Journal of System Simulation

Volume 30 | Issue 6 Article 39

6-14-2018

Online Synthesis Incremental Data Streams Classification Algorithm

Sanmin Liu
1. College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China;;

Yuxia Liu
2. Center of Modern Education Technology, Anhui Polytechnic University, Wuhu 241000, China;

Online Synthesis Incremental Data Streams Classification Algorithm

Abstract

Abstract: Online learning is the effective way to solve the sample's non-recurrence in data streams classification, and how to deal with the problem of sample deficiency is the critical point for improving online learning efficiency. *According to the mean square error decomposition theory of the model's parameter estimation and the idea of cluster, the new samples are constructed by linear synthesis with the class center and the sample, which can improve the distribution information of sample and reduce the lower bound of parameter value. The online incremental learning is executed and the class center point is continuously updated.* Through theory analysis and simulation experiment, it is suggested that the provided schema is feasible and has superiority over other algorithm.

Keywords

online learning, data streams classification, cluster, incremental learning

第 30 卷第 6 期 系统仿真学报© Vol. 30 No. 6
2018 年 6 月 Journal of System Simulation Jun., 2018

在线合成增量式数据流分类算法

刘三民[1]，刘余霞[2]
(1. 安徽工程大学计算机与信息学院，安徽 芜湖 241000；2. 安徽工程大学现代教育技术中心，安徽 芜湖 241000)

摘要：在线学习是解决数据流分类挖掘样本不可再现性的有效手段，如何解决在线学习过程中样本量不足问题是提高在线学习质量的关键点。基于分类模型参数估计的均方误差分解理论，结合聚类思想利用类中心和样本线性合成样本，增加样本分布信息，降低参数估计的下界值；在此基础上进行在线合成增量学习，并根据样本系列信息不断修正类中心位置。经理论分析与仿真实验结果表明所提方案是有效的，在噪声环境内相比其他算法更具优势。

关键词：在线学习；数据流分类；聚类；增量学习
中图分类号：TP393 文献标识码：A 文章编号：1004-731X (2018) 06-2315-06
DOI: 10.16182/j.issn1004731x.joss.201806039

Online Synthesis Incremental Data Streams Classification Algorithm

Liu Sanmin[1], *Liu Yuxia*[2]
(1. College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China;
2. Center of Modern Education Technology, Anhui Polytechnic University, Wuhu 241000, China)

Abstract: Online learning is the effective way to solve the sample's non-recurrence in data streams classification, and how to deal with the problem of sample deficiency is the critical point for improving online learning efficiency. *According to the mean square error decomposition theory of the model's parameter estimation and the idea of cluster, the new samples are constructed by linear synthesis with the class center and the sample, which can improve the distribution information of sample and reduce the lower bound of parameter value. The online incremental learning is executed and the class center point is continuously updated.* Through theory analysis and simulation experiment, it is suggested that the provided schema is feasible and has superiority over other algorithm.

Keywords: online learning; data streams classification; cluster; incremental learning

引言

随着计算机软硬件和网络通信技术发展，各行业数据源源不断产生，如社交网络、传感网络和电子商务等，这种连续、无限的数据系列称为数据流。数据流中蕴含大量有价值的信息，如何对数据流进行分类挖掘提取相关信息成为当务之急[1]。

当前数据流分类处理模式有两种：批量处理和在线处理。批量处理是先存储数据、后分析处理模式，实现简单、准确率较高，技术相对成熟。文献[2-4]利用聚类、集成学习和增量学习思想构建批量处理模型，较好地解决数据流中概念漂移和噪声问题，实现数据分类功能。但某些场景受设备、环境限制无法大量缓存样本，具有不可再现性、要求能够及时分析处理，此时在线处理模式具有更好的适应性。Wang 等[5]提出过采样或欠采样的方法，

收稿日期：2017-07-13 修回日期：2017-09-07；
基金项目：国家自然科学基金(61300170，71371012)，安徽省自然科学基金(1608085MF147)，安徽省教育厅提升计划一般项目(TSKJ2016B05)；
作者简介：刘三民(1978-)，男，安徽岳西，博士，副教授，研究方向为机器学习。

并利用集成学习实现在线学习模型，解决了数据流中概念漂移和样本不平衡问题。文献[6]结合欠采样方法训练类子分类器，构建基于类子分类器的在线集成方法 CBCE，解决数据流中类概念演化现象。针对数据流中概念重现和概念漂移检测具有滞后性的特点，文益民等[7]提出一种在线迁移学习框架，其在线学习特点主要体现在选择源分类器方面，而并非传统的基于样本在线学习框架。文献[8]基于泊松分布逼近原理，建立在线装袋采样和自助采样方法，在此基础上构建在线集成学习模型，模型的有效性依赖于分布之间逼近的可靠性。文献[9]结合凸优化理论建立一种在线被动-主动学习算法(Online Passive-Aggressive Algorithm: PA)，该算法根据分类模型对当前样本预测损失情况，来寻找与当前超平面最相邻分类面作为最优分类面，该方法是基于单样本的在线学习模型，具有较强的学习能力。在此基础上，文献[10]把被动-主动学习算法用在大规模网络流量分类当中，并与其他算法进行对比，结果表明该方法具有较好的泛化能力。文献[11]针对具体应用场景利用聚类算法、结合增量集成学习方法实现在线学习方案，解决了量大、稀疏和数据流中概念周期性变化问题。

综上可知基于在线学习方法解决了数据流样本不可再现性和分类模型实时性问题，但在线学习模型的准确性易受到样本数量影响和制约。因此，在文献[12]对主流在线学习方案分析的基础上结合上述相关工作，本文提出一种新的在线样本合成增量学习方法。该方法借鉴聚类思想提出样本合成方案，解决在线学习时样本数量不足问题，而且该方案不需要保留历史样本以节约存储资源，同时克服样本不可再现问题。

1 在线被动-主动 PA 算法

为便于理解本节先对 PA 算法进行简单说明。PA 算法是基于单样本错误驱动在线学习模型，其目标是寻找满足条件且与当前分类器最相邻的超平面，作为新的分类模型。

数据流 $ds=\{(x_1,y_1),(x_2,y_2),\ldots,(x_t,y_t),\ldots\}$，其中 $(x_t,y_t)\in(R^d\times\{+1,-1\})$，$t$ 表示时间标签信息。基于此数据流信息，PA 算法可表示成下述最优化问题。

$$\begin{aligned}&\arg\min_{w}\frac{1}{2}\|w-w_t\|^2\\&s.t.\ \ l_t(w;(x_t,y_t))=0\end{aligned}\tag{1}$$

式中：w 表示所寻找最优分类面的法向量；w_t 表示 t 时刻分类面；公式(1)中的损失函数 $l_t=\max\{0,1-y_t(w_t\cdot x_t)\}$。利用拉格郎日函数上式可写成如公式(2)所示。

$$L(w,\tau)=\frac{1}{2}\|w-w_t\|^2+\tau(1-y_t(w.x_t))\tag{2}$$

式中：τ表示拉格朗日乘子，对 w 求偏导可得：$w=w_t+\tau y_t x_t$。把结果代入公式(2)可得目标函数的对偶形式。

$$L(\tau)=-\frac{1}{2}\tau^2\|x_t\|^2+\tau(1-y_t(w_t.x_t))\tag{3}$$

对公式(3)求导，可得：

$$\tau=\frac{1-y_t(w_t.x_t)}{\|x_t\|^2}=\frac{l_t}{\|x_t\|^2}\tag{4}$$

由式(4)可得出分类超平面参数 w 的更新计算方法。

$$w=w_t+\frac{l_t}{\|x_t\|^2}y_t x_t\tag{5}$$

在原文献中针对τ的两种改进计算方法，分别形成 PA-I、PA-II 在线学习方法。

$$\begin{aligned}&\tau_t=\min\{C,\frac{l_t}{\|x_t\|^2}\}\ \ \text{(PA-I)}\\&\tau_t=\frac{l_t}{\|x_t\|^2+\frac{1}{2C}}\ \ \text{(PA-II)}\end{aligned}\tag{6}$$

在公式(6)中 C 表示惩罚系数，根据经验值进行设定。根据内积特性，容易形成基于核函数的非线性分类模型，相应的预测函数如(7)式所示。

$$\hat{y}=sign(\sum_{i=1}^{t-1}\tau_i y_i K(x_t,x_i))\tag{7}$$

在式(7)中 $K(x_t,x_i)$表示核函数。

2 在线合成增量学习算法模型

基于单样本的在线学习方法存在样本量少，难

DOI: 10.16182/j.issn1004731x.joss.201806039

以全面反映样本分布信息。因此，文中结合聚类思想实现一种在线样本合成的增量学习方法，并通过模型参数有效性分析从理论上说明样本合成的必要性。

2.1 模型参数有效性分析

在衡量模型参数有效性时可采用均方误差分解方法进行评估，如公式(8)所示。

$$M^2(\tilde{\theta}) = E\{(\tilde{\theta}-\theta)^2\} = \mathrm{Var}(\tilde{\theta}) + \mathrm{b}^2(\tilde{\theta}) \tag{8}$$

式中：$Var(\tilde{\theta})$ 表示参数 $\tilde{\theta}$ 的估计方差；$b(\tilde{\theta}) = E(\tilde{\theta}) - \theta$ 表示 $\tilde{\theta}$ 的估计偏差；θ表示未知的真实值。

根据 Fisher 信息不等式，均方误差值具有 Cramer-Rao 下界：

$$M^2(\tilde{\theta}) \geqslant \frac{[g'(\theta)]^2}{J(\theta)} \tag{9}$$

若参数估计无偏，即 $b(\tilde{\theta}) = E(\tilde{\theta}) - \theta = 0$。则上述公式(8)和(9)可转化如公式(10)。

$$M^2(\tilde{\theta}) = Var(\tilde{\theta}) \geqslant \frac{1}{J(\theta)} \tag{10}$$

式中：$J(\theta)$表示 Fisher 信息，值越大表示数据集中反映参数θ的信息越多，此时得出的参数θ的估计值有更好的辨识度。当数据集中样本服从独立同分布，则参数θ的估计值的辨识度随样本数增多而增加。

数据流在线学习过程中，由于可能存在的概念漂移和每次处理一个样本的特点，难以满足上述理论结果要求。基于此，本文在尽可能减少额外负担的前提下(如存储历史样本或分类模型)，用类中心和当前样本信息合成新样本，在一定程度上保证样本之间的相似性。同时通过增加相似样本数量来提高对参数估计的辨识度，保证动态数据流环境下分类的准确性。

2.2 样本合成方法

由模型参数有效性理论分析可知，若能获取更多同分布样本将有助于提升参数估计的有效性。按聚类思想可知，同类样本分散在类中心的周围，样本之间具有更强的相似性。基于此，结合文献[13]构建出本文在线样本合成方法，提高在线增量学习时相似样本的数量。

若正类样本中心点为 c_+，当前数据流中 t 时刻到达的同类样本为 x_t，则新样本 x_{new} 合成计算如公式(11)所示。

$$x_{new} = r \times c_+ + (1-r) \times x_t \tag{11}$$

公式(11)表示在类中心和新到样本之间随机线性合成新样本，其中 $r \in (0,1)$为随机值。按此方法合成的新样本，保证新样本位于类中心与新到样本之间。由聚类思想可知合成的新样本具有较高概率与新到样本具有相同类别概念。增加了相似样本数量，提高 Fisher 信息量，保证在线增量学习过程中分类模型参数估计的有效性。

2.3 在线合成增量学习算法

本文的在线增量学习算法核心是融合聚类中心信息，依托数据流中同类样本信息，通过合成新样本进行在线增量更新分类模型。在算法执行过程中，每当分类模型错分数据流中样本时就触发上述新样本合成操作。

本文主要考虑二值分类情况，故所提在线合成增量式学习算法如下所示：

在线合成增量学习算法

输入：数据流样本

$ds=\{(x_1,y_1),(x_2,y_2),\ldots,(x_t,y_t),(x_{t+1},y_{t+1}),\ldots\}$，新样本合成个数值 num，分类正确样本数 count

输出：总体分类准确率

1) 计算初始正负类样本中心 c_+、c_-，并获取初始分类器 L；置变量 count=0

2) For t=1,2,…

2.1 获取真实样本标签 y_t

2.2 If $y_t=L(x_t)$ /*分类器 L 对样本 x_t 预测*/

count++

2.3 Else /*预测错误在线更新分类模型*/

For j=1, num /*合成 num 个新样本*/

产生随机数 r，按公式(11)进行样本合成，并利用新样本在线更新分类模型 L

2.4 结合 x_t 样本标签值更新对应类中心信息

3) 输出准确率

算法中步骤(3)输出的准确率是在线学习模型在数据流上预测整体准确率，分类模型训练采用贝叶斯学习算法。

3 仿真实验

为验证本文所提融合聚类思想的在线样本合成增量学习方法(Sample Synthesis based Online Incremental Learning: SSBOIL)的有效性，在公开的大型在线仿真平台 MOA 上生成移动超平面数据集[14]，与经典的在线被动-主动算法 PA 及其两种改进型算法(PA-I,PA-II)和传统基于增量式贝叶斯分类器的在线学习方法(Traditional Online Incremental Learning: TrOIL)进行比较，并以数据流上总体准确率作为评价指标。

3.1 移动超平面数据集

数据流由 d 维超平面 $\sum_{i=1}^{d} a_i x_i = a_0$ 构建，其中特征值 x_i 在区间(0,1)上随机取值，且 $a_0 = (\sum_{i=1}^{d} a_i)/2$。当 $\sum_{i=1}^{d} a_i x_i \geqslant a_0$，特征矢量$(x_1,x_2,\cdots x_d)$构成正类样本集；当 $\sum_{i=1}^{d} a_i x_i < a_0$，特征矢量$(x_1,x_2,\cdots x_d)$构成负样本集。在 MOA 中移动超平面数据流样本特征个数取值为 5，其中两个特征包含概念漂移，参数 m 取值 0.1 和参数 s 为 10%，按此设置生成两万条无噪声移动超平面数据集 h_0。为验证算法的抗噪能力，向数据集中随机注入 10%、20%、30%和 40%的噪声样本分别构成数据集 h_1、h_2、h_3 和 h_4。

3.2 仿真实验环境及结果分析

本文在线合成增量学习实验仿真环境是 Weka 和 Eclipse，学习器采用 Weka 内置的增量式贝叶斯学习器。在线被动-主动学习 PA 算法在 MATLAB 中完成，其中的核函数采用高斯核，相关参数均按默认值设置。传统基于增量式贝叶斯分类器的在线学习方法是直接利用 Weka 中的增量式贝叶斯学习器进行在线学习，与本文所提算法的区别是没进行样本合成操作。

在具体实验操作过程中，先在少量初始样本系列中训练分类器，主要是避免可能出现前面样本均为同类样本时出现的严重不平衡学习现象。对比实验一个关键参数是合成个数大小问题，在对 20 以内的值测试后，综合衡量选择合成个数 5 作为后续实验参数值。该值设置过大与批量处理相似，过小可能达不到想要的结果，所以针对不同的实验数据环境该值是一个经验值。

实验一：对比无噪声数据集 h_0 上准确率

5 种不同的在线增量学习方法首先在无噪声数据集 h_0 上进行实验对比，仿真实验的准确率曲线如图 1 所示。

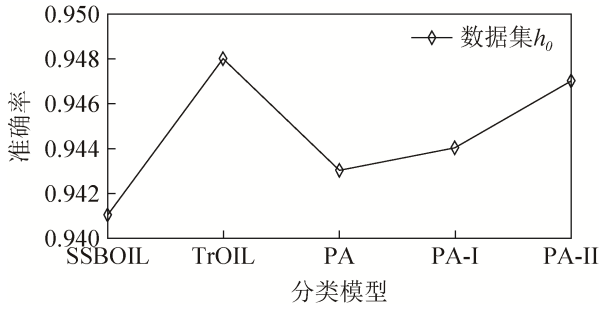

图 1 无噪声数据集上准确率曲线

Fig. 1 Curve of accuracy on noise-free data set

分析图 1 纵轴准确率值可知，文中所提的融合聚类中心在线合成增量式学习算法是可行的，其准确率值达 94%左右。纵观图 1 可得，5 种基于错误驱动的在线增量学习途径能够跟踪到数据流中的概念变化并及时调整分类模型，保证与数据流样本系列中的概念保持一致。但相比其他 4 种增量学习方案而言，本文所提方案的实验准确率值稍低，这是由于在数据流样本系列中概念漂移现象较明显时，通过文中所提的线性合成新样本方法并不能完全充分地满足模型参数估计有效性的理论基础，即样本独立同分布特性。

实验二：对比噪声环境下分类准确率

由于数据流产生环境的复杂性，数据流样本系

DOI: 10.16182/j.issn1004731x.joss.201806039

列中不仅有概念漂移现象，还存在着噪声问题，是当前数据流分类研究中面临的主要问题之一。为验证本文所提的在线合成增量学习算法 SSBOIL 的性能，在 4 种不同噪声数据集上进行仿真实验，并与其他几种在线学习方法进行对比，实验结果如表 1 所示。

表 1 噪声数据集上分类准确率
Tab. 1 Accuracy on noisy dataset

实验方案	数据集			
	h_1	h_2	h_3	h_4
SSBOIL	0.84	0.705	0.602	0.524
TrOIL	0.748	0.63	0.56	0.514
PA	0.739	0.617	0.546	0.507
PA-I	0.796	0.679	0.588	0.52
PA-II	0.763	0.642	0.565	0.509

从表 1 中的准确率值可知，在具有噪声的数据流环境里，本文所提的在线合成增量式学习算法明显具备一定的优势。在含噪声数据流样本集 h_1、h_2(含噪比低于 30%)，本文算法的分类准确率要比其他方法明显要好，准确率最大相差 10%左右；而在高更噪声比的环境(数据集 h_3、h_4)，SSBOIL 也表现出一定的优势。这是由于经典的在线增量式学习方法均是基于错误样本驱动，强调对错分样本的关注。故仅用噪声样本进行在线学习(如方案 TrOIL、PA 算法及 PA 两种改进算法)，容易误导分类面的移动方向，使其尽可能靠近噪声样本而偏离正确位置，产生较大误差。本文的融合聚类中心样本合成方法，充分考虑到了类别历史平均分布信息，减少噪声样本对分类模型的影响；同时通过在线合成样本，根据分类模型参数估计有效性的误差分解理论可知，增加相似样本数量时，能够提高分类模型参数估计的有效性，这为 SSBOIL 的性能的性能在噪声环境下占据优势提供理论依据。

4 结论

本文所提融合聚类中心的在线样本合成增量学习算法切实可行，解决了在线学习样本数不足问题，同时在较复杂的噪声环境中相比其他学习方法更具优势。相比其他的增量学习或在线学习方案，不需要额外存储资源来保存数据流中的历史样本或者历史分类模型信息，而且通过类中心点的统计信息解决样本不可再现性难题。通过实验总结可知，数据流在线增量学习过程中，如何避免数据流系列中样本的不平衡问题是提高分类模型预测准确性的关键，是下一步研究在线增量学习的重点。针对数据流在线学习过程中，有效区分概念漂移和噪声样本值得关注，同时探索合成样本个数大小与其之间的关系也可进一步深入研究。

参考文献:

[1] 孙大为, 张广艳, 郑纬民. 大数据流式计算: 关键技术及系统实例[J]. 软件学报, 2014, 25(4): 839-862.
Sun D W, Zhang G Y, Zheng W M. Big data stream computing: Technologies and instances[J]. Journal of Software, 2014, 25(4): 839-862.

[2] DM Farid, Z Li, A Hossain, et al. An adaptive ensemble classifier for mining concept drifting data streams[J]. Expert Systems with Applications (S0957-4174), 2013, 40(15): 5895-5906.

[3] Peng Zhang, Xingquan Zhu, Yong Shi, et al. Robust ensemble learning for mining noisy data streams[J]. Decision Support Systems (S0167-9236), 2011, 50: 469-479.

[4] Peipei Li, Xindong Wu, Xuegang Hu, et al. concept-drifting data streams with random ensemble decision trees[J]. Neurocomputing (S0925-2312), 2015, 166: 68-83.

[5] Shuo Wang, Leandro L M, Xin Yao. Resampling based ensemble methods for online class imbalance learning[J]. IEEE Transactions on Knowledge and Data Engineering (S1041-4347), 2015, 27(5): 1356-1367.

[6] Yu Sun, Ke Tang, LL Mink, et al. Online ensemble learning of data streams with gradually evolved classes[J]. IEEE Transactions on Knowledge and Data Engineering (S1041-4347), 2016, 28(6): 1532-1545.

[7] 文益民, 唐诗淇, 冯超, 等. 基于在线迁移学习的重现概念数据流分类[J]. 计算机研究与发展, 2016, 53(8): 1781-1791.
Wen Yimin, Tang Shiqi, Feng chao, et al. Online transfer learning for mining recurring concept in data stream classification[J]. Journal of Computer Research and Development, 2016, 53(8): 1781-1791.

(下转第 2327 页)