数据流分类算法流程中，对算法的准确率起决定因素的是两大机制：漂移检测和噪音过滤。

2.2.1 漂移检测

本文使用假设检验中大子样检验母体平均数-μ 检验的原理来进行漂移检测。它的基本思想是认为小概率事件在一次试验中是几乎不可能发生的。若发生了，则有理由怀疑这一假设的真实性。

定义 1(μ 检验) 设有母体 X，它的分布是任意的，且一、二阶矩存在。记 $EX=\mu$，$DX=\sigma 2$(σ 未知)。在母体上做假设 $H_0:\mu\leqslant\mu_0$(μ_0 已知)，$H_1:\mu>\mu_0$。用 $\bar{X}$ 做检验，当 n 很大时，

统计量 $U=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$（$\overline{X}$ 为子样平均数，S 为子样标准差）近似地服从标准正态分布 $N(0,1)$。给定显著水平 α，存在 μ_α 使 $P\{U>\mu_\alpha\}\approx\alpha$。

单个样本被误分类的概率为 $P\{X=1\}=p$，正确分类的概率为 $P\{X=0\}=1-p$，则 X 服从两点分布 $B(1,p)$。设有 n 个样本，其中有 m 个被误分类，则 $\overline{X}=m/n$，$S^2=\overline{X}(1-\overline{X})$。定义中的统计量可写为 $U=\frac{\overline{X}-\mu_0}{\sqrt{\overline{X}(1-\overline{X})/n}}$。

CDSMM 使用上述原理检测概念漂移，对当前数据块上的分类错误率 err 进行假设检验，μ_0 是一个常量，初始化为前 K 个数据块的分类错误率的均值。因此当统计量 $U=\frac{err-\mu_0}{\sqrt{err(1-err)/n}}\geqslant\mu_\alpha$ 时，错误率出现较大程度的变化，从而说明发生了概念漂移；反之认为概念分布平稳。

2.2.2 噪音过滤

为降低噪音对数据流概念漂移检测的影响，本文引入朴素贝叶斯模型进行噪音过滤。原因在于：基于朴素贝叶斯的算法简单、运行速度快、抗噪性强、分类精确度高，被广泛应用于文本分类与信息检索领域。利用其统计特性，能够发现数据中的异常点，可有效达到过滤噪音数据的效果[17]。

朴素贝叶斯的主要思想描述如下：设每个数据样本用一个 n 维特征向量来描述 n 个属性的值，即 $X=\{x_1,x_2,\cdots,x_n\}$，假定有 m 个类，分别用 $C_1,C_2,\cdots,C_m$ 表示。给定一个未知的数据样本 X（即没有类标号），X 属于类 $C_k(k=1,2,\cdots,n)$，当且仅当 $P(C_k|X)>P(C_j|X)$，$1\leqslant j\leqslant m$，贝叶斯定理可以表示为

$$P(C_k|X)=\frac{P(X|C_k)P(C_k)}{P(X)},(k=1,2,\cdots,m)$$

由于 $P(X)$ 对于所有类为常数，则最大化后验概率 $P(C_k|X)$ 可转化为最大化先验概率 $P(X|C_k)P(C_k)$。朴素贝叶斯分类器假定各个特征项之间是相互独立的，则

$$P(X|C_k)P(C_k)=P(x_1,x_2,\cdots,x_n|C_k)=\prod_{j=1}^{n}P(x_j|C_k)P(C_k)$$

算法根据每个类别所占的比例求得 $P(C_k)$，然后根据每个类别中各个属性值所占的比例求得 $P(x_i|C_k)$。若是连续属性，则先对属性值进行离散化再求解。

基于以上计算方法，在所设计的 CDSMM 算法中，假设混合模型中共有 $K+1$ 个分类器，其中 K 个为 C4.5 分类器、1 个为朴素贝叶斯分类器。在投票时，若一个实例被 $K/2$ 以上的 C4.5 分类器和朴素贝叶斯分类器同时误分类，则认为它是具有潜在概念漂移的实例，并放入误分类实例缓冲区；否则认为它可能是噪音，将其丢弃，从而减少噪音对概念漂移检测的影响。

2.3 算法分析

已有理论和实验证明，在处理存在概念漂移的数据流数据时，使用集成分类器方法比仅使用单个分类器的方法具有更好的适应性和精确性[9]。另外，在概念漂移检测方面，CDSMM 以在数据块 S 上的错误率为指标，将 C_i 在 S 上的精度作为基分类器的权值，一方面能有效地评判该基分类器的分类效果，另一方面在发生概念漂移时，减小了过时历史数据对分类器的负面影响。此外，收集误分类实例并添加新实例，建立新分类器的方法使算法能快速地适应新概念。而在噪音处理方面，若实例被误分类，则可能是数据流中概念发生了变化，也可能是噪音的影响导致的分类误差。采用朴素贝叶斯分类器能有效过滤误分类实例，大大降低了噪音数据的干扰。

复杂度分析：假设在大小为 d 的数据块 S 上建立一个分类器的复杂度为 $f(d)$，数据块 S 被单个分类器分类的复杂度为线性的。假设数据流由 n 个数据块组成，则算法 CDSMM 的复杂度为 $O(n*f(d)*p+K*n*d)$，其中 $0<p<1$，表示 n 个数据块中需要重建分类器的数据块的比例，K 表示基分类器的个数。与经典算法 weighted-bagging 相比，CDSMM 并不是对每个数据块均建一个基分类器，仅在误分类数据达到一定阈值时才重建一个基分类器，从而一定程度上减少了构建分类器的复杂度。

3 实验与性能分析

为验证算法对数据流概念漂移与噪音处理的有效性，本文选取 4 个具有代表性的数据流集成算法 Weighted-bagging[9]，OzaBag[4]，OzaBoost[5] 和 Bag10-ASHT-W+R[12] 在不同的数据集上与 CDSMM 进行对比。实验中所使用的 HyperPlane，Waveform-Drift，LED-Drift 及 SEA 等概念漂移基准数据集均由开源资源 MOA[14] 中相应的数据生成器生成。HyperPlane 数据是处理渐进式数据流概念漂移的基准数据集；Waveform-Drift 数据集和 LED-Drift 数据集则是在原有的 Waveform 和 LED 上加入了概念漂移的数据集；SEA 数据集是突变式概念漂移数据集，具有三维属性，包含 4 个概念。

本文实验所在的环境是：Intel Pentium 双核 1.6Hz，2G 内存的 PC 机；操作系统是 Windows XP；开发环境为 Weka 3.6平台，编译运行环境是 jdk1.6。

3.1 参数分析

本文进行了大量的实验，其结果表明，当设置数据块大小 $d=500$，集成分类器中基分类器的个数 $K=8$，误分类缓冲区实例个数的阈值 $e=100$，μ 检验中的 $\alpha=0.95$ 时，算法具有较优的分类效果。

另外，关于贝叶斯分类器个数选择：为了提升算法的抗噪性，在算法中添加朴素贝叶斯分类器进行噪音的过滤。设计了 3 种策略：①在缓冲区中仅保存一个数据块，建立一个贝叶斯分类器，每个数据块到来后重建该分类器；②在缓冲区中保存多个数据块，但只建立一个贝叶斯分类器，利用窗口机制保存最新的数据块并更新分类器；③在缓冲区中保存多个数据块，每个数据块都建立一个贝叶斯分类器，利用窗口机制保存最新的几个分类器。为了选取更有效的噪音过滤方法，分别采用了这 3 种策略进行大量实验。实验结果表明，第二种策略，即用多个数据块建立一个贝叶斯分类器的方法效果最佳。分析原因如下：若使用第一种策略，虽每次使用的是最新的数据，但仅是一个数据块，信息量较小。而用贝叶斯过滤噪音的本质是统计的方法，数据较小导致准确率不高。若使用第三种策略，虽保存了多个数据块有大量历史信息，而每个分类器的数据量较少，且分类器之间相似度较大，因而难以达到理想的效果。故本文使用了第二种策略。

3.2 实验分析

3.2.1 概念漂移检测

实验选用含噪率在 0%～30% 的 HyperPlane 数据集，考

察在不同噪音环境下 CDSMM 算法的概念漂移检测情况。图 1 描述了对 100k 条数据进行训练，每 5k 条数据进行测试的曲线图。其中横坐标是数据块的下标，纵坐标是分类精度。从图得知，当数据中无噪音时，准确度基本保持在 90% 至 95%之间；而随着噪音的增加，准确度逐渐下降，且波动变大。分析原因：噪音率越高，数据集中含有真实类别的数据越少，在 CDSMM 构建 C4.5 基分类器时影响了分裂属性的选择，且局部数据可能出现概念漂移，所以准确度不高；当检测到概念漂移后，集成分类器根据当前的数据分布进行调整，且朴素贝叶斯分类器过滤了噪音，同时新建了拟合当前概念分布的分类器，从而提高了分类精度。可以说，CDSMM 算法在一定的噪音环境中能够检测到局部概念漂移并根据当前概念能有效更新模型。

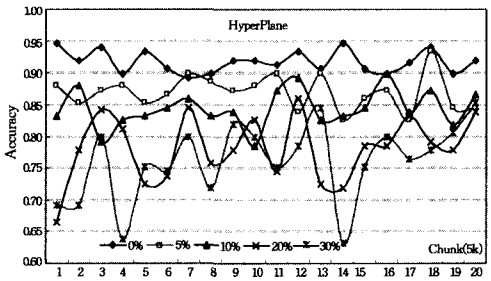

图 1　CDSMM 在 HyperPlane 上的漂移检测

3.2.2　朴素贝叶斯分类器的噪音过滤效果

表 1 列出了 CDSMM 算法使用朴素贝叶斯分类器前后在分类精度上的对比。其中"——"表示未使朴素贝叶斯分类器，"NB"表示使用了朴素贝叶斯分类器。所用数据集为含噪率从 0%到 30%变化的 HyperPlane 数据集，训练集大小分别为 50k，100k，200k，300k，测试集大小为 100k。由表 1 知，加入朴素贝叶斯分类器后，在不含噪音的数据集及含噪率为 5%的数据集上精度略有提高；在含噪率为 10%～20%的数据集上，精度提高 1%～3%；而当含噪率为 30%时，精度可提高 2.8%～4.3%。由此可见，贝叶斯分类器在含噪率较大的数据集上噪音过滤的效果较为显著。

表 1　使用朴素贝叶斯分类器前后在分类精度上的对比

	0		5		10		20		30	
	——	NB	——	NB	——	NB	——	NB	——	NB
50	97.5	98.5	94.9	95.4	92.9	94.8	90.0	92.2	88.9	91.8
100	97.9	98.2	95.8	94.9	92.2	95.5	90.9	93.3	88.6	91.5
200	97.2	97.4	94.0	94.8	92.1	93.9	91.1	92.2	86.8	91.2
300	97.6	96.6	93.8	94.6	93.2	94.7	91.3	94.2	88.9	92.7

3.2.3　CDSMM 与其他算法的性能对比

图 2 描述了 CDSMM 与 Weighted-bagging，OzaBag，OzaBoost 和 Bag10-ASHT-W＋R 算法在 HyperPlane 数据集上的对比实验结果。训练集大小为 200k，测试集大小为 100k，噪音率分别为 0%，5%，10%，20%，30%。由图可知，在不同的噪音率上，CDSMM 算法的分类准确率比 Weighted-bagging 提高了 8%～36%，比 OzaBag 和 OzaBoost 提高了 24%，比 Bag10-ASHT-W＋R 提高了 2%～27%，且所含噪音越多提升幅度越大。甚至当噪音率为 30%时，分类精度反而超过了噪音率为 20%的情况。这是由于所采用的基分类器是 weka 中提供的 C4.5 分类器，已经过剪枝和优化，也在一定程度上增强了抗噪性。

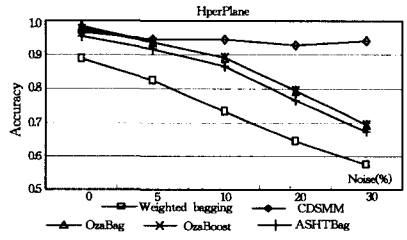

图 2　HyperPlane 数据集上检测情况

图 3 是一组算法在 Waveform-Drift，LED-Drift 和 SEA 4 个数据集上的实验对比结果。Waveform-Drift 数据集具有 21 维属性，其中 20 维属性带有漂移；SEA 数据的 4 个概念分别对应 4 个测试集。以下实验均采用含噪率为 10%、训练集大小为 200k、测试集大小为 100k 的数据集。由图可知，在 Waveform-Drift 数据库中，CDSMM 算法的正确率比 OzaBag，OzaBoost 和 Weighted-bagging 提高了 12%，比 Bag10-ASHT-W＋R 提高了 10%；在 SEA 数据库的 4 个测试集上，CDSMM 算法的正确率比其余几种算法也具有显著的优势，提高幅度在 1%到 19%之间。由此可见，CDSMM 算法在分类正确率上都明显优于其余 4 种算法。然而在 LED-Drift 数据库中，CDSMM 算法的正确率却比其余 4 种算法分别降低了 3.6%，4%，3.7%，3.7%。分析原因如下：Waveform-Drift，HyperPlane 和 SEA 数据均为连续型数据，而 LED-Drift 数据为离散型数据，CDSMM 在实验中采用的是已经过剪枝和优化的 C4.5 决策树，并对树的高度进行限制，从而导致该算法在离散数据上的分类正确率不如同类算法。

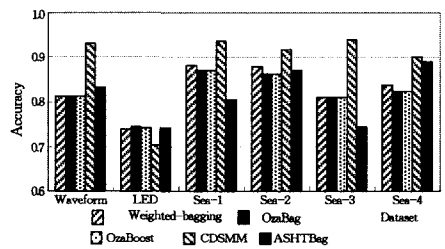

图 3　与其他方法在多个数据集上的比较

结束语　针对数据流中的噪音与概念漂移问题，本文提出一种基于 C4.5 和 Naïve Bayes 的混合集成分类算法 CDSMM。它采用假设检验的方法判定概念漂移，利用贝叶斯分类器去除误分类实例中的噪音，并及时地更新基分类器适应概念漂移。研究表明，与 Weighted-baggging 等基于单一模型的集成方法相比，CDSMM 的分类正确率显著提高，且具有较强的抗噪性。然而，如何克服算法在离散数据集上的劣势，以及探索其它分类器在混合模型中的作用，将是未来研究的重点。

参 考 文 献

[1]　Widmer G，Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts[J]. Machine Learning，1996，23(1)：69-101

[2]　Schlimmer J C，Granger R H. Incremental leaning from noisy data[J]. Machine Learning，1986，1(3)：317-354

[3]　Domingos P，Hulten G. Mining High-speed Data Streams[C]// Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston，MA，USA：

ACM Press,2000:71-80

[4] Oza N C,Russell S. Online bagging and boosting[C]// Proc of Artificial Intelligence and Statistics. Morgan Kaufmann. San Francisco,2001:105-112

[5] Oza N C, Russell S. Experimental comparisons of online and batch versions of bagging and boosting[C]// Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California: ACM Press, 2001: 359-364

[6] Gama J, Rocha R, Medas P. Accurate decision trees for mining high-speed data streams[C]// Proc. of the 9th ACM SIGKDD International Conference in Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 523-528

[7] Hulten G, Spencer L, Domingos P. Mining Time-changing Data Streams[C]// Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM Press, 2001: 97-106

[8] Street W, Kim Y. A *streaming ensemble algorithm* (SEA) for large-scale classification[C]// Proc. of the 7th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM Press, 2001: 377-382

[9] Wang Haixun, Fan Wei, Yu P S, e al. Mining concept-drifting data streams using ensemble classifiers[C]// Proc. of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington: ACM Press, 2003: 226-235

[10] Kolter J Z, Maloof M A. Dynamic weighted majority: an ensemble method for drifting concepts[J]. Machine Learning Research, 2007, 8: 2755-2790

[11] Zhang P, Zhu X Q, Shi Y. Categorizing and mining concept drifting data streams[C]// Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, Nevada, USA: ACM Press, 2008: 812-820

[12] Bifet A, Holmes G, Pfahringer B, et al. New ensemble methods for evolving data streams[C]// Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 139-148

[13] Chu Fang, Wang Yizhou, Zaniolo C. An adaptive learning approach for noisy data streams[C]// Proc. of 4th IEEE International Conference on Data Mining. Brighton: IEEE Computer Science, 2004: 351-354

[14] Wang Yong, Li Zhanhuai, Zhang Yang. Classifying noisy data streams[C]// Proc. of 3rd International Conference of Fuzzy Systems and Knowledge Discovery (FSKD). Xi'an, China: Springer, 2006: 549-558

[15] Gama J, Fernandes R, Rocha R. Decision trees for mining data streams[J]. *Intelligent Data Analysis*, 2006, 10: 23-45

[16] Li Peipei, Hu Xuegang, Wu Xindong. Mining concept-drifting data streams with multiple semi-random decision trees[C]// Proc. of 4th International Conference on Advanced Data Mining and Applications. Chengdu, China: Springer, 2008: 733-740

[17] Tan Pang-ning, Steinbach M. 数据挖掘导论[M]. 范明, 等译. 北京: 人民邮电出版社, 2006

(上接第 133 页)

键属性个数越少,ADVR 算法的优越性就越大。其主要原因是关键属性个数越少,DVR 算法构造的查询语句中存在冗余的可能性就越大,即 ADVR 算法移除冗余的可能性就越大,因此其性能就越好。

结束语 通过对 DVR 算法分析,得知 DVR 算法没有考虑用户提交的 SQL 语句和细粒度访问控制策略的特性,使得最终生成的查询语句中存在冗余,从而影响了查询语句的执行性能。本文给出了两类冗余,并提出了移除冗余的方法。最终提出了 ADVR 算法,该算法移除了 DVR 算法所产生的冗余。通过实验得知,ADVR 算法优于 DVR 算法。

参 考 文 献

[1] Agrawal R, Kiernan J, Srikant R, et al. Hippocratic Database [C]// Proceedings of VLDB. Hong Kong, China, 2002: 563-574

[2] Agrawal R, Bird P, Grandison T, et al. Extending relational database systems to automatically enforce privacy policies [C]// Proceedings of ICDE 2005. Tokyo, Japan: IEEE, 2005: 1013-1022

[3] Bertino E, Sandbu R. Database Security-Concepts, Approaches, and Challenges [J]. IEEE Transactions on Dependable and Secure Computing, 2005, 2(1): 2-19

[4] Byun J W, Bertino E, Li N. Purpose Based Access Control of Complex Data for Privacy Protection [C]// Proceedings of SACMAT 2005. Stockholm, Sweden: ACM, 2005: 102-110

[5] Dwivedi S, Menezes B, Singh A. Database Access Control for E-business-A Case Study [C]// Proceedings of Int. Conf. on Management of Data. Bombay, India: IEEE, 2005: 168-175

[6] Rizvi S, Mendelzon A, Sudarshan S, et al. Extending Query Rewriting Techniques for Fine-grained Access Control [C]// SIGMOD 2004. Paris, France: ACM, 2004: 551-562

[7] Chaudhuri S, Dutta T, Sudarshan S. Fine Grained Authorization Through Predicated Grants [C]// Proceedings of ICDE 2007. Istanbul, Turkey: ACM, 2007: 1174-1183

[8] Wang Q, Yu T, Li N, et al. On the Correctness Criteria of Fine-grained Access Control in Relational Databases [C]// Proceedings of VLDB 2007. Vienna, Austria, 2007: 555-566

[9] Stonebraker M, Wong E. Access control in a relational database *management system by query modification* [C]// *Proceedings of* the 1974 ACM/CSC-ER. 1974: 180-186

[10] Corporation O. Oracle Virtual Private Database [EB/OL]. http://www.oralce.com/technology/deploy/security/db_security/virtual-private-database/index.html

[11] LeFevre K, Agrawal R, Ercegovac R, et al. Limiting disclosure in Hippocratic databases [C]// Proceedings of VLDB 2004. Roronto, Canada, 2004108-119

[12] Zhu H, Shi J, Wang Y, et al. Controlling Information Leakage of Fine-grained Access Control Model in DBMSs [C]// Proceedings of WAIM 2008. Zhangjiajie, China: IEEE, 2008: 583-590

[13] Shi J, Zhu H, Fu G, et al. On the Soundness Property for SQL Queries of Fine-grained Access Control in DBMs [C]// Proceedings of ICIS/ACIS 2009. Shanghai, China: IEEE, 2009: 469-474

[14] DeWitt D J. The Wisconsin benchmark: Past, present, and future [M]. The Benchmark Handbook, 1993