一种基于混合集成方法的数据流概念漂移检测方法

桂 林 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 近年来,数据流分类问题研究受到了普遍关注,而漂移检测是其中一个重要的研究问题。已有的分类模型有单一集成模型和混合模型,其漂移检测机制多基于理想的分布假设。单一模型集成可能导致分类误差扩大,噪音环境下分类效果受到了一定影响,而混合集成模型多存在分类精度和时间性能难以两者兼顾的问题。为此,基于简单的WE集成框架,构建了基于决策树和bayes混合模型的集成分类方法WE-DTB,并利用典型的概念漂移检测机制Hoeffding Bounds和μ检验来进行数据流环境下概念漂移的检测和分类。大量实验表明,WE-DTB能够有效检测概念漂移且具有较好的分类精度及时空性能。

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

Data Stream Concept Drift Detection Method Based on Mixture Ensemble Method

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

Abstract Mining with data stream concept drift is a hot topic in data mining. Existing classification approaches consist of ensemble method based on single base classifiers and ensemble method based on hybrid base classifiers, which depend on the stationary assumption and learnable assumption. However, the former probably causes the larger classification deviation and the performance on accuracy is impacted in the noisy data streams, while the latter performs worse on the classification accuracy or the time consumption. Motivated by this, an ensembling classification method WE-DTB was proposed, based on hybrid based models with decision trees and Naive Bayes. It is an extended framework of WE model. Meanwhile, we utilized the popular concept drift detection mechanisms based on Hoeffding Bounds and μ test to implement the detection on concept drifts. Extensive experiments demonstrate that our proposed method WE-DTB can detect concept drift effectively while maintaining the good performance on classification accuracy and consumptions on time and space.

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,已成为数据挖掘领域中的一项热点和难点。

近年来,数据流分类最流行的方法是集成学习方法。因集成学习保存了一个概念描述的集合,被广泛用于概念漂移的处理。集成分类器的优点在于它能将所有分类器的信息联合起来。王等[3]证明采用集成分类器方法比采用单一分类器方法具有更好的效果。在基于"稳定分布假设"的条件下,证明WE集成模型[3-5]具有很高的精度。"稳定分布假设"是数据平稳分布、训练数据块和测试数据块的分布相同或者相似,但是在真实的数据流环境中概念会发生持续的变化,这种基于"稳定分布假设"下的模型并不能适应新的概念,所以WE集成模型分类效果并不理想。为此,高等[6]于2007年提出了基于"可学习假设"的集成模型AP,"可学习假设"允许数据流中的概念随机地变化。尽管集成模型AP比"稳定性假设"有更宽松的条件,但是它没有考虑到噪音数据块存在的情况。如果当前数据块是一个噪音数据块,则在这个噪声数据块上建立分类器将会比使用其他缓存块更加糟糕。实际上,"可学习假设"仍然不是一个对数据流的真实描述。由于以上两种假设的不足,张等[7]提出了一个"噪声情况的假设",即在真实的数据流中,可能同时含有概念漂移和噪音。

基于这种假设,进一步提出了一种基于混合集成方法的数据流概念漂移检测方法WE-DTB。该模型以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。在SEA,HyperPlane概念漂移基准数据集和KDDCUP'99[11]入侵检测数据集上结合概念漂移处理机制进行试验。本文采

到稿日期:2011-01-27 返修日期:2011-04-27 本文受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。
桂 林(1986-),男,硕士生,主要研究领域为数据挖掘、数据流,E-mail:1986_520@sina.com;张玉红(1979-),女,博士,讲师,主要研究领域为数据挖掘、数据流等;胡学钢(1961-),教授,博士生导师,主要研究领域为人工智能、数据挖掘。

用典型的 Hoeffding Bounds[8] 和 μ 检验[9] 来进行数据流环境下概念漂移的检测和分类。实验结果表明，与其它集成模型相比，WE-DTB 能够有效检测概念漂移且具有较好的分类精度及时空性能。

本文第 2 节介绍相关工作；第 3 节详细介绍所提出的混合集成方法及理论分析；第 4 节分析实验结果；最后是小结与展望。

1 相关工作

数据流的连续性、快速性、无限性和多变化等，特别是概念漂移的特点，使得传统挖掘算法难以及时有效地对其处理。因此，研究者们提出了一系列用于数据流挖掘的算法。其中，使用最多的是集成学习方法。以下是对集成学习的说明。

假设一个包含无限数据块 $D_i(i=-\infty,\cdots,+\infty)$ 的数据流 S。由于空间限制，仅能在缓冲区中存储最新的 n 个数据块，每个数据块包含确定数量的事例。假设当前时间戳观察到第 n 个数据块 D_n 以及被记为 $D_1,D_2,\cdots,D_n$ 的 n 个缓冲区中的数据块。为了预测下一个到达的数据块 D_{n+1}，选择一个或者多个学习算法学习缓冲区中的数据块，建立若干个基分类器 $f_i(i=1,2,\cdots,m)$，然后通过模型的平均机制形成一个集成分类器，对数据块 D_{n+1} 进行预测。

典型的构建方法有：水平集成框架 WE、垂直集成框架 AP、混合集成框架 AE。下面对这 3 种典型的集成框架作简单的介绍。

1.1 水平集成框架(WE)

图 1 是水平集成框架的图示说明。水平集成方法使用单一的分类算法学习缓冲区中的 n 个数据块，得到 n 个基分类器，构成集成模型。优点有两个：1)可以多次使用缓冲区中数据块所包含的信息，这对以后的测试数据块可能有用；2)最后的决策是由基于不同数据块训练的不同分类器共同决定的，具有一定的抗噪性。然而，水平集成框架是基于“稳定分布假设”提出的，它适用于数据分布相同或者相似的环境。但是，真实的数据流中是存在概念漂移的，概念会持续地发生变化，所以水平集成框架不适合对数据流进行概念漂移检测。

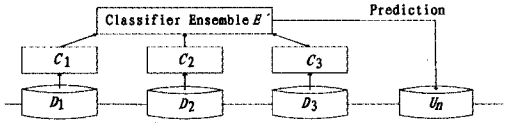

图 1 水平集成框架

1.2 垂直集成框架(AP)

图 2 给出垂直集成框架的图示说明。垂直集成框架是选择 m 个不同的学习算法 $L_j(j=1,2,\cdots,m)$ 在当前数据块 D_n 上建立 m 个不同的基分类器 $f_j=L_j(D_n)$，然后通过模型平均组合所有基分类器。当数据块未到达时，其先验知识是未知的，取最新到达的数据块上的模型平均可以在测试集上获得最小的期望错误。同时，使用不同的学习算法建立多个分类器，能降低期望错误偏差。尽管垂直集成方法比水平集成方法有更宽松的条件，即它允许数据流中的概念随机变化，可以应用到更多类型的数据流中，但是这个假设是没有考虑噪音的，而真实的数据流是含有噪音的。由于垂直集成的方法仅仅在当前的数据块上建立分类器，如果当前是一个噪音数据块，则结果可能非常糟糕。由于对噪音数据这一问题没有考虑周全，垂直集成框架被其自身限制在仅能处理概念漂移的情况，而不适用于在含噪音的数据流环境下处理概念漂移。

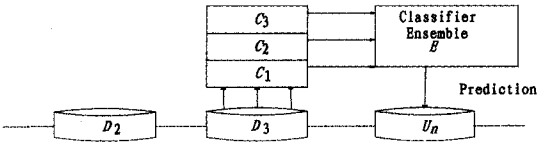

图 2 垂直集成框架

1.3 混合集成框架(AE)

图 3 是混合集成框架的图示说明。混合集成框架首先使用 m 个不同的学习算法 $L_i(i=1,2,\cdots,m)$ 在 n 个缓冲数据块 $D_j(j=1,2,\cdots,n)$ 上建立不同的分类器，训练 $m*n$ 个基分类器 $f_{ij}=L_i(D_j)$，其中 i 表示第 i 个算法，j 表示第 j 个数据块。把所有的这些基分类器组合起来，通过模型平均的方式组合成一个集成模型。它是水平集成模型与垂直集成模型的混合体，其中的基分类器组成了一个分类器矩阵 CM。

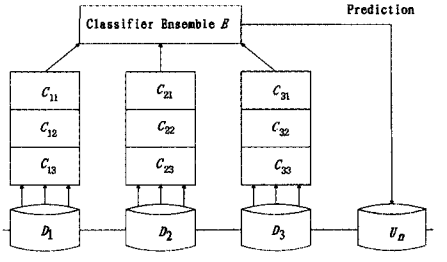

图 3 混合集成框架

$$CM=\begin{bmatrix} f_{11} & f_{12} & \cdots\cdots & f_{1n} \\ f_{21} & f_{22} & \cdots\cdots & f_{2n} \\ f_{31} & f_{32} & \cdots\cdots & f_{3n} \\ f_{41} & f_{42} & \cdots\cdots & f_{4n} \end{bmatrix}_{m*n}$$

CM 中的每个元素 f_{ij} 表示使用算法 i 在数据块 j 上建立的一个基分类器。如垂直分类器框架中的描述，CM 中每个列上的分类器(即在同一数据块上使用不同的算法建立的分类器)可以减少在一个未知的测试数据块上的期望分类错误偏差。CM 的每一行上的分类器(即在不同的数据块上使用相同的算法建立的不同分类器)可以消除噪音数据块对分类模型的影响。但是建立这样一个分类器矩阵 CM 需要耗费大量的时间，效率较低。

为此，基于“噪声情况的假设”提出了混合集成方法 WE-DTB，用于数据流中的概念漂移检测。

2 WE-DTB：基于混合集成方法的数据流概念漂移检测方法

在概念漂移数据流的处理中，面临的两大难题就是如何有效地检测概念漂移和快速地适应以及如何降低噪音数据的影响。虽然已存在的一些算法能够部分或全部考虑这些问题，但是少有算法较好地实现了二者的兼顾，尤其在含噪音较多的情况下，其正确率大幅下降。针对这一问题，本节设计了一种基于混合集成方法的数据流概念漂移检测方法 WE-DTB，它能有效处理噪音环境中的概念漂移数据流分类问题。

由于数据流的快速性，这里使用 C4.5 和 Naïve Bayes 两个分类算法来构建混合集成方法。

2.1 WE-DTB 方法描述

图 4 是 WE-DTB 方法的图示说明。首先对 WE-DTB 方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 表示采用 C4.5 建立的集成分类器，K 描述

EC 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。

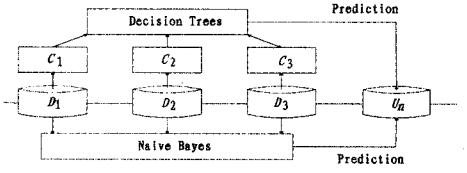

图 4　WE-DTB 方法

方法流程如下：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 分类 S 中的每一个实例，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 WE-DTB：
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} 并将其加入 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；
　利用 err 进行概念漂移检测；
　if(发生了漂移 && $num==K$){
　　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 S 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；
}
}

本模型以水平集成方法为基础，加入了一个朴素贝叶斯分类器。如图 4 所示，WE-DTB 方法首先采用 C4.5 分类算法学习每个在缓冲区中的数据块 D_i（$i=1,2,\cdots,n$），得到 n 个基分类器，然后将缓冲区中的数据块 D_i（$i=1,2,\cdots,n$）组成一个大的训练集 $D=\sum_{i=1}^{n} D_i$，用训练集 D 建立一个 Naïve Bayes 分类器，得到了由 n 个 C4.5 分类器和一个 Naïve Bayes 分类器组成的混合集成模型。

在对数据块 D_{n+1} 的预测中，每个事例 x 若被 $n/2$ 以上个数的 C4.5 分类器和 Naïve Bayes 分类器同时误分类，则该事例被误分类。模型对每个 C4.5 分类器都动态地加权，概念漂移发生后，用误分类缓冲区的数据构建新的 C4.5 分类器，来替换权值最小的一个 C4.5 分类器。

2.2　WE-DTB 方法分析

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，使用混合集成方法比仅使用单一模型的集成方法具有更好的适应性和精确性[7]。另外，在概念漂移检测方面，WE-DTB 方法以在数据块 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。

在含噪音的数据流环境下，假设混合模型 WE-DTB 中共有 $n+1$ 个分类器，其中 n 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 $n/2$ 个以上的 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。将误分类实例放入临时误分类缓冲区，收集误分类实例并添加新实例建立新分类器的方法能快速地适应新概念；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。

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 WE-DTB 方法的复杂度为 $O(n*f(d)*p+K*n*d)$。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。与基于同一假设的混合集成框架 AE 相比，WE-DTB 的分类器规模仅为 AE 的 $1/m$（m 为 AE 使用分类算法的个数），降低了构建分类器的复杂度。

3　实验与性能分析

为验证 WE-DTB 方法对数据流概念漂移处理的有效性，本文选取 3 个具有代表性的数据流集成方法 WE[3]，AP[6] 和 AE[7]，并结合典型的漂移检测指标 Hoeffding Bounds[8] 和 μ 检验[9] 在不同的数据集上与 WE-DTB 进行了对比。

3.1　数据集的选择

人工数据集：为了详细对比各种算法的性能，构造了两个人工数据集。

1）SEA：SEA 概念数据集的基本结构为 $\langle f_1, f_2, \cdots, C\rangle$，其中 f_1，f_2 是条件属性（取值 0～10），C 是决策属性（当样本属性满足条件 $f_1+f_2\leqslant\theta$ 时，属于第一类，否则属于第二类）。随机产生 110k 条样本（100k 的训练数据和 10k 的测试数据），包含 4 个 SEA 概念、3 次漂移。因此，它可以被用来作为评价标准容量、突变式的概念漂移挖掘的公共数据集。

2）超平面（Hyperplane）：一个 m 维的超平面上的样本的属性值 x_i 随机生成并且均匀分布在[0,1] 区间上。当产生的随机样本满足 $\sum_{i=1}^{m} a_i x_i \geqslant a_0$ 时，则将此样本标记为正样例；否则标记为负样例。产生 150k 条含噪率为 10% 的样本（100k 的训练数据和 50k 的测试数据），其中每 1k 的数据发生一次漂移。因此，它可以用来检测噪音环境下渐进式概念漂移挖掘。

真实数据集：为了评估真实情况下的算法性能，在 KDDCUP'99[10] 网络入侵数据上进行对比测试。该数据也是一个比较常用的数据流测试库。这个数据集包含了一系列 TCP 访问记录。这些记录主要分类两种：正常访问和 DOS 攻击。过去很多的实验结果已经表明，该数据集中的概念呈线性可分性(仅仅使用 10%的样本就可以达到 97%的预测精度)。这里采用 10%的样本，大小为 49w。

3.2 实验性能分析

为了验证本文所提方法在概念漂移数据流分类的有效性，以下实验中将分别从概念漂移检测、分类精度和时间性能这 3 个方面对比 WE-DTB 与其他经典的 3 种集成方法的优劣。

本文进行了大量的实验。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 $K=4$ 时，算法具有较优的分类效果。实验中 μ 检验的参数 $\alpha=0.95$。实验环境是：Intel Core 双核 2.93GHz，2G 内存的 PC 机；操作系统是 Windows XP；开发环境为 Weka3.6 平台，编译运行环境是 jdk1.6。

以下实验中，SEA-H 代表 SEA 数据集在 Hoeffding Bounds 概念漂移检测方法下进行的实验，SEA-μ 代表 SEA 数据集在 μ 检验概念漂移检测方法下进行的实验。

3.2.1 概念漂移检测

表 1 列出的是 4 种集成方法在 SEA，HyperPlane 和 KDDCUP'99 这 3 个数据集上漏报次数的实验对比结果。由表 1 知，WE-DTB 集成方法是具有较大的优势的。在 HyperPlane 数据集上，使用 Hoeffding Bounds 检测机制时，WE-DTB 方法比 WE 和 AP 少漏报 33 次，比 AE 少漏报 5 次，它仅只有 2 次漏报；使用 μ 检验检测机制时，WE-DTB 方法和 WE 持平，比 AP 少漏报 26 次，比 AE 少漏报 16 次，具有较好的表现。在 KDDCUP'99 数据集上，WE-DTB 方法比其它 3 种集成模型少漏报 1～2 次，也具有较好的效果。仅在 SEA 数据集上，WE-DTB 方法有 1 次的漏报，和另外 3 种模型表现相当。而且，使用 Hoeffding Bounds 或者 μ 检验检测机制对 WE-DTB 在概念漂移检测上的影响并不大。

表 1 漏报次数的对比

	WE	AP	AE	WE-DTB
SEA-H	1	0	1	1
SEA-μ	0	0	0	0
HyperPlane-H	35	35	7	2
HyperPlane-μ	1	27	17	1
KDDCUP'99-H	5	3	5	3
KDDCUP'99-μ	1	2	1	0

表 2 列出的是 4 种集成方法在误报率上的实验对比结果。表中分数的分母表示漂移检测的次数，分子表示误报的次数。从整体来看，使用 Hoeffding Bounds 概念漂移检测机制比使用 μ 检验的效果好。在 SEA 数据集上，WE-DTB 方法的误报次数比 AP 少 23 次，与 WE，WEAP 和 AE 基本持平。在 KDDCUP'99 数据集上，WE-DTB 与 WE 和 AE 表现相当，AP 效果最好。这是因为 AP 其实更适合处理动态的数据流，该种数据流假设只有概念漂移而没有噪音，KDDCUP'99 数据集正好符合这一假设，因此 AP 有更好的效果。在 HyperPlane 数据集上，使用 Hoeffding Bounds 检测机制时这 4 个模型在误报率上表现相当，只相差 1 次；使用 μ 检验概念漂移检测机制时 WE-DTB 效果较差。分析原因，μ 检验概念漂移检测机制是基于初始给定的一个平均错误率，而 HyperPlane 数据集是渐进式地进行概念漂移，在渐进的过程中分类器平均错误率上升，导致误报次数过大。

表 2 误报次数的对比

	WE	AP	AE	WE-DTB
SEA-H	0	0	0	0
SEA-μ	26/100	45/100	18/100	22/100
HyperPlane-H	0	0	0	1/200
HyperPlane-μ	0	0	0	32/200
KDDCUP'99-H	20/987	16/987	20/987	28/987
KDDCUP'99-μ	80/987	26/987	74/987	104/987

3.2.2 分类精度

表 3 列出的是分类精度的实验对比结果。从整体来看，WE-DTB 方法的效果较好。在 SEA 数据集上，这 4 种集成方法表现相当，WE-DTB 方法略有优势。在 HyperPlane 数据集上，WE-DTB 方法的优势较明显，分类精度比 WE 和 AE 高 7.7%～9.2%，比 AP 高 11.2%～15.9%。在含有噪音的 HyperPlane 数据集上的概念漂移检测，WE-DTB 方法减少了噪音的影响，较快适应了概念的变化。在网络入侵检测数据集 KDDCUP'99 数据集上，WE-DTB 的分类精度比 AP 略高，比 WE 高 6%～15.4%，比 AE 高 8.7%～12%，表现较好。而且，从分类精度上看，使用 Hoeffding Bounds 或者 μ 检验检测机制对 WE-DTB 在概念漂移检测上的影响微乎其微。

表 3 分类精度的对比

	WE	AP	AE	WE-DTB
SEA-H	99	97.96	98.55	99.3
SEA-μ	99.2	97.8	98.7	99.4
HyperPlane-H	82.2	75.5	83	91.4
HyperPlane-μ	82.8	79.3	82.2	90.5
KDDCUP'99-H	84.5	98.5	87.9	99.9
KDDCUP'99-μ	94	99	91.2	99.9

3.2.3 时间性能

表 4 列出的是 4 种集成方法在时间上的实验对比结果。在 SEA 和 HyperPlane 这种小数据集上，WE-DTB 的时间性能和 WE 表现相当，和 WE 只相差 1～5s，大概是 AP 和 WEAP 的 1/4，还不到 AE 集成框架的 1/10。在 KDDCUP'99 这个较大的数据集上，WE-DTB 方法的时间大概是 AE 的 1/4，比 AP 和 WEAP 效果略好(快 16～90s)，但是比 WE 慢 550～570s。WE-DTB 方法比 WE 慢是因为它加了一个朴素贝叶斯分类器，在分类器的个数上比 WE 要多一个，所以在时间性能上 WE-DTB 略差。可以说，在这 4 种集成方法中，WE-DTB 方法在时间性能上还是具有优势的。

表 4 时间性能的对比

	WE	AP	AE	WE-DTB
SEA-H	0.87	10	35	2.4
SEA-μ	1.4	18	37	2.4
HyperPlane-H	4.3	32	109	8
HyperPlane-μ	5.4	31	105	10
KDDCUP'99-H	16	611	2242	587
KDDCUP'99-μ	15	581	2391	565

以上实验分析研究表明：与其它 3 种集成方法相比，WE-DTB 方法在检测概念漂移、分类精度及时空性能上均有较好的表现。该方法的不足之处在于处理概念漂移时误报率较高。

(下转第 181 页)

表 3　实验二的统计结果

算法	最好解	最差解	平均值	最好解出现次数
模拟退火算法	151	160	152.19	35
文献[12]中算法	151	155	151.56	60
α-平坦化调度算法	152	152	152	100

通过分析实验二的统计结果可以得出，α-平坦化调度算法同模拟退火算法与文献[12]所提出的算法相比较而言，虽然不能求得该问题的最优解，但所求得的近似解与最优解之间仅相差 1 个单位，且在 100 次的重复求解过程中，所求得的解是一样的、确定的；而模拟退火算法和文献[12]中的算法虽然有时能够求得最优解，但最优解的出现具有随机性，且所求得的最差解同最优解之差较大。所以，综合而言，本文所提出的 α-平坦化调度算法的调度性能是较优的。

结束语　本文在研究多处理机调度问题的基础上，提出 α-平坦的概念，并基于此提出了一种基于 α-平坦的多处理机调度算法——α-平坦化调度算法。该算法运用将处理时间最小与次小的作业不断地相互合并的方法对非 α-平坦的作业集进行平坦化处理；在将作业集平坦化之后，采用"伪蛇行"的调度方式对其进行调度分配，从而求得原问题的一个近似解。通过这种方式可以有效地求得原问题的一个较优的近似解，很好地避开了直接对原调度问题进行求解时所遇到的问题，且不存在运用模拟退火算法等启发式算法进行求解时所带来的随机性问题。

实验结果表明，本文所提出的 α-平坦化调度算法在对多处理机调度问题求解中表现出较好的效果，但是如何确定参数 α 以使所求得的调度策略较优是较困难的，这将是我们下一步的研究重点。

参 考 文 献

[1] Garey M R, Johnson J S. Computers and Intractability: A Guide to the Theory of NP-Completeness [M]. San Francisco, CA: Freeman, 1979

[2] Ahmad I, Kwok Y-K. On Parallelizing the Multiprocessor Scheduling Problem [J]. IEEE Transactions on Parallel and Distributed System, 1999, 10(4): 414-432

[3] Yang Xiao-guang. A glass of generalized multiproces-sor scheduling problems[J]. Systems Science and Mathematical Sciences, 2000, 13(4): 385-390

[4] Huang K L, Liao C J. Ant colony optimization combined with taboo search for the job shop scheduling problem[J]. Computers & Operations Research, 2008, 35: 1030-1046

[5] 张志强，张璟，张翔，等，解决作业车间调度问题的改进蚁群优化算法[J]. 应用科学学报，2010，28(2)：182-188

[6] 徐立芳，莫宏伟. 基于自适应克隆启发算法的作业车间调度[J]. 计算机工程，2009，35(4)：207-209

[7] 蔡斌，毛帆，傅鹂. 等，解决作业车间调度的微粒群退火算法[J]. 计算机应用研究，2010，27(3)：856-859

[8] 唐海波，叶春明。一种求解作业车间调度的混合粒子群算法[J]. 计算机应用研究，2011，28(3)：883-889

[9] Brucker P. Scheduling Algorithms [M]. Springer, 2006

[10] 冯斌，孙俊. 一种多处理机任务分配的启发式算法[J]. 计算机工程，2004，30(14)：63-65

[11] 王剑波，陈内萍. 多处理机独立任务调度问题的 DNA 计算机算法[J]. 湖南师范大学自然科学学报，2009，32(3)：36-41

[12] 高尚，杨静宇. 多处理机调度问题的粒子群优化算法[J]. 计算机工程与应用，2005，41(27)：72-73

(上接第 155 页)

结束语　针对数据流概念漂移问题，本文提出一种基于混合集成方法的数据流概念漂移检测方法 WE-DTB。首先用水平集成的方法建立集成分类器，然后利用朴素贝叶斯分类器去除噪音。实验表明，WE-DTB 方法能够有效检测概念漂移且具有较好的分类精度以及时空性能。然而，如何克服方法在误报率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] Wang H X, Fan W, Yu P S, et al. Mining Concept-Drifting Data Streams Using Eensemble Classifiers[C]// Proc of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington: ACM Press, 2003: 226-235

[4] Scholz M, Klinkenberg R. An Ensemble Classifier for Drifting Concepts[C]// Proc of 16th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Porto, 2005

[5] Fan W. Systematic Data Selection to Mine Concept-Drifting Data Streams[C]// Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, 2004: 128-137

[6] Gao J, Fan W, Han J W. On Appropriate Assumptions to Mine Data Streams: Analysis and Practice[C]// Proc of the 7th IEEE International Conference on Data Mining. Omaha, 2007: 143-152

[7] Zhang P, Zhu X Q, Shi Y, et al. An Aggregate Ensemble for Mining Concept Drifting Data Streams with Noise[C] //Proc of the 13th Pacific-Asia Conference on Knowledge Discovery. Bangkok, 2009: 1021-1029

[8] Li P P, Hu X G, Wu X D. Concept Drifting Detection on Noisy Streaming Data in Random Ensemble Decision Trees[C] //Proc of the 6th International Conference on Machine Learning and Data Mining. 2009: 236-250

[9] Li Y, Zhang Y H, Hu X G. A Classification Algorithm for Noisy Data Streams [C] // Proc of 3rd International Conference of Fuzzy Systems and Knowledge Discovery (FSKD). Yantai, China: Springer, 2010: 2239-2244

[10] ACM Special Interest Group on Knowledge Discovery and Data Mining. KDDCUP99 data set [EB/OL]. http://kdd. ics. uci. edu/databases/ kddcup99, 1999

[11] Tan P-N, Steinbach M. 数据挖掘导论[M]. 范明，等，译. 北京：人民邮电出版社，2006

[12] Wang Y, Li Z-H, Zhang Y. Classifying Noisy Data Streams[C]// Proc of 3rd International Conference of Fuzzy Systems and Knowledge Discovery (FSKD). Xi'an: Springer, 2006: 549-558