

# 学界 | 南京大学周志华等提出DFOP算法：无分布一次通过学习

2017年06月10日 13:15:14 机器之心

0| | | |

选自arXiv

机器之心编译

作者：赵鹏、周志华

参与：吴攀、黄小天

在线机器学习应用中，数据总是会随时间增多，怎么开发能有效应对这种动态情况的算法是一个值得关注的热门研究主题。近日，南京大学研究者赵鹏和周志华在 arXiv 发布了一篇题为《Distribution-Free One-Pass Learning》的论文，提出了一种有望解决这一问题的算法 DFOP。机器之心对该论文进行了摘要介绍，更多详情请参阅原文。

论文：无分布一次通过学习（Distribution-Free One-Pass Learning）

论文地址：<https://arxiv.org/abs/1706.02471>

## Distribution-Free One-Pass Learning

Peng Zhao, Zhi-Hua Zhou\*


*National Key Laboratory for Novel Software Technology  
Nanjing University, Nanjing 210093, China*

在许多大规模机器学习应用中，数据会随着时间而累积，因此，一个合适的模型应当能以一种在线的范式而进行更新。此外，因为在构建模型时，总的数据量是未知的，因此我们希望使用独立于数据量的存储来对每个数据项进行仅一次的扫描。另外值得注意的是在数据累积过程中，其基础分布可能会发生改变。为了应对这样的任务，在这篇论文中，我们提出了 DFOP——无分布一次通过学习方法（distribution-free one-pass learning approach）。这种方法在数据累积过程中分布发生变化时效果良好，且无需有关该变化的先验知识。每个数据项一旦被扫描后就可以被抛弃了。此外，理论保证（theoretical guarantee）也表明在一个轻微假设下的估计误差（estimate error）会下降，直到高概率地收敛。我们通过实验验证了 DFOP 在回归和分类上的表现。

### 3 预备工作

这一部分简要介绍了静态场景中的流回归模型（streaming regression model）。

在一个流场景（streaming scenario）中，我们用  $\{x(t), y(t)\}$  表示一个有标签的数据集，其中  $x(t)$  是第  $t$  个实例的特征， $y(t)$  是一个实值输出。此外，我们假设了一个如下的线性模型：

 **机器之心**

专业的人工智能媒体与产业服务平台。

### 热文排行

- 日榜周榜月榜
- 1 连马云都不见 却被马化腾请出来了
  - 2 为什么自助餐不会赔钱？餐厅老板自曝行..
  - 3 中国人被骗几十年，吹上天的德国制造，..
  - 4 拒了马化腾，看不上马云，雷军的世界首..
  - 5 新华社记者：我为什么守在一个买不到防..
  - 6 房产登记在孩子名下，6大隐患你可能想...
  - 7 房子、票子、位子都不要！就要个妹子，..
  - 8 他靠揉面团赚了1个亿！拿下世界冠军，...
  - 9 10万美元就能买下美加两国国籍，还送一..
  - 10 重磅：未来八年中国房价翻倍铁定板上钉..



$$y(t) = \mathbf{x}(t)^T \mathbf{w}(t-1) + \epsilon(t)$$

(1)

其中  $\{\epsilon(t)\}$  是噪声序列， $\{\mathbf{w}(t)\}$  是我们估计的。

当在一个静态场景中时，序列  $\{\mathbf{w}(t)\}$  是一个常数向量，用  $\mathbf{w}_0$  表示。然后，可以采用最小二乘法来最小化其残差平方和，其有一个闭式解（closed-form solution）。但是，当添加一个在线的/一次通过的约束（其要求原始项在被处理之后就被抛弃）时，它就无法工作。递归最小二乘法（RLS/recursive least square）和随机梯度下降（SGD）是以在线的范式解决这一问题的两种经典方法。

当在一个非静态环境中时，尤其是基础分布改变时，传统的方法是不合适的，因为我们永远不期望经典的 i.i.d 假设还能继续发挥效用。在下一节，我们提出了基于指数遗忘机制（exponential forgetting mechanism）来处理这种场景，而无需对数据流的演化进行明确的建模；我们也给出了理论支持和实证论证。

在下面， $\|\cdot\|$  表示在

$$\mathbb{R}^n$$

空间中的 L2 范数。同时，对于有界实值序列  $\{x(t)\}$ ， $x^*$  表示该序列的上界，即

$$x^* = \sup_{t=1,2,\dots} x(t).$$

#### 4 无分布一次通过学习

因为序列  $\{\mathbf{w}(t)\}$  在动态环境中会随时间改变，所以使用前面介绍的方法来估计当前（即时间  $t$  时）概念。相反，我们引入了一个贴现因子（discounted factors） $\{\lambda(t)\}$  序列来对旧数据的损失降权，如下：

$$\hat{\mathbf{w}}(t) = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^t \left( \prod_{j=i+1}^t \lambda(j) \right) [y(i) - \mathbf{x}(i)^T \mathbf{w}]^2,$$

(2)

其中  $\lambda(i) \in (0, 1)$  是一个贴现因子，可以平滑地给更旧的数据加上更少的权重。如果我们把所有  $\lambda(i)$  都简化成一个常量  $\lambda \in (0, 1)$ ，那么就可以更直观地理解，则此时该函数就为：

$$\hat{\mathbf{w}}(t) = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^t \lambda^{t-i} [y(i) - \mathbf{x}(i)^T \mathbf{w}]^2,$$

(3)

数量

$$\mu \triangleq 1 - \lambda$$

被称为遗忘因子（forgetting factor）[Hay08]。遗忘因子的值实际上是过去条件的稳定性（stability of past condition）和未来演化敏感度（sensitivity of future evolution）之间的权衡。

需要指出，这个基于指数遗忘因子的遗忘机制也可以被看作是滑动窗口方法（sliding window approach）的某种程度的连续类比。带有足够小权重的旧数据或多或少可被看作是从窗口中排除的。更多关于窗口大小和遗忘因子的关系的讨论可见于第 5.4 节。

##### 4.1 算法

对于 (3) 中提出的优化问题，显然，通过取该函数的导数，我们可以直接得到其最优的闭式解：

$$[\mathbf{w}(t)]_{opt} = \left[ \sum_{i=1}^t \lambda^{t-i} \mathbf{x}(i) \mathbf{x}(i)^T \right]^{-1} \left[ \sum_{i=1}^t \lambda^{t-i} \mathbf{x}(i) y(i) \right] \tag{4}$$

但是，上述表达式是一个离线的估计（off-line estimat），亦即 t 之前的所有数据项都需要。我们没有重复求解 (4)，而是基于新进入数据项的信息为之前的估计增加了一个校正项，从而对其基础概念（underlying concept）进行估计。使用遗忘因子递归最小二乘法 [Hay08]，我们可以以一次通过的范式（one-pass paradigm）求解目标 (3)。而就我们所知，这是第一次采用传统的遗忘因子 RLS 来在一次通过的约束条件下解决这样的带有分布改变的任务。我们将其命名为 DFOP（Distribution-Free One-Pass 的缩写），并总结为如下算法 1：

**Algorithm 1** Distribution Free One-Pass Learning

**Input:** A stream of data with  $\{\mathbf{x}(t), \{y(t)\}_{t=1 \dots T}\}$ , forgetting factor  $\mu \in (0, 1)$ ;

**Output:** Prediction  $\{\hat{y}(t)\}_{t=1 \dots T}$  (real value for regression and discrete-value for classification).

Initialize  $P_0 > 0$ ;

**for**  $t = 1$  **to**  $T$  **do**

$P(t) = \frac{1}{1-\mu} \left\{ P(t-1) - \mu \frac{P(t-1)\mathbf{x}(t)\mathbf{x}(t)^T P(t-1)}{1-\mu+\mathbf{x}(t)^T P(t-1)\mathbf{x}(t)} \right\}$ ;

$L(t) = P(t)\mathbf{x}(t)$ ;

$\hat{\mathbf{w}}(t) = \hat{\mathbf{w}}(t-1) + \mu L(t)[y(t) - \hat{\mathbf{w}}(t-1)^T \mathbf{x}(t)]$ ;

$\hat{y}(t) = \hat{\mathbf{w}}(t)^T \mathbf{x}(t)$ .      // for regression;

$\hat{y}(t) = \text{sign}[\hat{\mathbf{w}}(t)^T \mathbf{x}(t)]$ . // for classification

**end for**

另外，应该指明， $\{\lambda(t)\}$  被选作常量无论如何都是必要的，我们为动态贴现因子序列  $\{\lambda(t)\}$  提供了一个泛化的 DFOP（缩写为 G-DFOP），对应于 (11) 式，其也被证明是一个一次通过（one-pass）算法。第 1 节的详细证明参阅补充材料。

对于分类场景， $y(t)$  不再是一个实值输出，而是一个离散值，出于方便我们假设  $y(t) \in \{-1, +1\}$ 。在分类时，会在原来的输出步骤上进行一点细微的调整，其效果在下一节中通过实验获得了验证。

假设特征是 d 维的，在算法处理步骤中我们只需要记住：

$$P(t) \in \mathbb{R}^{d \times d}$$

易言之，存储总是  $O(d^2)$ ，其与训练实例的数量无关。此外，在第 t 时间戳（time stamp）时， $\hat{\mathbf{w}}(t)$  的更新也与先前的数据项不相关，即每一个数据项一旦被扫描，即被舍弃。

4.2. 理论保证

这一节中，我们在一个非平稳回归场景中开发了一个估计的误差界（error bound）。

5. 实验

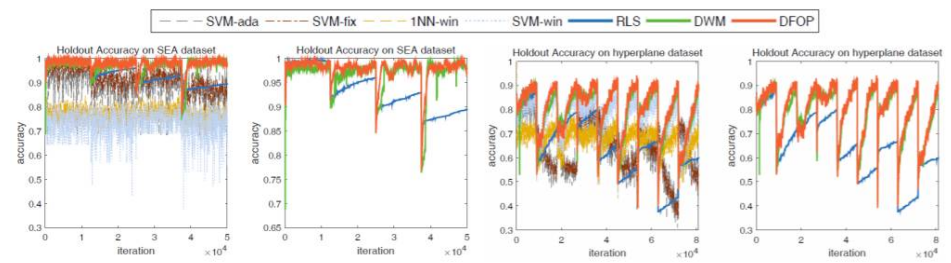


图 1: 在合成数据集上，7 种方法在 holdout 精确度方面的表现对比。左边是全部的 7 种方法；为了清晰，右边只绘制了 RLS、DWM 和 DFOP。

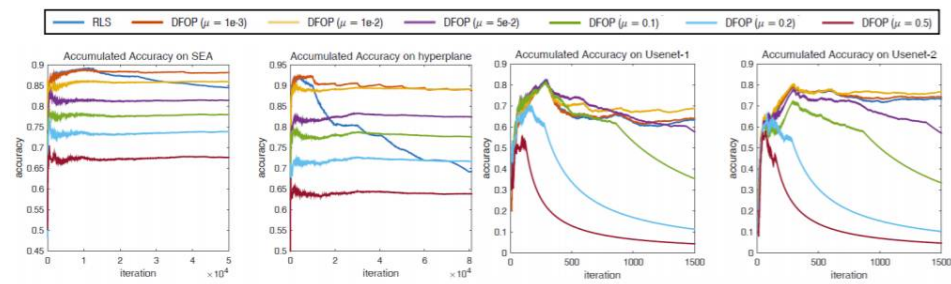


图 2：在带有分布变化的 4 个数据集上使用不同的遗忘因子的累积精确度



本文为机器之心编译，转载请联系本公众号获得授权。

加入机器之心（全职记者/实习生）：[hr@jiqizhixin.com](mailto:hr@jiqizhixin.com)

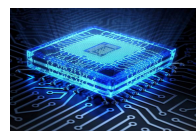
投稿或寻求报道：[editor@jiqizhixin.com](mailto:editor@jiqizhixin.com)

广告商务合作：[bd@jiqizhixin.com](mailto:bd@jiqizhixin.com)

点击阅读原文，查看机器之心官网↓↓↓

作者历史文章


研学社·系统组 | 实时深度学习的推理加速和持续训练



机器之心原创作者：YanchenWang参与：蒋思源、李亚洲作者Yanchen毕业于普林斯顿大学机器学习方向，现就职于微软Redmond总部，从事大规模分布式机[详细]

2017年 06月21日 10:45


教程 | 如何使用JavaScript构建机器学习模型



选自：hackernoon作者：AbhishekSoni参与：李泽南目前，机器学习领域建模的主要语言是Python和R，前不久腾讯推出的机器学习框架Angel则[详细]

2017年 06月21日 10:45


## 李飞飞高徒Andrej Karpathy加盟特斯拉，担任人工智能与自动驾驶视觉总



选自TechCrunch机器之心编译今日，特斯拉宣布前OpenAI研究员、斯坦福大学博士生AndrejKarpathy担任特斯拉人工智能和自动驾驶视觉总监（Di[详细]

2017年 06月21日 10:45


## 一个模型库学习所有：谷歌开源模块化深度学习系统Tensor2Tensor



选自Google.research机器之心编译参与：黄小天、李泽南在谷歌提交热点论文《AttentionIsAllYouNeed》和《OneModelToLea[详细]

2017年 06月20日 12:15


## 资源 | 斯坦福CS231n Spring 2017详细课程大纲（附完整版课件下载）



选自Stanford机器之心编译参与：Smith、蒋思源CS231n近几年一直是计算机视觉领域和深度学习领域最为经典的课程之一。而最近才刚刚结课的CS231nS[详细]

2017年 06月20日 12:15


## 全球人工智能黑客马拉松一触即发，北京站 15 支战队首曝光！



全球最大规模黑客松，北京，上海，旧金山，西雅图，纽约，巴黎，东京，温哥华，多伦多，慕尼黑，伦敦，阿姆斯特丹.....全球15个城市同步上演，世界Geek们的狂欢盛宴！[详细]

2017年 06月20日 12:15

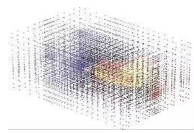
## 学界 | SIGIR2017满分论文：IRGAN



「每周一起读」是由PaperWeekly发起的协同阅读小组。我们每周精选一篇优质好文，利用在线协同工具进行精读并发起讨论，在碎片化时代坚持深度阅读。目前已成立的[详细]

2017年 06月20日 12:15


## 教程 | 深度学习初学者必读：张量究竟是什么？



选自Kdnuggets作者：TedDunning机器之心编译参与：晏奇、吴攀今天很多现有的深度学习系统都是基于张量代数（tensoralgebra）而设计的，但[详细]

2017年 06月20日 12:15

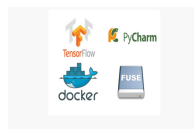
## 资源 | 谷歌全attention机器翻译模型Transformer的TensorFlow实现



选自GitHub机器之心编译参与：黄小天、Smith谷歌前不久在arXiv上发表论文《AttentionIsAllYouNeed》，提出一种完全基于attent[详细]

2017年 06月19日 12:45

## 我的深度学习开发环境详解：TensorFlow + Docker + PyCharm等，你的



选自Upflow.co作者：Killian机器之心编译参与：NurhachuNull、李亚洲在这篇文章中，研究员Killian介绍了自己的深度学习开发环境：Te[详细]

2017年 06月19日 12:45

- 1
- 2
- 3
- 4
- 5
- 
-