一种基于分类器相似性集成的数据流分类研究

刘余霞[1] 吕 虹[1,2] 刘三民[3]
(安徽工程大学电气工程学院 芜湖 241000)[1] (安徽建筑工业学院电子与信息工程学院 合肥 230022)[2]
(安徽工程大学计算机与信息学院 芜湖 241000)[3]

摘 要 数据流分类已成为当前研究热点之一,如何解决其中的概念漂移和噪声是关键问题,为此提出了一种新的基于分类器相似性的动态集成算法。由于数据流中相邻数据具有相同概念的概率较大,因此用最新基分类器代表数据流中即将出现的概念,同时基于此分类器求出基分类器之间的相似性作为权值进行加权多数投票,并根据相似性大小淘汰较弱基分类器以适应概念漂移和噪声。在标准仿真数据集上进行了仿真实验,结果表明该算法相比其他集成方法在抗噪性能和分类准确性方面均得到显著提高。

关键词 概念漂移,相似性,集成学习,数据流分类,加权多数投票
中图法分类号 TP393 **文献标识码** A

Classifier-similarity-based Ensemble Classification Research for Data Streams

LIU Yu-xia[1] LV Hong[1,2] LIU San-min[3]
(College of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China)[1]
(College of Electronic and Information, Anhui Institute of Architecture and Industry, Hefei 230022, China)[2]
(College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China)[3]

Abstract Classification of data streams has become one of hot research spots, and a new similarity-based dynamic ensemble algorithm was presented to deal with two critical problems, namely, concept drift and noise. Because adjacent data in data stream has the same concept with more probability, the new sub-classifier stands for the coming concept. Based on it, ensemble classifier is got by similarity weighted majority voting, and sub-classifier with worst performance is deleted for suiting for concept drift and noise. Experiment result on simulation data set shows the algorithm is best than other schema in classification accuracy and anti-noise.

Keywords Concept drift, Similarity, Ensemble learning, Data stream classification, Weighted majority voting

目前数据流已广泛出现在如网络安全、传感网络、金融和天气预测等应用领域,成为当前研究热点。同时数据流具备量大、快速、持续性、概念漂移和包含噪声等特点,导致建立在静态、集中数据集上的分类模型难以适应。

关于数据流分类,在国内外已取得较多研究成果。按分类模型是否具备概念漂移检测功能,可将概念漂移数据流分类研究分为两大类:显式探测分类和隐式探测分类。显式探测分类采用统计推断[1-4]、信息论[5]等方法来设置概念漂移检测模块,检测数据流发生漂移时间点,并把漂移时间点后的数据缓存,重新训练新的分类器。该类方法可归结为滑动窗口(Sliding Window)类型,其难点在于窗口大小难以确定,设置合适的区间精度和置信度等参数是关键。相比显示探测方法,通过集成学习实现隐式探测的分类方法具有较多优势,它克服了显式探测方法参数难以控制的缺陷。文献[6]首次用集成学习方法解决数据流分类问题并提出SEA算法,即通过投票法获得最终分类结果。Wang等[7]在SEA算法基础上,提出加权数据流集成分类器框架WCE,使分类准确率高的分类器具有较高投票权重,提高了集成分类器的准确率。文献[8,9]根据数据流概念漂移特点,提出集成学习框架Learn++.NSE,以实现样本权重、分类器权重的灵活配置,解决样本不平衡环境下的分类问题。文献[10]提出在线数据流集成分类器算法,它根据分类器分类情况制定分类器权重更新策略和分类器淘汰方法,有效解决了概念变化频繁的数据流分类问题。文献[11]设计了一种不平衡数据流集成分类算法,用以解决由数据集不平衡引起分类器分类性能下降的问题。从现有文献分析可知,其基本是用分类器在当前数据集上的分类性能作为集成投票权重。但各分类器是基于历史数据集训练所得,同时由于概念漂移和噪声,使得上述方法不能准确地反映出分类器知识与数据流当前概念的相关性。

针对具有噪声的概念漂移数据流分类问题,本文采用基分类器间的相似性来代替上述文献中用分类器的准确率作为权重进行加权投票。实验结果分析表明,用相似性作为权重具有更好的分类性能和抗噪特点,说明相似性更能表明基分类器所拥有的知识与数据流概念间具有更强的相关性。

到稿日期:2012-02-25 返修日期:2012-07-23 本文受国家自然科学基金(61071001),安徽省教育厅自然科学基金(KJ2008A010),安徽省高等学校青年教师科研计划项目(2012SQRL220)资助。
刘余霞(1980—),女,硕士生,主要研究方向为信号处理、模式识别、计算机应用;吕 虹(1959—),女,教授,主要研究方向为信号处理、计算机应用。

1 分类器相似性

文中所定义的分类器相似性是一种基于后验概率的计算模型，它具有两大优点：一是分类模型间的独立性，即可以在异态子分类器间实现基于相似性集成；二是预测能力的独立性，即可把子分类器功能看作黑盒子，以便实现子分类器的共享。

为便于描述分类器相似性，先定义分类器在单个样本上的相似性。设 e 是样本，E 是样本验证集合，满足 $e\in E$。符号 $s_e(M_x,M_y)$ 表示分类器 x,y 关于样本 e 的相似性，同时相似性满足下述 3 条性质：

(1)有界性：$0\leqslant s_e(M_x,M_y)\leqslant 1$

(2)对称性：$s_e(M_x,M_y)=s_e(M_y,M_x)$

(3)自相似性：$s_e(M_x,M_x)=1$

受文献[12]启发，在文中用高尔相似系数(Gower's Similarity Coefficient)作为分类器间相似性计算模型。

基于单样本相似性的计算方法如式(1)所示：

$$s_e(M_x,M_y)=1-\delta_e(M_x,M_y) \tag{1}$$

式中，$\delta_e(M_x,M_y)$ 的计算见式(2)：

$$\delta_e(M_x,M_y)=\frac{1}{|C|}\sum_{j=1}^{|C|}\frac{|PD_{xj}(e)-PD_{yj}(e)|}{R_j(e)} \tag{2}$$

式中，$|C|$ 表示类别数，概率 $PD_{xj}(e)$ 表示子分类器 x 在样本 e 上关于类 j 的后验概率，$PD_{yj}(e)$ 表示子分类器 y 在样本 e 上关于类 j 的后验概率，$R_j(e)$ 表示样本 e 关于类 j 的后验概率极差，计算方法如式(3)所示：

$$R_j(e)=\max\{PD_{1j}(e),\cdots,PD_{Nj}(e)\}-\min\{PD_{1j}(e),\cdots,PD_{Nj}(e)\} \tag{3}$$

在单样本相似性基础上，很容易导出分类器间的相似性，即在整个验证集合上样本相似性的算术平均值：

$$s_E(M_x,M_y)=\frac{1}{|E|}\sum_{e\in E}s_e(M_x,M_y) \tag{4}$$

式中，$|E|$ 表示验证集合样本个数。

2 动态相似性集成分类算法

本文分类算法的实现基于后述的假设：最新分类器能代表当前数据流出现的概念，且数据流中相邻数据间具有较大概率保证两者概念相似。基于此假设，定义在最近数据块上建立的分类器作为参考分类器，其他分类器与参考分类器在当前数据块上的相似性作为各分类器的投票权重，以此进行集成学习。

在第 1 节相似性定义的基础上，结合动态集成思想，可描述出算法 DSWMV(DS,M,EC)伪代码。其中算法参数 DS 表示数据流；M 表示当前基分类器个数，初始值为 0；EC 表示集成分类器，初始值为空，最大规模容量 20。

步骤：

(1) M=1
(2) 形成数据块，并训练基分类器 Cls_1，EC=EC∪Cls_1
(3) While(DS is not end)
(4)　形成数据块 DB
(5)　调用函数 computeSim(EC,DB,M)计算相似性，返回相似性权重向量 w
(6)　基于相似性权重多数投票方法对数据块 DB 进行分类预测
(7)　If (M=20)
(8)　　用基分类器 Cls_M 取代相似性值最小的基分类器
(9)　　M=M−1
(10)　End if
(11)　If (M<20)
(12)　　M=M+1
(13)　　在数据块 DB 上训练新的基分类器 Cls_M，EC=EC∪Cls_M
(14)　End if
(15) End while

其中，computeSim(EC,DB,M)表示当前基分类器相对于参考分类器的相似性，其算法可描述如下：

(1) For (i=1,…,|DB|)
(2)　For (j=1,…,M)
(3)　　p_{ji}=classify(Cls_j,DB_i)
(4)　For (k=1,…,|C|)
(5)　　根据式(3)计算 $R_k(i)$
(6)　For (j=1,…,M−1)
(7)　　基于样本 DB_i，根据式(2)计算基分类器 Cls_M 和 Cls_j 相似性
(8)　　累加步骤(7)所得相似性
(9)　End for
(10) End for
(11) 根据式(4)计算相似性向量 w，并返回

其中，$|DB|$ 表示数据块大小，$|C|$ 表示样本类别个数，p_{ji} 表示基分类器 j 关于样本 DB_i 的预测概率。

3 仿真实验

3.1 数据集

实验所用数据集源自平台 MOA (Massive Online Analysis)[13] 环境中的移动超平面数据集[14] 和 SEA[6]，分别代表缓漂移和突漂移数据流类型。

(a)移动超平面数据集

该数据集通过 m 维超平面 $\sum_{i=1}^{m}a_ix_i=a_0$ 随机生成，样本的属性值 x_i 在区间[0,1]上随机均匀生成。正类样本是由满足 $\sum_{i=1}^{m}a_ix_i\geqslant a_0$ 的属性向量($x_1,x_2,\cdots,x_m$)构成，反之为负类样本，即 $\sum_{i=1}^{m}a_ix_i<a_0$，其中 $a_0=(\sum_{i=1}^{m}a_i)/2$。数据集形成过程主要受 3 个参数影响，噪声参数(n)表示改变样本类标号引入噪声数据量；参数 t 表示权值的改变量(每 N 个样本)。$s_i\in\{-1,1\}$ 表示每个权 a_i 方向改变量，即每隔 N 个样本，权 a_i 改变 s_i*t/N。参数 s 表示以概率 s 使每隔 N 个样本超平面方向发生翻转，即 s_i 变成 $-s_i$。在本实验数据集中共有 5 个特征属性，其中 2 个特征属性发生漂移，每个数据集含 100W 个样本。为验证本文研究目标，在固定参数 $t=0.1$ 和参数 $s=10\%$时，改变噪声参数 n(5%、10%、15%)，依次生成 3 个数据集(记 H1、H2、H3)进行实验。

(b)SEA 数据集

SEA 数据集样本向量结构为(f_1,f_2,f_3,C)，属性 f_1,f_2 是条件属性，C 是样本类别属性。当 $f_1+f_2\leqslant\theta$ 时，属于正类，否则属于负类样本。在数据集中共有 4 种概念，即 θ 依次取值 8,9,7,9.5。数据集总样本数为 100W，各概念样本数分别为 25W，并在各概念中分别加入 10%的噪声，记样本数据集为 SEA。

3.2 实验方案

实验是基于 WEKA 平台在 Ecplise 环境中实现的，在数据流分类时，文献[13]提出的先测试后训练策略使集成分类算法总能测试未遇到的样本，可以有效地利用样本数据，更真

实地反映实际环境。为便于与基于动态相似性集成方案对比，结合现有文献共实现3种集成方案算法，分别是基于相似性动态集成(Dynamic Similarity-based Weighted Majority Voting，DSWMV)、基于准确率动态集成(Dynamic Accuracy-based Weighted Majority Voting，DWMV)、简单投票集成(Majority Voting，MV)。本文所涉及的动态性含义是指是否淘汰性能指标(如相似性、准确率)较差的分类器。其中方案MV是静态投票方法，指各基分类器固定不变，投票权重相同，均设为1。上述3种集成方案分别在两种概念漂移类型中含有不同噪声的4个数据集(H1、H2、H3、SEA)上进行实验分析。在实验过程中，集成分类器最大规模为20个基分类器，基分类器用贝叶斯学习器通过批处理方式训练生成，每个数据块大小为1000个样本。

在上述思想的指导下，4个数据集上的实验结果分别如图1—图4所示。

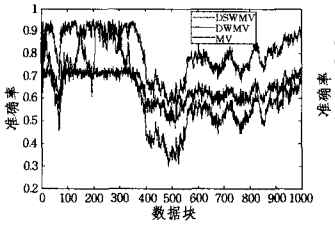

图1 数据集H1实验数据

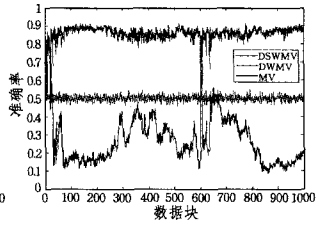

图2 数据集H2实验数据

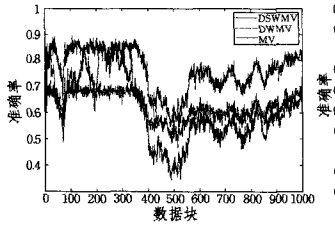

图3 数据集H3实验数据

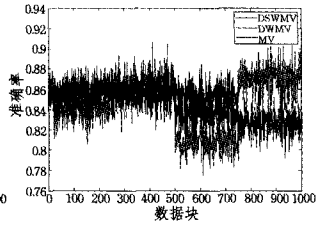

图4 数据集SEA实验数据

从图1—图3可清晰看出，基于动态相似性的集成算法DSWMV准确率较其他方案要好，且从表1准确率的平均值上分析可知约高出20%左右，说明采用基分类器间的相似性作为权值的方法能准确捕获数据流中出现的概念漂移现象，且随着噪声增大相比其他方案表现出较好的稳定性，具备较强的抗噪特点。这是因为DSWMV及时训练最新概念的基分类器，用相似性作为历史基分类器投票权重。对与当前概念相关的基分类器给予较高的话语权是合情合理的，可提高集成分类精度，同时能充分利用历史分类器所拥有的知识来解决数据流概念漂移呈现周期性的问题。从3幅图的曲线分析可知，在分类能力方面，动态调整子分类器(DSWMV、DWMV)比静态方式(MV)表现要好，说明当概念漂移现象出现时，应该及时调整集成分类器拥有的知识，这是概念漂移现象研究的关键点。3幅图内的曲线位置证明了本文假设前提是可行的，当参考分类器概念准确时，其集成更能反映数据流中的概念变化情况。

表1 准确率统计量

数据集		DSWMV	DWMV	MV
H1	平均值	0.81	0.65	0.61
H2	平均值	0.86	0.50	0.27
H3	平均值	0.77	0.63	0.61
SEA	平均值	0.86	0.85	0.84

图4反映的是在突漂移数据流上的实验结果，图中曲线变化反映数据流中概念漂移情况，能及时地识别4个基本概念。由于概念漂移出现在基本概念之间，在每一个基本概念内仅受噪声影响，因此各种分类方法的分类性能较为稳定，体现集成学习的优势，但从图中仍可看出，DSWMV性能相比其他方法要好，而且从表1的平均值也可验证这一现象。

结束语 本文采用基分类器的相似性代替传统基于准确率的集成方法进行概念漂移数据流分类研究。仿真实验结果表明，该方案是可行的。方案的关键之处是寻找相似性度量机制来准确描述分类器所拥有的知识与当前数据流概念的相关度。算法的动态特征是直接淘汰相似性较小的分类器，对于相似性动态调节在文中未涉及。寻找新的相似性计算方法，实现在线学习，如何把本文集中学习方式迁移到分布式环境将是后续工作重点。

参考文献

[1] Bifet A, Gavalda R. Learning from time changing data with adaptive windowing [R]. Universitat Polit ecnica de Catalunya, 2006

[2] Baena-Garcia M, del Campo-Avila J, et al. Early Drift Detection Method [C]// Proc of 4th International Workshop on Knowledge Discovery from Data Streams. 2006:77-86

[3] 朱群，张玉红，胡学钢，等. 一种基于双层窗口的概念漂移数据流分类算法[J]. 自动化学报，2011，37(9)：1077-1084

[4] 柴玉梅，周驰，王黎明. 数据流上概念漂移的检测和分类[J]. 小型微型计算机系统，2011，32(3)：421

[5] Vorburger P, Bernstein A. Entropy-based Concept Shift Deteciton [C]// Proc of the 6th International Conference on Data Mining. HongKong, Dec. 2006:1113-1118

[6] Street W N, Kim Y. A Streaming Ensemble Algorithm (SEA) for Large-scale Classification[C]// Proc of the 7th International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001:377-382

[7] Wang H, Fan W, Yu P S, et al. Mining concept-drifting data stream using ensemble classifiers[C]// Proc of the 9th international conference on Knowledge discovery and data mining. Washington D. C, ACM, 2003:226-235

[8] Ditzler G, Polikar R. An ensemble based incremental learning framework for concept drift and class imbalance[C]// International Joint Conference on Neural Networks. Barcelona, Spain, 2010:1-8

[9] Elwell R, Polikar R. Incremental Learning of Concept Drift in Nonsta- tionary Environments [J]. IEEE Transactions on neural networks, 2011, 22(10):1517-1531

[10] 孙岳，毛国君，刘旭，等. 基于多分类器的数据流中的概念漂移挖掘[J]. 电子学报，2008，34(1)：93-97

[11] 欧阳震诤，罗建书，胡东敏，等. 一种不平衡数据流集成分类模型[J]. 电子学报，2010，38(1)：184-189

[12] Krzanowski W J. Priniciple of Multivariate Analysis: A user's perspective [M]. Oxford Science Publications, 1993

[13] Kirkby R. Improving Hoeffding Trees [D]. New Zealand: University of Wakato, 2007

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