使用多目标回归模型预测用户访问网站的倾向性*

侯建鹏，郭天佑，李萍，徐君，程学旗
中国科学院计算技术研究所，北京市，100190
{houjianpeng,guotianyou,liping}@software.ict.ac.cn, {junxu, cxq}@ict.ac.cn

摘　要：根据用户的上网记录预测用户未来对不同网站的访问倾向性，可以为用户建模和网站优化提供依据，现有方法大多将此问题归结为基于用户-网站对的单目标回归问题进行解决。然而单目标回归模型忽略了同一用户访问不同网站的行为间的关联关系，因而影响了预测精度。本文基于梯度提升模型(Gradient Boosting Machine)[4]提出了一种新的多目标回归模型：在训练过程中，算法通过最优化用户级别损失函数得到所有的模型参数，因而有效地建模了用户访问不同网站的行为间的关联关系；在预测过程中，对于相同用户访问不同网站的倾向性采用不同的预测函数进行预测。基于真实的商业视频网站用户访问数据集合上的实验表明，与现有的单目标回归方法相比较，本文所提出的方法显著提升了预测精度。本文所提出的方法在中国电信举办的大数据算法应用大赛中获得第 1 名。
关键词：倾向性预测；多目标回归；梯度迭代；机器学习

Predicting User Preferences on Accessing the Websites with Multi-Target Regression

Jianpeng Hou, Tianyou Guo, Ping Li, Jun Xu, and Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
E-mail: {houjianpeng,guotianyou,liping}@software.ict.ac.cn, {junxu, cxq}@ict.ac.cn

Abstract: Predicting user's preferences on accessing different websites based on historical user behavior data is an important task, as it is helpful for user profiling and website optimization etc. Usually the task is formulated as a single-target regression problem in which each user-website is considered as an indepdent instance. However, the formulation does not take the relationship among the behaviors of a user when she accessing different websites. In this paper, we propose to alleviate the problem wieth a multi-target regression model based on gradient boosting machine. Speicifically, in the test phase, the model use different functions to predict a user's preferences on accessing different websites. In the training phase, a user level loss function is constructed and optimized for achieving the model parameters. In this way, the relationship of the activities when a user accesing different website is modeled effectively because the parameters in different functions are estimated simultanously. The experimental results on a real dataset show that the proposed model can significatnly outperform the baseline of single-target regression model. The model won the first prize in the big data algorithm applicaiton contest organized by China Telecom.
Keywords: Preference prediction; Multi-target regression; Gradient boosting machine

1　引言

通过访问互联网网站获取信息已成为人们获取各类信息资源的重要手段，一般来说用户访问不同的网站的行为具有很大的随机性，如何对用户访问网站的行为进行建模，尤其是准确预测用户未来访问不同网站的倾向性成为了一个难题，对此问题的解决可以为用户

*单击此处输入基金或资助机构的名称（项目编号），或删去此行

画像和用户行为预测提供基础，也为网站的优化提供依据，因此具有重要意义。

在现有解决方案中，机器学习中的监督学习方法被广泛应用，其依据用户的历史行为数据训练预测模型并基于训练好的模型预测用户未来对不同网站访问的倾向性。通常来说，问题被抽象成一个基于用户-网站对的单目标回归问题，即将每一个不同的用户-网站对看成不同的数据点，用同一个回归模型独立预测本用户对目标网站的访问倾向。

虽然上述方法对问题的建模简单直观，在实际应用系统中也取得了一定的成功，但是也存在着明显的缺陷：为了问题建模和模型训练的方便，其假设同一用户对不同网站的访问行为相互独立，因而同一用户对不同网站的访问倾向预测可以独立进行。在现实应用中这一假设有明显的不合理之处，比如当某用户查找某个视频时，该用户很可能去相关的多个视频网站上进行该视频的检索。

通过在真实的商业视频网站用户访问数据集合上对用户历史行为进行了统计分析我们发现，用户访问不同网站的次数间的皮尔逊相关系数较高，表明用户对不同网站的访问行为间确实存在一定的统计相关性。为了进一步提高预测精度，有必要在模型训练和预测的过程中，建模用户访问不同网站的行为间的关联关系。

为了解决上述问题，本文提出了 Multi-Target Gradient Boosting Machine（MTGBM）算法对用户访问多网站间的关联性进行建模，有效解决了上述问题。具体来说，MTGBM 算法将用户对多网站的访问倾向性预测问题抽象成一个多目标回归问题，每个用户对应多个预测目标（即该用户对不同网站的访问倾向性）。在训练的过程中，它通过引入用户级别失函数的方式，考虑了建立模型时不同目标变量之间的相互影响；在预测的过程中，该模型为每个用户预测输出多个目标变量（即用户对多网站的访问倾向性）。很明显，MTGBM 算法中的多元损失函数对用户访问不同网站的行为间的关联关系进行了建模，有效弥补了现有方法的不足之处。

为了验证 MTGBM 算法在网站访问倾向性预测的有效性，我们在真实的商业视频网站用户访问数据集合（包含 211,117 名用户，7,577,880 条访问记录）上进行了实验，预测用户对 10 个视频网站的访问倾向性，并与现有的方法进行了对比。实验结果表明，本文所提出的 MTGBM 多目标回归算法对用户未来行为预测的准确率要明显高于单目标回归方法。我们进一步将 MTGBM 算法运用到中国电信集团公司主办的大数据算法应用大赛（http://bdg.ctyun.cn/）上，击败了来自微软、百度等各企业和高校的 1,112 名参赛选手，获得第 1 名。

在本文的后续章节中，第 2 节介绍了现有方法在网站访问预测方面的相关研究；第 3 节介绍了 Gradient Boosting Machine 算法在多网站访问倾向性预测问题中的应用；第 4 节分析用户的多网站访问行为的特点；第 5 节详细叙述了本文的基于用户行为的多网站访问预测算法的实现细节；第 6 节介绍相关实验及结果；最后是对本文的总结及概述未来的研究方向和问题。

2 相关工作

现有方法在解决多网站访问预测问题时，通常将其归结为用户-网站对的单目标回归问题，即结合用户历史行为数据独立地预测用户在接下来一段时间内对不同网站的访问量，最常用到的模型有 AdaBoost[1-3]、Gradient Boosting Machine 以及 Random Forest[5,6]等。这些集成方法[8-10]将多个弱学习器通过某种策略结合在一起，从而得到比单一个体学习器更好的预测效果。

这些算法在回归问题场景中只能对单个目标变量建模，也就是说只适用于解决单目标

回归问题。在多网站访问倾向性预测问题中，由于对于每个用户来说需要预测的变量有多个，因此该问题属于多目标回归问题。现有方法在解决该问题时必须对其进行转化，从而获得单目标回归问题。通常有以下两种转化方式：独立建模和特征扩展。在独立建模的问题转化方式中，现有方法将根据网站 ID 对用户访问历史日志进行划分，获得多个互不相交的访问历史日志集合，并根据不同的日志集合对应建立多个不同的模型，最后使用这些模型来分别预测用户未来对多网站的访问情况。在特征扩展的问题转化方式中，现有方法将对网站 ID 进行 One-Hot 编码，并将编码结果作为特征加入到特征向量中去，从而对训练样本中分属不同网站的用户访问记录进行区分，从而建立一个可以预测不同网站访问情况的模型。

上述方法通过采用不同的转化问题的方式，都可以解决多网站访问预测问题。从本质上来讲，它们均对用户访问不同网站的行为进行了独立建模，即都是从解决单目标回归问题的角度进行出发，并未考虑用户访问多个网站时行为间的关联关系，即多目标变量之间的相关性，因而影响了预测精度。

近年来，基于用户历史行为的多网站访问倾向性预测问题因其包含的巨大潜在商业价值而得到相关企业的广泛关注。在 2015 年 11 月由中国电信集团公司举办的算法应用大赛中，参赛选手需要根据用户访问视频网站的历史数据，预测未来每个用户每天分别访问不同视频网站的情况。

3 基于单目标回归的多网站访问倾向性预测

在本节中，我们将以 Gradient Boosting Machine 算法为例，介绍现有方法如何基于单目标回归解决多网站访问倾向性预测问题。在训练的过程中，该算法对用户访问不同网站的行为分别建模，在预测的过程中，它使用不同的模型分别预测用户对不同网站的访问倾向性，即我们只需要使用 Gradient Boosting Machine 算法预测用户对单个网站访问倾向性。具体来说，令$\{(\mathbf{x}_i, y_i)\}_1^N$表示基于用户对单个网站的访问行为历史数据构建的训练样本，则用户对该网站的访问倾向度定义如下：

$$y_i = F_M(\mathbf{x}_i)$$

其中，$F_M(\mathbf{x})$是根据 Gradient Boosting Machine 算法迭代 M 轮后得到的固执函数。为了求解为了求解估值函数$F_M(\mathbf{x})$，Gradient Boosting Machine 算法需要在有限的训练样本$\{(\mathbf{x}_i, y_i)\}_1^N$上进行 M 轮迭代，每轮迭代过程中试图找到一个新的假设$h(\mathbf{x}; \mathbf{a}_m)$，从而在原先的假设$F_{m-1}(\mathbf{x})$基础上得到新的估值函数$F_m(\mathbf{x})$，使得损失更小，结果更加逼近真实答案。Gradient Boosting Machine 算法每轮迭代的策略如公式(1)所示：

$$(\beta_m, \mathbf{a}_m) = \arg\min_{\beta, \mathbf{a}} \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a})), m = 1, 2, .., M \tag{1}$$

根据公式(1)的计算结果更新估值函数，如公式(2)所示：

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{2}$$

但是，对于一些损失$L(y, F(\mathbf{x}))$和假设$h(\mathbf{x}; \mathbf{a})$来说，式(1)的求解会非常困难。为了得到这个问题的解，Gradient Boosting Machine 算法将损失函数对估值函数求导，得到$\{-g_m(\mathbf{x}_i)\}_1^N$作为训练样本$\{y_i, \mathbf{x}_i\}_1^N$的前进方向：

$$-g_m(\mathbf{x}_i) = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$$

在每一轮的迭代中，找到一个假设$h(\mathbf{x}; \mathbf{a}_m)$去拟合训练样本的前进方向：

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} \left[-g_m\left(\mathbf{x}_i\right) - \beta h\left(\mathbf{x}_i;\mathbf{a}\right)\right]^2 \tag{3}$$

这样，就将原先复杂的最优化问题(1)转化为容易求解的最小二乘问题(3)，最终只需要求解单变量的最优解即可：

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L\left(y_i, F_{m-1}\left(\mathbf{x}_i\right) + \rho h\left(\mathbf{x}_i;\mathbf{a}_m\right)\right), m = 1,2,..,M \tag{4}$$

根据式(3)、式(4)的推导结果，估值函数(2)采用如下方式进行更新：

$$F_m\left(\mathbf{x}\right) = F_{m-1}\left(\mathbf{x}\right) + \rho_m h\left(\mathbf{x};\mathbf{a}_m\right) \tag{5}$$

在预测问题中，过度拟合目标值往往会造成事与愿违的效果。因此，我们对式(5)引入正则化系数ν来防止过拟合现象的发生：

$$F_m\left(\mathbf{x}\right) = F_{m-1}\left(\mathbf{x}\right) + \nu\rho_m h\left(\mathbf{x};\mathbf{a}_m\right)$$

对于任意形式的损失L及假设h，Gradient Boosting Machine 算法都可以通过多次迭代计算逼近最优解，算法伪代码如算法 1 所示：

Algorithm 1 Gradient Boosting

Input: $\{(\mathbf{x_i}, y_i)\}, 1 \le i \le N$

Output: F_M

1: $F_0\left(\mathbf{x}\right) = \arg\min_{\rho} \sum_{i=1}^{N} L\left(y_i, \rho\right)$
2: **for** $m = 1$ **to** M **do**
3: $\quad \tilde{y}_i = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, i = 1, N$
4: $\quad \mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} \left[\tilde{y}_i - \beta h\left(\mathbf{x}_i;\mathbf{a}\right)\right]^2$
5: $\quad \rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L\left(y_i, F_{m-1}\left(\mathbf{x}_i\right) + \rho h\left(\mathbf{x}_i;\mathbf{a}_m\right)\right)$
6: $\quad F_m\left(\mathbf{x}\right) = F_{m-1}\left(\mathbf{x}\right) + \nu\rho_m h\left(\mathbf{x};\mathbf{a}_m\right)$
7: **end for**

算法 1：Gradient Boosting Machine 算法伪代码

4 多网站访问倾向性分析

虽然基于单目标回归的方式可以解决用户对多网站的访问倾向性预测问题，但是其背后的假设是用户对不同网站的访问行为相互独立，很明显在现实生活中这个假设并不总是成立。例如，当某用户希望观看某个视频时，该用户很可能去相关的多个视频网站上进行该视频的检索。

为了更加进一步探索这个问题，本文在真实的商业视频网站用户访问历史数据集合上进行统计分析。

我们统计了 1 周内各用户对 10 个网站的访问次数$\{\mathbf{s}_k\}_1^K$，其中$\mathbf{s}_k = \{s_{ik}\}_1^N$表示各用户对第$k$个网站的访问情况，并根据公式(6)计算用户访问 10 个网站行为之间的皮尔逊相关系数[14,15]。

$$\rho_{\mathbf{s}_i,\mathbf{s}_j} = \frac{cov\left(\mathbf{s}_i, \mathbf{s}_j\right)}{\sigma_{\mathbf{s}_i}\sigma_{\mathbf{s}_j}} = \frac{E\left[\left(\mathbf{s}_i - \mu_{\mathbf{s}_i}\right)\left(\mathbf{s}_j - \mu_{\mathbf{s}_j}\right)\right]}{\sigma_{\mathbf{s}_i}\sigma_{\mathbf{s}_j}} \tag{6}$$

计算结果如表 1 所示，依据表 1 中的数据可知，用户访问网站 1 和用户访问网站 5 的行为间存在弱相关关系（介于 0.2-0.4 之间），用户访问网站 7 和用户访问网站 10 的行为间存在中等程度的相关关系（介于 0.4-0.6 之间）。因而可以得出结论，对于某些网站，用户

访问它们倾向性确实存在着一定的关联

表 1 不同网站访问行为间的相关性统计

网站 ID	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	0.007	1.000								
3	0.121	0.003	1.000							
4	0.006	0.000	0.002	1.000						
5	**0.362**	0.006	0.067	0.006	1.000					
6	0.124	0.004	0.021	0.002	0.133	1.000				
7	**0.205**	0.065	0.042	0.003	**0.233**	0.139	1.000			
8	0.101	0.002	0.009	0.001	0.056	0.041	0.065	1.000		
9	0.136	0.003	0.034	0.003	0.099	0.054	0.044	0.050	1.000	
10	0.074	0.016	0.007	0.001	0.053	0.013	**0.519**	0.015	0.023	1.000

而现有方法在预测用户的未来访问倾向性时，并没有考虑用户行为的这一特点，因而影响了预测精度，现有方法还有提升空间。为此，我们对现有方法进行改进，提出了多目标回归模型。

5 本文所提出的方法：基于多目标回归的多网站访问倾向性预测

为了建模用户对不同网站访问行为间的关联关系，本文基于 Gradient Boosting Machine 提出一种多目标回归算法——Multi-Target Gradient Boosting Machine（MTGBM）。从第 3 节对 Gradient Boosting Machine 算法的介绍中我们可以发现，Gradient Boosting Machine 算法有一个显著的特点：损失函数对估值函数求导。本文利用该算法的这个特点，将 Gradient Boosting Machine 算法的单个估值函数扩展为多个估值函数并同时进行模型参数估计，从而达到了多目标回归的目的。

具体的，假设在过去一段时间内，用户i对K个网站的访问数据表示为：

$$\mathrm{D}_i = \{(\mathbf{x}_{i1}, y_{i1}), ..., (\mathbf{x}_{iK}, y_{iK})\}$$

其中$\mathbf{x}_{ik}$为描述用户i与网站k的特征向量，y_{ik}为用户i对网站k的倾向性标签（依据用户历史访问数据得出）。用$D = \{D_i\}_1^N$表示含有 N 个用户的训练数据。

网站访问倾向性预测的目标是基于历史训练数据 D，预测未来任意用户 i 对任意网站 k 的访问倾向性。用倾向性函数$F_k(\mathbf{x}_{ik})$描述用户i对网站k的访问倾向性，定义如下：

$$y_{ik} = F_k(\mathbf{x}_{ik})$$

其中，$F_k(\mathbf{x})$是 MTGBM 算法迭代M轮之后得到的估值函数，即$F_k(\mathbf{x}) = F_{Mk}(\mathbf{x})$。

与 Gradient Boosting Machine 算法类似，为了求解估值函数$\{F_{Mk}(\mathbf{x})\}_1^K$，MTGBM 算法需要进行 M 轮迭代。由于对于每个用户来说，需要预测的目标变量有多个$\{y_{ik}\}_1^K$，因此对损失函数进行扩展并以向量作为输入，得到$L(\mathbf{y}, \mathbf{F})$。在每轮迭代过程中需要找到$K$个假设$\mathbf{h}_m = \{h(\mathbf{x}, \mathbf{a}_{mk})\}_1^K$，并在原先的假设基础上得到新的估值函数$\mathbf{F}_m = \{F_{mk}(\mathbf{x})\}_1^K$，使得损失进一步减小。MTGBM 算法每轮迭代的策略如下所示：

$$(\beta_m, \mathbf{h}_m) = \arg\min_{\beta, \mathbf{h}} \sum_{i=1}^{N} L(\mathbf{y}_i, \mathbf{F}_{m-1} + \beta\mathbf{h}), m = 1, M \tag{7}$$

根据式(7)的计算结果更新估值函数，如下所示：

$$\mathbf{F}_m = \mathbf{F}_{m-1} + \beta_m \mathbf{h}_m \tag{8}$$

由于对于不同的损失函数及假设，式(7)的求解非常困难。因此，我们将上一轮的损失函数$L(\mathbf{y},\mathbf{F})$对$\{F_k(\mathbf{x})\}_1^K$分别求偏导，得到$\{-\mathbf{g}_{mi}\}_1^N$作为各样本点的前进方向，其中$-\mathbf{g}_{mi}=\{-g_{mk}(\mathbf{x}_{ik})\}_1^K$：

$$-g_{mk}(\mathbf{x}_{ik})=-\left[\frac{\partial L(\mathbf{y}_i,\mathbf{F})}{\partial F_k(\mathbf{x}_{ik})}\right]_{\mathbf{F}=\mathbf{F}_{m-1}}$$

在每一轮迭代中，找到K个假设$\mathbf{h}_m=\{h(\mathbf{x},\mathbf{a}_{mk})\}_1^K$去拟合训练样本的前进方向：

$$\mathbf{a}_{mk}=\arg\min_{\mathbf{a},\beta}\sum_{i=1}^{N}\left[-g_{mk}(\mathbf{x}_{ik})-\beta h(\mathbf{x}_{ik};\mathbf{a})\right]^2,k=1,K \tag{9}$$

这样，将原先复杂的最优化问题(7)转化成了K个最小二乘问题(9)，最终只需要求解单变量的最优解即可：

$$\rho_m=\arg\min_{\rho}\sum_{i=1}^{N}L(\mathbf{y}_i,\mathbf{F}_{m-1}+\rho\mathbf{h}_m),m=1,M \tag{10}$$

根据式(9)、式(10)的推导结果，估值函数更新方法如下所示：

$$\mathbf{F}_m=\mathbf{F}_{m-1}+\rho_m\mathbf{h}_m \tag{11}$$

在预测问题中，过度拟合目标值往往会造成事与愿违的效果。因此，我们对式(11)引入正则化系数ν来防止过拟合现象的发生，如下所示：

$$\mathbf{F}_m=\mathbf{F}_{m-1}+\nu\rho_m\mathbf{h}_m,0<\nu\leq 1$$

经过上述变化，对于多目标回归问题中任意的多元损失L及假设h，MTGBM 算法都可通过多元损失函数对多估值函数分别求偏导并每轮训练多个弱学习器的方式逼近最优解。该算法如算法 2 所示。

Algorithm 2 Multi-Target Gradient Boosting

Input: $\{\{(\mathbf{x}_{ik},y_{ik})\},1\leq i\leq N\},1\leq k\leq K$
Output: $\mathbf{F}_M$

$\mathbf{F}_0=\arg\min_{\vec{\rho}}\sum_{i=1}^{N}L(\mathbf{y}_i,\vec{\rho}),|\vec{\rho}|=1$
for $m=1$ **to** M **do**
  $\tilde{\mathbf{y}}_i=\left\{\left[\frac{\partial L(\mathbf{y}_i,\mathbf{F})}{\partial F_k(\mathbf{x}_{ik})}\right]_{\mathbf{F}=\mathbf{F}_{m-1}}\right\}_1^K,i=1,N$
  for $k=1$ **to** K **do**
    $\mathbf{a}_{mk}=\arg\min_{\mathbf{a},\beta}\sum_{i=1}^{N}\left[\tilde{y}_k-\beta h(\mathbf{x}_{ik};\mathbf{a})\right]^2$
  end for
  $\mathbf{h}_m=\{h(\mathbf{x},\mathbf{a}_{mk})\}_1^K$
  $\rho_m=\arg\min_{\rho}\sum_{i=1}^{N}L(\mathbf{y}_i,\mathbf{F}_{m-1}+\rho\mathbf{h}_m)$
  $\mathbf{F}_m=\mathbf{F}_{m-1}+\nu\rho_m\mathbf{h}_m$
end for

算法 2：MTGBM 算法伪代码

6 实验结果与分析

本节基于中国电信集团公司提供的用户访问视频网站的真实历史数据，通过两部分实验来验证本文提出的多网站访问预测方法的有效性，第 1 部分实验重点讨论不同的参数对结果的影响，第 2 部分利用被广泛使用的 Scikit-learn 工具[11,12]以及 XGBoost 工具[13]中的 AdaBoost、Gradient Boosting Machine、Random Forest 算法与 MTGBM 算法进行对比实

验，检验 MTGBM 算法相较于现有方法的提升效果。

6.1 实验环境

本文实验的环境是单台服务器，系统的配置为 6 核主频为 1.2GHz 的 GenuineIntel 处理器，128G 的 RAM 和 2TB 的硬盘，使用的操作系统版本为 CentOS release 6.6(Final)。

为了使实验结果具有说服力，同一组实验运行 10 次之后取平均值记为最终结果。

6.2 实验数据

实验数据采用中国电信集团公司提供的用户 7 周访问 10 个视频网站的行为数据。该数据以 0.5 小时为粒度统计，数据集的详细描述如表 2 所示。

表 2 实验数据集详细描述

属性	属性值
大小/MB	287
记录数目/条	7,577,880
用户数目/人	211,117
视频网站数目/个	10

实验数组集共包含 4 个字段，对数据集的详细格式描述如表 3 所示。

表 3 实验数据集格式描述

字段	类型	描述	样例
用户标识	String	用户唯一标识	4WFkFrSYjTs=
访问时间	String	widjhhmm，表示第 i 周第 j 天第 hh 小时第 mm 分开始的 30 分钟时段	w3d51900
视频网站标签	String	vi，表示第 i 个视频网站	v10
访问次数	Integer	在访问时间标识时段内访问该网站的次数	3

6.3 实验设计

中国电信集团公司提供的用户访问视频网站数据集共包含的时间段为 7 周。因此，我们将根据前 6 周用户访问 10 个视频网站数据预测用户第 7 周对 10 个视频网站访问的倾向性。

为了评估预测准确度，算法需要对每个用户第 7 周访问各视频网站的可能性进行打分，记为$\{p_i\}_1^{10}$，并以用户第 7 周对各视频网站的真是访问次数作为标准答案，记为$\{t_i\}_1^{10}$。

对于单个用户，预测的准确度计算如下：

$$precision_i = \frac{\sum_1^{10}(t_i \times p_i)}{\sqrt{\sum_1^{10} t_i^2} \times \sqrt{\sum_1^{10} p_i^2}}$$

最终得分计算如下：

$$precision = \sum_{i=1}^{n} precision_i$$

为了预测用户第 7 周对十个视频网站访问的倾向性，我们使用用户第 6 周对十个视频网站的访问次数构建训练数据，并从第 1 周到第 5 周的数据中抽取训练数据的特征向量。这样，每个用户可以构造一个实例$(\mathbf{X}_i, \mathbf{y}_i)$，其中$\mathbf{y}_i$是一个 10 维向量，分别对应用户对 10

个视频网站的访问次数。因此，原先的问题是一个多目标回归问题。

对于现有方法而言，训练数据的目标值必须为单个变量。因此，我们将每个用户构造的实例$(\mathbf{X}_i, \mathbf{y}_i)$拆解为$\{(\mathbf{x}_i, y_{ij})\}_1^{10}$的形式，并使用 10 个模型分别对 10 个网站进行预测。对于本文提出的 MTGBM 算法，训练数据的目标值不需要额外的处理，可直接输出向量。

在实验数据集上，我们共抽取用户访问各网站的相关特征 9 类，共计 30 维。具体描述如表 4 所示。

表 4 实验特征表

特征名	标识	维度	描述
平均间隔时间	f_1	1	各用户访问各网站的平均间隔天数
最近访问间隔	f_2	1	各用户访问各网站的最近间隔天数
平均值	f_3	1	各用户历史访问各网站次数的平均值
中位数	f_3	1	各用户历史访问各网站次数的中位数
方差	f_4	1	各用户历史访问各网站次数的方差
历史访问标识	f_5	5	分别统计前五周各用户对各网站是否有访问记录
历史访问次数	f_6	5	分别统计前五周各用户对各网站访问次数的总和
历史访问平均值	f_7	5	分别统计前五周各用户对各网站访问次数的平均值
历史访问中位数	f_8	5	分别统计前五周各用户对各网站访问次数的中位数
历史访问方差	f_9	5	分别统计前五周各用户对各网站访问次数的方差

在本次实验中，MTGBM 算法采用决策树作为弱学习器，同时损失函数设计如下：

$$L(\mathbf{y}, \mathbf{F}) = -\frac{\mathbf{y} \cdot \mathbf{F}}{\|\mathbf{y}\| \, \|\mathbf{F}\|}$$

6.4 实验结果

为了验证 MTGBM 算法相较于现有方法的性能提升，本小节首先对 AdaBoost, Gradient Boosting Machine 以及 Random Forest 等算法在相同的数据集上进行测试，并取其最佳结果与 MTGBM 算法的结果进行对比，测试结果如表 5 所示。

表 5 实验数据集详细描述

算法	Precision
MTGBM	**0.560119**
AdaBoost(from sklearn)	0.521356
Gradient Boosting Machine(from xgboost)	0.548356
Gradient Boosting Machine(from sklearn)	0.540148
Random Forest(from sklearn)	0.506730

通过比较 MTGBM 算法的测试结果与现有算法预测的最优测试结果可以看出，MTGBM 算法的预测结果相较于 AdaBoost、Gradient Boosting Machine 以及 Random Forest 等现有方法有显著的提升。

表 6 不同网站访问行为间的相关性统计

	1	2	3	4	5	6	7	8	9	10
真实倾向性	0.000	0.000	0.000	**0.447**	0.000	0.000	0.000	0.000	0.000	0.894
MTGBM	0.150	0.000	0.000	**0.308**	0.174	0.036	0.066	0.000	0.101	0.914
GBM	0.001	0.000	0.000	**0.003**	0.002	0.001	0.001	0.000	0.002	1.000

我们从 MTGBM 算法和 xgboost 中的 Gradient Boosting Machine（GBM）算法的最优运

算结果中抽取了某用户对 10 个网站访问倾向性的预测结果进行对比，如表 6 所示。从中可以看出，在考虑了用户访问行为的相关性之后，MTGBM 算法可以捕获该用户未来对 ID 为 4 的网站的访问倾向性，从而使得多网站访问倾向性的预测结果变得更加精准。

进一步观察 MTGBM 算法与现有 Gradient Boosting Machine 算法随迭代次数的变化情况，如图 1 所示。可以发现，MTGBM 算法的收敛速度略慢于 xgboost 中实现的 Gradient Boosting Machine 算法，但是最终结果优于任意一种 Gradient Boosting Machine 算法的实现在测试集上的最优结果，且收敛过程更加稳定。

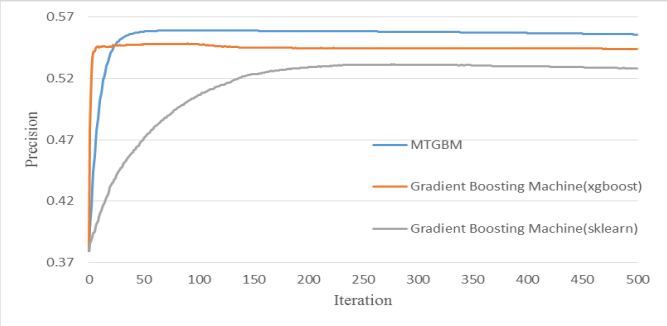

图 1 算法性能对比

为了观察 MTGBM 算法的不同参数对结果的影响，我们令 MTGBM 算法的正则化系数取不同大小的值$\nu \in \{1.0, 0.25, 0.125, 0.06\}$，在测试集上算法性能表现如图 2 所示。从中可以发现，过大的正则化系数会造成模型对训练集合的过拟合现象发生，从而降低其在测试集合上的效果，适当减小正则化系数的大小，可以避免这一现象的发生，同时会增加模型收敛所需要的时间。

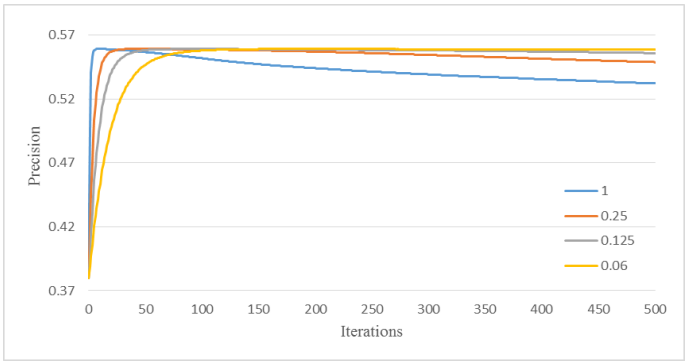

图 2 MTGBM 算法性能

为了观察 MTGBM 算法的不同参数对结果的影响，我们令 MTGBM 算法的正则化系数取不同大小的值$\nu \in \{1.0, 0.25, 0.125, 0.06\}$，在测试集上算法性能表现如图 1 所示。从中可以发现，过大的正则化系数会造成模型对训练集合的过拟合现象发生，从而降低其在测试集合上的效果，适当减小正则化系数的大小，可以避免这一现象的发生，同时会增加模型收敛所需要的时间。

6.5 实际应用

本文提出的 MTGBM 算法出色地解决了中国电信集团公司在 2015 年 12 月主办的大数据算法应用大赛上提出的用户上网行为预测问题，在 1,112 名来自微软、百度等各企业及高校的参赛选手中排名第 1。该赛题涉及 299,993 个用户的数据，数据量总计 25.38GB。Top 3 成绩如表 7 所示。

表 7 Top 3 排名及评分

排名	队长	所在团队	评分
1	HouJP	科院南路 6 号	28.9081%
2	Gloria	lala	28.8072%
3	dada	hhhh	28.7840%

7 总结与展望

基于用户历史行为的多网站访问预测问题是一个有重要实用价值的研究问题。以往的解决方案基于用户-网站对进行单目标回归预测，忽视了用户浏览不同网站行为间的相互影响，从而限制了预测精度的提升。本文研究了如何利用不同网站浏览行为间的相关性进行多网站访问预测。通过实验对比分析，我们发现基于用户行为的多网站访问预测算法要显著优于现有方法。

在后续工作中，我们考虑基于用户行为的多目标网站预测算法与网站特征相结合，进一步提升预测效果。

参 考 文 献

[1] Freund Y, Schapire R, Abe N. A short introduction to boosting[J]. Journal-Japanese Society For Artificial Intelligence, 1999, 14(771-780): 1612.

[2] Solomatine D P, Shrestha D L. AdaBoost. RT: a boosting algorithm for regression problems[C]//Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. IEEE, 2004, 2: 1163-1168.

[3] Shrestha D L, Solomatine D P. Experiments with AdaBoost. RT, an improved boosting scheme for regression[J]. Neural computation, 2006, 18(7): 1678-1710.

[4] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

[5] Ho T K. Random decision forests[C]//Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. IEEE, 1995, 1: 278-282.

[6] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[7] Ho T K. The random subspace method for constructing decision forests[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1998, 20(8): 832-844.

[8] Drucker H, Cortes C, Jackel L D, et al. Boosting and other ensemble methods[J]. Neural Computation, 1994, 6(6): 1289-1301.

[9] Opitz D, Maclin R. Popular ensemble methods: An empirical study[J]. Journal of Artificial Intelligence Research, 1999: 169-198.

[10] Dietterich T G. Ensemble methods in machine learning[M]//Multiple classifier systems. Springer Berlin

Heidelberg, 2000: 1-15.

[11] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. The Journal of Machine Learning Research, 2011, 12: 2825-2830.

[12] Prettenhofer P, Louppe G. Gradient Boosted Regression Trees in Scikit-Learn[J]. 2014.

[13] Chen T, He T. xgboost: eXtreme Gradient Boosting[J]. R package version 0.4-2, 2015.

[14] Benesty J, Chen J, Huang Y, et al. Pearson correlation coefficient[M]//Noise reduction in speech processing. Springer Berlin Heidelberg, 2009: 1-4.

[15] Sedgwick P. Pearson's correlation coefficient[J]. BMJ, 2012, 345: e4483.