ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 5,Sep. 2017,pp. 853－860
DOI:10.16337/j.1004-9037.2017.05.001

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种基于 Tri-training 的数据流集成分类算法

胡学钢　马利伟　李培培

(合肥工业大学计算机与信息学院数据挖掘与智能计算实验室,合肥,230009)

摘　要: 数据流分类是数据挖掘领域的重要研究任务之一,已有的数据流分类算法大多是在有标记数据集上进行训练,而实际应用领域数据流中有标记的数据数量极少。为解决这一问题,可通过人工标注的方式获取标记数据,但人工标注昂贵且耗时。考虑到未标记数据的数量极大且隐含大量信息,因此在保证精度的前提下,为利用这些未标记数据的信息,本文提出了一种基于 Tri-training 的数据流集成分类算法。该算法采用滑动窗口机制将数据流分块,在前 k 块含有未标记数据和标记数据的数据集上使用 Tri-training 训练基分类器,通过迭代的加权投票方式不断更新分类器直到所有未标记数据都被打上标记,并利用 k 个 Tri-training 集成模型对第 $k+1$ 块数据进行预测,丢弃分类错误率高的分类器并在当前数据块上重建新分类器从而更新当前模型。在 10 个 UCI 数据集上的实验结果表明:与经典算法相比,本文提出的算法在含 80% 未标记数据的数据流上的分类精度有显著提高。

关键词: 数据流分类;Tri-training;未标记数据;集成;加权投票

中图分类号: TP274^{+}.3　　**文献标志码:** A

Data Stream Ensemble Classification Algorithm Based on Tri-training

Hu Xuegang, Ma Liwei, Li Peipei

(Data Mining and Intelligence Computing Laboratory, School of Computer and Information, Hefei University of Technology, Hefei, 230009, China)

Abstract: Data stream classification is one of important research tasks in the field of data mining. Most existing data stream classification algorithms require the labeled data for training. However, there are few labeled data in data streams in real applications. To solve this problem, the labeled data can be obtained by manual labeling, but it is very expensive and time consuming. Considering the unlabeled data are huge and full of information, a data stream ensemble classification algorithm based on Tri-training for labeled and unlabeled data is proposed in this paper. The proposed algorithm divides data stream into chunks by sliding windows and trains base classifiers with Tri-training on the first coming k chunks with labeled and unlabeled data. Then the classifiers are iteratively updated by weighted voting until all unlabeled data are labeled. Meanwhile, the $k+1$ data chunk is predicted by using the ensemble model of k Tri-training classifiers and the classifier with higher classification error is discarded, which reconstructs a new classifier on current data chunk to update the model. Experiments on 10 UCI data sets show that the

基金项目: 国家重点研发计划课题(2016YFC0801406)资助项目;国家自然科学基金(61673152,61503112)资助项目;教育部博士点博导基金(20130111110011)资助项目。
收稿日期: 2015-10-18;**修订日期:** 2015-12-11

proposed algorithm can significantly improve the classification accuracy of data stream even with 80% unlabeled data in comparison with traditional algorithms.

Key words: data stream classification; Tri-training; unlabeled data; ensemble; weighted voting

引　言

随着网络技术的迅速普及,电子商务、网上娱乐等实际应用要处理的数据量越来越大,数据形式也越来越多,如:淘宝网上每天成千上万的购物记录,QQ 相关娱乐软件上无时无刻不在产生的新注册用户、聊天记录和游戏日志等等。这类连续到达、高速动态和规模巨大的数据被称为数据流[1]。这些海量数据中含有潜在的有价值信息,如何从海量数据中获取这些有用信息具有重要的现实意义。

数据流分类就是对数据流进行建模和预测。这是一个具有挑战性的课题,其原因在于:现实世界的数据长度无限,变化的特征,以及如何建立与当前特征相一致的分类模型[2]。集成分类技术提供了解决这些难题的方法,其优势在于它们的分类器可以被有效地更新并很容易地适应数据流的改变[3]。一般集成方法[4-6]的实现是将数据流划分为若干个等大小的数据块,用这些数据块训练出分类器,并组合起来去预测未标记数据。不过这些集成方法都是在完全标记数据集上进行训练的。

为了使传统的数据流分类算法更有效,需要大量的标记数据,这就需要人工标注,但人工标注成本昂贵且耗时长。能从根本上解决这一问题的方法是找到能有效使用未标记数据的方法,因为数据流中大部分数据没有标记信息,只有少量数据有标记[1],这些未标记数据同样有效可用。半监督技术提供了能有效使用未标记数据的切实可行的方法。常用的半监督技术有 Co-training[7] 范式、基于生成模型的算法[8-10]和转移半监督支持向量机(Support vector machine, SVM)的 K-means 算法[11]等。其中最具代表的是 Co-training 范式,其要求数据集有两个充足和冗余的视图,在两个不同的视图上分别训练分类器,其中一个分类器对未标记数据进行预测并加标记,用得到的新的标记数据来增大另一个分类器的训练集。然而,实际情况中的数据很难满足 Co-training 的条件。Tri-training[12] 是对 Co-training 范式的一种扩展,它克服了前者的不足,因此应用范围更广泛。然而,Tri-training 是批处理分类算法,当面向数据流环境时,由于时空性能的问题其难以直接用于数据流分类。

鉴于 Tri-training 算法对处理不完全数据的优势,本文提出一种适应于数据流环境的基于 Tri-training 的集成分类算法(Tri-training based data stream ensemble classification algorithm, TriTDS)。该算法采用滑动窗口机制将数据流分块,在前 k 块数据集上利用未标记数据和标记数据训练 Tri-training 基分类器,通过加权投票方式迭代地为未标记数据打上标记并加入到标记数据中,从而不断地更新分类器直到所有未标记数据都被打上标记。同时,为适应数据流环境,当第 $k+1$ 块数据块到来时,利用 k 个 Tri-training 的集成模型对其进行预测,丢弃分类错误率高的分类器并在当前数据块上重建新分类器从而更新当前模型。在 10 个 UCI 数据集上的实验结果表明:TriTDS 可以使用高达 80%以上的未标记数据进行训练,并保持良好的分类精度,对于某些特殊的数据集甚至可以使用 99%的未标记数据进行训练。与经典算法相比,该算法既能有效地处理数据流分类问题又能有效使用未标记数据以降低训练数据成本,同时分类精度也有显著提高。

1　不完全数据流分类方法

实际应用中产生的数据流含有标记数据和未标记数据,传统的数据流分类算法要求训练数据全标记,不足以解决不完全数据流分类问题,而基于半监督技术的数据流分类算法则能有效使用标记数据和未标记数据进行训练。常见的基于半监督技术的数据流分类算法有很多,比如使用聚

类来选择可信的未标记数据,并用它们来增量地重训练分类器[13]。首先在标记数据上训练初始分类器,然后从当前流中抽取一定数量的未标记数据,使用 K-prototype 聚集对其加标记,并用当前分类器调整聚类标记。选择当前分类器和 K-prototype 聚集两种方法对未标记数据标记一致的样例来更新分类器。文献[3]使用微簇和 k 近邻分类算法建立适合数据流特征的分类模型。文献[14]更深入地适应数据流分类特征来处理概念漂移,是基于一种进化的决策树,根据历史概念簇和通过一种成熟的聚类算法 K-Modes 生成的新簇之间的偏差来在叶子上区别概念漂移和噪音。文献[15]结合分类器和聚类器,提出了一种新的标记传递方法,即使用从分类器得到的类标记信息和簇内部结构信息来推断每个簇的类标记,并提出了一种新的加强机制,即依据所有基模型与其最新基模型之间的相容性为其加权,通过加权平均机制将所有分类器和簇能组合在一起。文献[16]用基于相似度的方法产生概念簇来处理未标记数据,通过检测和调整趋势与值的变化,自适应地构建电力需求供应和定价学习模型,用于解决由于电力市场反常,电价不固定导致的类标记不可用及潜在有价值信息丢失等问题。文献[3,13-16]通过聚类方法对未标记数据进行处理,再使用标记传递等方法结合标记数据的信息来处理数据流分类问题。

文献[17]提出的 Udeed 算法在标记数据上尽可能地提高分类器的分类精度,同时开采未标记数据来增强集成的多样性。它使用未标记数据对使用的模型进行改进,并不直接预测未标记数据的标记信息。文献[11]提出一种在数据流上建立预测模型的框架,依据基于转移半监督 SVM 的 K-means 算法进行学习。文献[18]综合了单视图树结构和 Co-training 算法,可以有效地使用未标记数据和标记数据处理多类问题。文献[19]提出了一种在线数据流分类方法,使用选择自训练对有限标记数据进行学习,只需要很少(20%左右)的标记数据就可以达到理想效果。文献[2]通过检测仅使用有限标记数据训练分类器的可信度的显著变化来动态决定数据块的边界,并集成合适的分类技术提出了一种完整的半监督框架。文献[2,11,18,19]通过估计易出错的伪标记的未标记数据概率来增大标记数据集进而提高分类精度。本文提出的算法也属于该类。

2 基于 Tri-training 的数据流集成分类算法

本文提出了一种基于 Tri-training 的数据流集成分类算法 TriTDS,将 Tri-training 算法作为基分类器,实现了半监督技术和集成模型相结合处理数据流分类的思想。

2.1 Tri-training 算法

Tri-training 算法是通过对 Co-training 算法的扩展所得,相比于 Co-training 算法对数据集的严苛要求,Tri-training 算法不要求充足和冗余的视图,不要求使用两个不同的监督学习算法将样例空间划分成两个等价类的集合。它建立 3 个基分类器,只需要用一种监督学习算法在不同的训练集上训练即可轻易得到。其思想是:首先在原标记数据集上训练得到 3 个基分类器 h_1,h_2 和 h_3,任何一个分类器 $h_i(i\in\{1,2,3\})$欲对未标记样例 x 打标记,只需另外两个分类器对该样例 x 的标记加注一致,即若两个分类器对该样例 x 的标记都是 class,则 h_i 标记未标记样例 x 为 class,用于 h_i 的更新[12]。

Tri-training 算法为经典的半监督技术,它可以有效地使用未标记数据和标记数据训练分类器。但限于其是静态数据分类算法,当数据流数据量巨大时,Tri-training 算法将耗费大量内存,因此不能直接用于流环境下的数据分类问题。基于 Tri-training 算法能有效使用未标记数据的特性,本文提出了基于其的一种数据流分类算法。

2.2 基于 Tri-training 的数据流集成分类算法分析

图 1 展示了本文所提算法的基本思路。如图 1 所示,本算法借鉴了传统的基于协同训练策略处理

不完全标记数据的方法。设数据流 DS 包含标记数据和未标记数据，TriTDS 采取固定窗口机制将数据流 DS 划分成一系列数据块，记为 $D_i=\{l_i^1,l_i^2,\cdots,l_i^z,u_i^{z+1},u_i^{z+2},\cdots,u_i^{|D_i|}\}$，其中包含 z 个标记样例和 $|D_i|-z$ 个未标记样例，从数据流自身考虑，数据流的属性特征和数据块的属性特征一致。这些数据块随时间顺序产生，在最早到达的 k 个数据块 $D_1,D_2,\cdots,D_k$ 上使用半监督技术 Tri-training 算法训练得到 k 个基分类器 $T_1,T_2,\cdots,T_k$。此外，因为随着数据块的不断到来，数据的概念不断变化，最早建立的基分类器不能很好地适应最新到来的数据，因此需要不断更新基分类器。最后，通过等值加权投票策略对测试集进行测试得到分类精度。

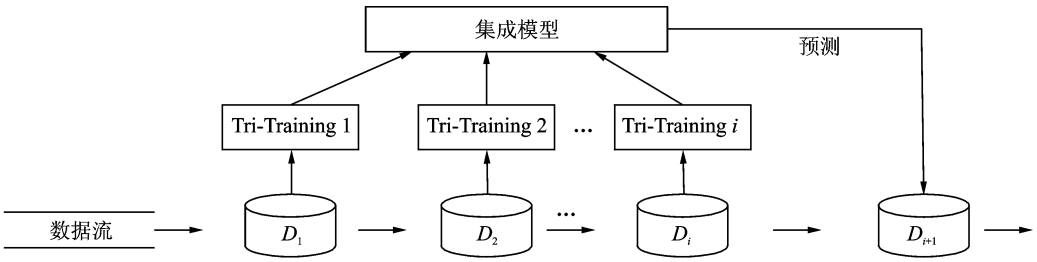


图 1 一种基于 Tri-training 的数据流集成分类算法示意图

Fig. 1 Sketch of data stream ensemble classification algorithm based on Tri-training

本算法的基分类器更新策略是：依次测试建立的 k 个基分类器 $T_1,T_2,\cdots,T_k$ 对最新到来数据块 D_{i+1} 的错误率 $\mathrm{err}_j(j\in\{1,2,\cdots,k\})$，若到第 $r(r\in\{1,\cdots,k\})$ 个基分类器 T_r 时其错误率 err_r 高于阈值 th，则 T_r 应在最新数据块 D_{i+1} 上重新训练；为了保持基分类器的多样性，用于更新剩下的基分类器的数据块应为 D_{i+2}，再依次测试剩下的基分类器 $T_{r+1},\cdots,T_k$ 在 D_{i+2} 上的错误率 $\mathrm{err}_j(j\in\{r+1,\cdots,k\})$，若到第 $s(s\in\{r+1,\cdots,k\})$ 个基分类器 T_s 时其错误率 err_s 高于阈值 th，则 T_s 在最新数据块 D_{i+2} 上重新训练；如此循环，直到更新到最后一个分类器。

算法 1 是本算法的伪代码。k 表示基分类器的个数；readChunk(DS,m)实现窗口机制，从 DS 中读取大小为 m 的一个数据块；err_i 表示分类器 T_i 的错误率；measureErr(D_{i+1},i,cL)计算分类器 T_i 的错误率；cL 存放数据集 D_{i+1} 的正确标记值；th 代表阈值($th\geqslant 0.5$)。

算法 1 一种基于 Tri-training 的数据流集成分类算法

FOR $i=1$ to k do

　　$D_i=$readChunk(DS,m)

　　使用 Tri-training 算法在 D_i 上训练得到基分类器 T_i

end FOR

　　$D_{i+1}=$readChunk(DS,m)

FOR $i=1$ to k do

　　$\mathrm{err}_i=$measureErr(D_{i+1},i,cL)

　　IF $\mathrm{err}_i>th$

　　Tri-training 在 D_{i+1} 上重新训练得到基分类器 T_i'，取代 T_i

　　$D_{i+1}=$readChunk(DS,m)

　　end IF

end FOR

3 实验与分析

3.1 实验数据集及相关设计

本实验使用 10 个 UCI 数据集,这些数据集只有两类,且都有标记,包含 351～100 000 条数据,其基本参数如表 1 所示。由于本算法是对不完全标记数据流进行分类,而 UCI 数据集为静态数据且带有标记。为了模拟不完全标记数据流,首先将 10 个数据集改为 ARFF 格式,对每个 ARFF 格式的数据集,使用 weka 的 Instances 类创建出只包含 UCI 数据集信息的结构文件,然后利用创建的结构文件逐条读取数据集,从而实现数据的流动性。当读取 m 条数据时建立一个数据块,为了确保足够的数据块个数,令 $m \leqslant |$数据集$|/12$。对每个数据块采取随机抽样方式抽取比例为 w 的数据,$w \in \{0.05, 0.1, \cdots, 0.9, 0.95, 0.99\}$,去掉这些数据的标记,如此数据块就是由比例为 w 的未标记数据和比例为 $1-w$ 的标记数据组成,去掉标记的数据样本数占每个数据块的比重称为未标记数据率 ur。

该算法使用 J48 作为 Tri-training 算法的基分类器,Tri-training 算法作为 TriTDS 算法的基分类器,TriTDS 在最新到达的 k 个数据块上训练,训练得到 k 个分类器。

表 1 实验数据集

Tab. 1 Experiment datasets

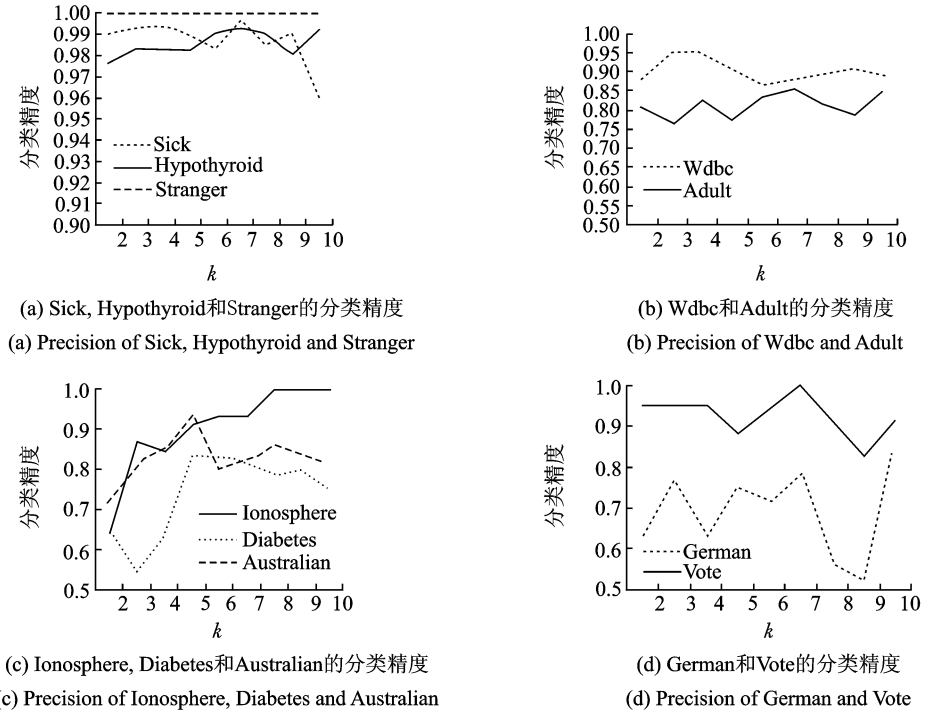
数据集	类数	属性数	大小	类属样例个数之比(%/%)
Ionosphere	2	34	351	35.9/64.1
Vote	2	16	435	61.4/38.6
Wdbc	2	30	569	37.3/62.7
Australian	2	14	690	44.5/55.5
Diabetes	2	8	768	65.1/34.9
German	2	20	1 000	70.0/30.0
Hypothyroid	2	25	3 163	4.8/95.2
Sick	2	29	3 772	92.3/7.7
Adult	2	14	32 561	24.1/75.9
Stranger	2	7	100 000	88.9/11.1

3.2 实验结果和分析

TriTDS 实验涉及到基分类器个数 k、未标记数据率 ur 和基分类器更新阈值 th 3 个参数。对于集成学习而言,组合的基分类器的个数不是越多越好,因为过多的分类器会导致冗余。为了确定合适的基分类器个数 k,当 $ur=0.2$,$th=0.5$ 时,分别对 k 取 2～10 进行实验,实验结果如图 2 所示。由图 2(a)可知,随着 k 的增大,算法在 Sick,Hypothyroid 和 Stranger 3 个数据集上的分类精度波动幅度不大,最高不超过 3%,且分类精度保持在 95%以上;图 2(b)显示算法在 Wdbc 数据集上的分类精度会在[0.867,0.950]之间波动,在 Adult 数据集上的分类精度会在[0.750,0.850]之间波动;图 2(c)显示算法在 Ionosphere,Diabetes 和 Australian 3 个数据集上的分类精度先增大后呈波动式平稳,而图 2(d)显示算法在 German 和 Vote 数据集上的分类精度当 $k>7$ 时会发生大幅度下降。当 $5 \leqslant k \leqslant 7$ 时,算法在 5 个数据集上的分类精度可保持在 80%以上。综上,参数基分类器个数的最优取值为 $5 \leqslant k \leqslant 7$,以下实验选择最优取值 $k=7$。

本文在 $k=7$ 时调整 th 值使基分类器更优,经过对比 10 组数据集的实验结果(由于数据量过大,限

于篇幅要求，此处仅给出实验结论）可知 $th=0.4$ 时 TriTDS 算法的鲁棒性最好，适应性最广。



(a) Sick, Hypothyroid和Stranger的分类精度

(a) Precision of Sick, Hypothyroid and Stranger

(b) Wdbc和Adult的分类精度

(b) Precision of Wdbc and Adult

(c) Ionosphere, Diabetes和Australian的分类精度

(c) Precision of Ionosphere, Diabetes and Australian

(d) German和Vote的分类精度

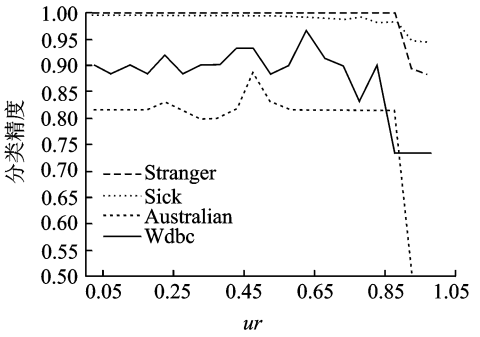
(d) Precision of German and Vote

图 2 当 $ur=0.2$，$th=0.5$ 时，TriTDS 算法在 10 个数据集上的分类精度

Fig. 2 Precision of TriTDS on ten datasets when $ur=0.2$ and $th=0.5$

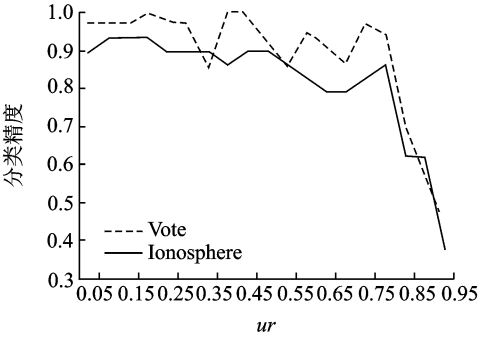
图 3 给出了当 $k=7$，$th=0.4$ 时 TriTDS 在 10 个数据集上的分类精度随未标记数据的变化情况。由图 3 可知：(1)图 3(a)显示 TriTDS 算法在 Stranger，Sick，Australian 和 Wdbc 4 个数据集的分类精度在 $ur\geqslant 0.95$ 时开始极速下降，图 3(b)显示算法在 Vote，Ionosphere 两个数据集的分类精度当 $ur\geqslant 0.85$ 时开始极速下降，算法在这 6 个数据集上的分类精度在没有极速下降之前都是在一定范围内波动；可知，使用未标记数据训练基分类器时，使用的未标记数据量存在限制，因为使用的未标记数据越多，即使用的标记数据的量越少，方法可以参考的标记信息越少，进而导致分类精度下降。(2)图 3(c)显示 TriTDS 算法在 Hypothyroid，Diabetes，German 和 Adult 4 个数据集上的分类精度随着 ur 的变大一直在一定范围内波动，没有明显下降趋势。分析可知：TriTDS 算法对于合适的数据集在保证分类精度的前提下可以使用近可能多的未标记数据。这 10 个数据集中有一个数据集的分类精度特殊，Stranger 数据集的分类精度可以高达 1，这是因为 stranger 为离散数据，与本算法的契合度高。综上，TriTDS 算法至少可以使用 80%未标记数据来训练基分类器。

表 2 给出了所提 TriTDS 算法与 Tri-training 和 Zhang 等所提算法[15]的实验对比结果（粗体表示最大值；为了节省篇幅，对比实验取 $ur\in\{0.2,0.4,0.6,0.8\}$）。由结果分析可知，在 $ur=0.2$ 时，TriTDS 算法在 7 个数据集上取得最大分类精度；在 $ur=0.4$ 时有 7 个；在 $ur=0.6$ 时有 8 个；在 $ur=0.8$ 时有 5 个。与基准算法相比，TriTDS 算法的优势很明显，由此可知，TriTDS 算法可以更加有效地使用未标记数据处理数据流分类问题。Zhang 等所提的算法也可以使用未标记数据处理数据流问题，本文的算法 TriTDS 优于其的原因在于在使用未标记数据时能保持较高的分类精度。



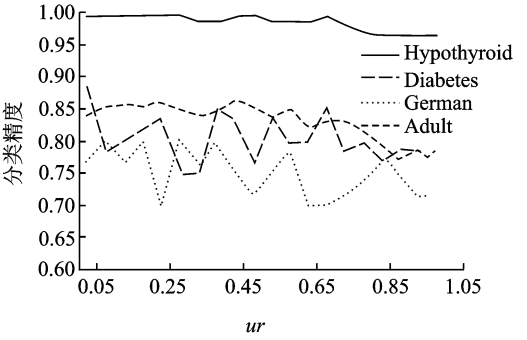
(a) Stranger, Sick, Australian和Wdbc的分类精度随未标记数据率的变化趋势

(a) Change tendency of precision of Stranger, Sick, Australian and Wdbc varying with *ur*



(b) Vote和Ionosphere的分类精度随未标记数据率的变化趋势

(b) Change tendency of precision of Vote and Ionosphere varying with *ur*



(c) Hypothyroid, Diabetes, German和Adult的分类精度随未标记数据率的变化趋势

(c) Change tendency of precision of Hypothyroid, Diabetes, German and Adult varying with *ur*

图 3 TriTDS 算法在 10 个数据集上分类精度随未标记数据率 *ur* 的变化趋势

Fig. 3 Change tendency of precision of TriTDS on 10 datasets with the unlabelled data rate *ur*

表 2 TriTDS, Tri-training 和文献[15]算法在不同未标记率 *ur* 数据集上的分类精度

Tab. 2 Datasets′ precision of TriTDS, Tri-training and the algorithom in Ref. [15] on different rates of unlabeled *ur*

Dataset	*ur*=0.2			*ur*=0.4			*ur*=0.6			*ur*=0.8		
	TriTDS	Tri-training	算法[15]	TriTDS	Tri-training	算法[15]	TriTDS	Tri-training	算法[15]	TriTDS	Tri-training	算法[15]
Stranger	1.000	1.000	0.884	1.000	1.000	0.889	1.000	1.000	0.889	0.917	1.000	0.888
Sick	0.997	0.987	0.914	0.997	0.988	0.924	0.993	0.984	0.923	0.993	0.979	0.981
Hypothyroid	0.993	0.900	0.963	0.987	0.989	0.952	0.987	0.989	0.956	0.970	0.989	0.930
Australian	0.889	0.852	0.839	0.889	0.852	0.830	0.933	0.838	0.850	0.867	0.807	0.800
Wdbc	0.911	0.951	0.913	0.889	0.934	0.908	0.933	0.923	0.903	0.911	0.925	0.908
Vote	0.867	0.960	0.624	0.867	0.951	0.623	0.822	0.951	0.634	0.800	0.945	0.613
Ionosphere	0.967	0.898	0.932	0.967	0.913	0.895	0.967	0.913	0.867	0.933	0.879	0.880
Diabetes	0.817	0.900	0.768	0.850	0.727	0.745	0.800	0.748	0.751	0.800	0.712	0.762
German	0.800	0.704	0.750	0.800	0.707	0.749	0.783	0.684	0.747	0.733	0.676	0.747
Adult	0.883	0.814	0.804	0.867	0.807	0.809	0.883	0.831	0.809	0.850	0.797	0.800

4 结束语

本文提出了一种适应数据流环境的基于 Tri-training 的集成分类算法。该算法受批处理算法 Tri-training 的启发，考虑其在静态未标记数据集上的分类优势，成功将其应用于数据流分类中。该算法能在保持分类精度的情况下有效使用未标记数据训练分类器，相比基于完全标记数据的传统数据流分类方法，该算法数据成本大幅度降低。通过与其他传统的基于半监督技术的数据流分类算法在 10 个 UCI 数据集上的对比实验结果表明：所提算法可以在高达 80%的未标记数据集上训练分类器，并保持很高的分类精度。未来工作将尝试将所提数据流分类算法用于井下传感器数据流分析。

参考文献：

[1] Hand D J. Statistics and data mining[J]. ACM Sigkdd Explorations Newsletter, 1999,1(1):16-19.

[2] Haque A, Khan L, Baron M. Semi-supervised adaptive framework for classifying evolving data stream[C]//19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Ho Chi Minh City, VIETNAM: John Neumann Inst, 2015:383-394.

[3] Masud M M, Gao J, Khan L, et al. A practical approach to classify evolving data streams: Training with limited amount of labeled data[C]//8th IEEE International Conference on Data Mining. Pisa, ITALY: IEEE, 2008:929-934.

[4] Scholz M, Klinkenberg R. An ensemble classifier for drifting concepts [C]//Proceedings of ICML/PKDD Workshop in Knowledge Discovery in Data Streams. [S. l.]: IEEE, 2005,6(11):53-64.

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

[6] Song Ge, Ye Yunming. A new ensemble method for multi-label data stream classification in non-stationary environment [C]//Neural Networks (IJCNN), 2014 International Joint Conference on. Beijing, China: [s. n.], 2014:1776-1783.

[7] Mitchell B T. Combining labeled and unlabeled data with co-training[C]// 11th Annual Conference on Computational Learning Theory by ACM. Madison, WI, USA: ACM, 1998:92-100.

[8] Shahshahani B, Landgrebe D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon[J]. IEEE Transactions on Geoscience and Remote Sensing, 1994,32(5):1087-1095.

[9] Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labeled and unlabelled data[J]. Medical Imaging IEEE Transactions on, 1997,9(5):571-577.

[10] Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000,39(2/3):103-134.

[11] Zhang Peng, Zhu Xingquan, Guo Li. Mining data streams with labeled and unlabeled training examples [C]//9th IEEE International Conference on Data Mining. Miami Beach, FL: IEEE,2009:627-636.

[12] Zhou Zhihua, Li Ming. Tri-training- exploiting unlabeled data using three classifiers[J]. Knowledge and Data Engineering, IEEE Transactions on, 2005,17(11):1529-1541.

[13] Wu Shuang, Yang Chunyu, Zhou Jie. Clustering-training for data stream mining[C]//6th IEEE International Conference on Data Mining. Hong Kong, China: IEEE,2006:653-656.

[14] Wu Xindong, Li Peipei, Hu Xuegang. Learning from concept drifting data streams with unlabeled data[J]. Neurocomputing, 2012,92:145-155.

[15] Zhang Peng, Zhu Xingquan, Tan Jianlong. Classifier and cluster ensembles for mining concept drifting data streams[C]//10th IEEE International Conference on Data Mining. Sydney, Australia: IEEE, 2010:1175-1180.

[16] Patil P, Fatangare Y, Kulkarni P. Semi-supervised learning algorithm for online electricity data streams[J]. Advances in Intelligent Systems and Computing, 2014,(324):349-358.

[17] Zhang Minling, Zhou Zhihua. Exploiting unlabeled data to enhance ensemble diversity[J]. Data Mining and Knowledge Discovery, 2013,26(1):98-129.

[18] Hady M F A, Schwenker F, Palm G. Semi-supervised learning for tree-structured ensembles of RBF networks with co-training [J]. Neural Networks, 2010,23(4):497-509.

[19] Loo H R, Marsono M N. Online data stream classification with incremental semi-supervised learning[C]//The Second ACM IKDD Conference on Data Sciences. New York, NY, USA: ACM, 2015:132-133.

作者简介：

胡学钢(1961-)，男，博士，教授，研究方向：知识工程、数据挖掘和数据结构，E-mail：jsjxhuxg@hfut.edu.cn。

马利伟(1988-)，女，硕士研究生，研究方向：数据流挖掘，E-mail：malwei@126.com。

李培培(1982-)，女，博士，副研究员，研究方向：数据流挖掘，E-mail：peipeili@hfut.edu.cn。