

# CTR预估中GBDT与LR融合方案

原创 腾讯大数据 腾讯大数据 2015-08-27

## 1 背景

CTR预估，广告点击率（Click-Through Rate Prediction）是互联网计算广告中的关键环节，预估准确性直接影响公司广告收入。CTR预估中用的最多的模型是LR（Logistic Regression）[1]，LR是广义线性模型，与传统线性模型相比，LR使用了Logit变换将函数值映射到0~1区间[2]，映射后的函数值就是CTR的预估值。LR，逻辑回归模型，这种线性模型很容易并行化，处理上亿条训练样本不是问题，但线性模型学习能力有限，需要大量特征工程预先分析出有效的特征、特征组合，从而去间接增强LR的非线性学习能力。

LR模型中的特征组合很关键，但又无法直接通过特征笛卡尔积解决，只能依靠人工经验，耗时耗力同时并不一定会带来效果提升。如何自动发现有效的特征、特征组合，弥补人工经验不足，缩短LR特征实验周期，是亟需解决的问题。Facebook 2014年的文章介绍了通过GBDT（Gradient Boost Decision Tree）解决LR的特征组合问题[3]，随后Kaggle竞赛也有实践此思路[4][5]，GBDT与LR融合开始引起了业界关注。

GBDT（Gradient Boost Decision Tree）是一种常用的非线性模型[6][7][8][9]，它基于集成学习中的boosting思想[10]，每次迭代都在减少残差的梯度方向新建立一颗决策树，迭代多少次就会生成多少颗决策树。GBDT的思想使其具有天然优势，可以发现多种有区分性的特征以及特征组合，决策树的路径可以直接作为LR输入特征使用，省去了人工寻找特征、特征组合的步骤。这种通过GBDT生成LR特征的方式（GBDT+LR），业界已有实践（Facebook，Kaggle-2014），且效果不错，是非常值得尝试的思路。下图1为使用GBDT+LR前后的特征实验示意图，融合前人工寻找有区分性特征（raw feature）、特征组合（cross feature），融合后直接通过黑盒子（Tree模型GBDT）进行特征、特种组合的自动发现。

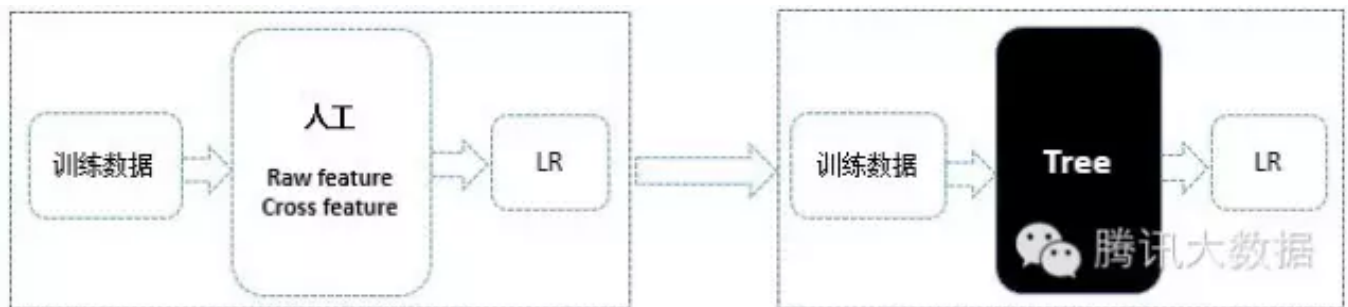


图1

## 2 GBDT与LR融合现状

GBDT与LR的融合方式，Facebook的paper有个例子如下图2所示，图中Tree1、Tree2为通过GBDT模型学出来的两颗树， $x$ 为一条输入样本，遍历两棵树后， $x$ 样本分别落到两颗树的叶子节点上，每个叶子节点对应LR一维特征，那么通过遍历树，就得到了该样本对应的所有LR特征。由于树的每条路径，是通过最小化均方差等方法最终分割出来的有区分性路径，根据该路径得到的特征、特征组合都相对有区分性，效果理论上不会亚于人工经验的处理方式。

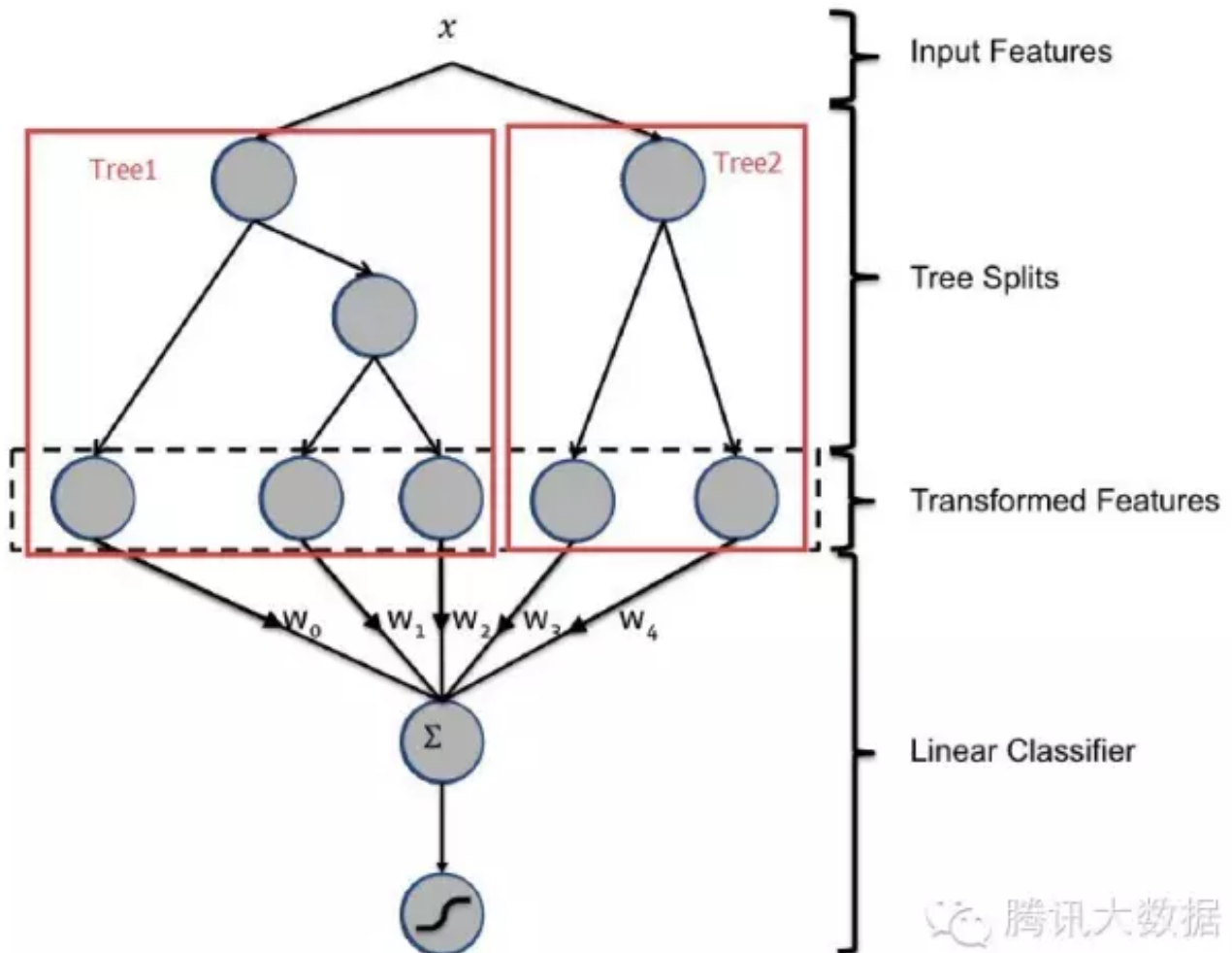


图2

GBDT模型的特点，非常适合用来挖掘有效的特征、特征组合。业界不仅GBDT+LR融合有实践，GBDT+FM也有实践，2014 Kaggle CTR竞赛冠军就是使用GBDT+FM，可见，使用GBDT融合其它模型是非常值得尝试的思路[11]。

笔者调研了Facebook、Kaggle竞赛关于GBDT建树的细节，发现两个关键点：采用ensemble决策树而非单颗树；建树采用GBDT而非RF（Random Forests）。解读如下：

### 1) 为什么建树采用ensemble决策树？

一棵树的表达能力很弱，不足以表达多个有区分性的特征组合，多棵树的表达能力更强一些。GBDT每棵树都在学习前面棵树尚存的不足，迭代多少次就会生成多少颗树。按

paper以及Kaggle竞赛中的GBDT+LR融合方式，多棵树正好满足LR每条训练样本可以通过GBDT映射成多个特征的需求。

## 2) 为什么建树采用GBDT而非RF?

RF也是多棵树，但从效果上有实践证明不如GBDT。且GBDT前面的树，特征分裂主要体现在对多数样本有区分度的特征；后面的树，主要体现的是经过前N颗树，残差仍然较大的少数样本。优先选用在整体上有区分度的特征，再选用针对少数样本有区分度的特征，思路更加合理，这应该也是用GBDT的原因。

然而，Facebook和Kaggle竞赛的思路是否能直接满足现在CTR预估场景呢？

按照Facebook、Kaggle竞赛的思路，不加入广告侧的AD ID特征？但是现CTR预估中，AD ID类特征是很重要的特征，故建树时需要考虑AD ID。直接将AD ID加入到建树的feature中？但是AD ID过多，直接将AD ID作为feature进行建树不可行。下面第三部分将介绍针对现有CTR预估场景GBDT+LR的融合方案。

## 3 GBDT与LR融合方案

AD ID类特征在CTR预估中是非常重要的特征，直接将AD ID作为feature进行建树不可行，故考虑为每个AD ID建GBDT树。但互联网时代长尾数据现象非常显著，广告也存在长尾现象，为了提升广告整体投放效果，不得不考虑长尾广告[12]。在GBDT建树方案中，对于曝光充分训练样本充足的广告，可以单独建树，发掘对单个广告有区分度的特征，但对于曝光不充分样本不充足的长尾广告，无法单独建树，需要一种方案来解决长尾广告的问题。

综合考虑方案如下，使用GBDT建两类树，非ID建一类树，ID建一类树。

**1) 非ID类树：**不以细粒度的ID建树，此类树作为base，即便曝光少的广告、广告主，仍可以通过此类树得到有区分性的特征、特征组合。

**2) ID类树：**以细粒度的ID建一类树，用于发现曝光充分的ID对应有区分性的特征、特征组合。

如何根据GBDT建的两类树，对原始特征进行映射？以如下图3为例，当一条样本x进来之后，遍历两类树到叶子节点，得到的特征作为LR的输入。当AD曝光不充分不足以训练树时，其它树恰好作为补充。

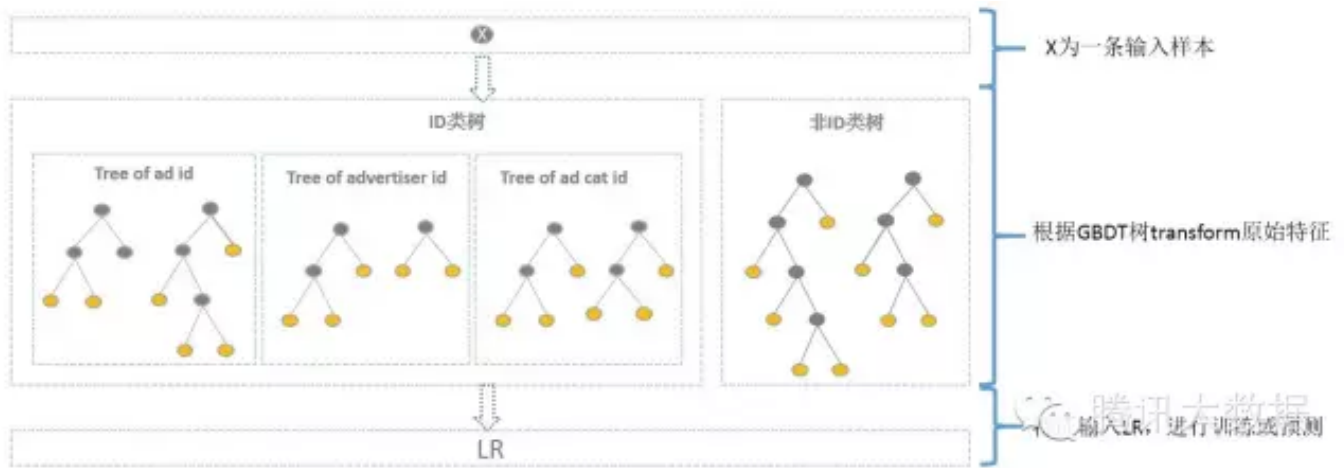


图3

### 通过GBDT 映射得到的特征空间维度如何？

GBDT树有多少个叶子节点，通过GBDT得到的特征空间就有多大。如下图4一颗树，一个叶子节点对应一种有区分性的特征、特征组合，对应LR的一维特征。这颗树有8个叶子节点，即对应LR的8维特征。估算一下，通过GBDT转换得到的特征空间较低，Base树、ID树各N颗，特征空间维度最高为 $N + N \times \text{广告数} + N \times \text{广告主数} + N \times \text{广告类目数}$ 。其中广告数、广告主数、广告类目数都是有限的，同时参考Kaggle竞赛中树的数目N最多为30，则估算通过GBDT 映射得到的特征空间维度并不高，且并不是每个ID训练样本都足以训练多颗树，实际上通过GBDT 映射得到的特征空间维度更低。

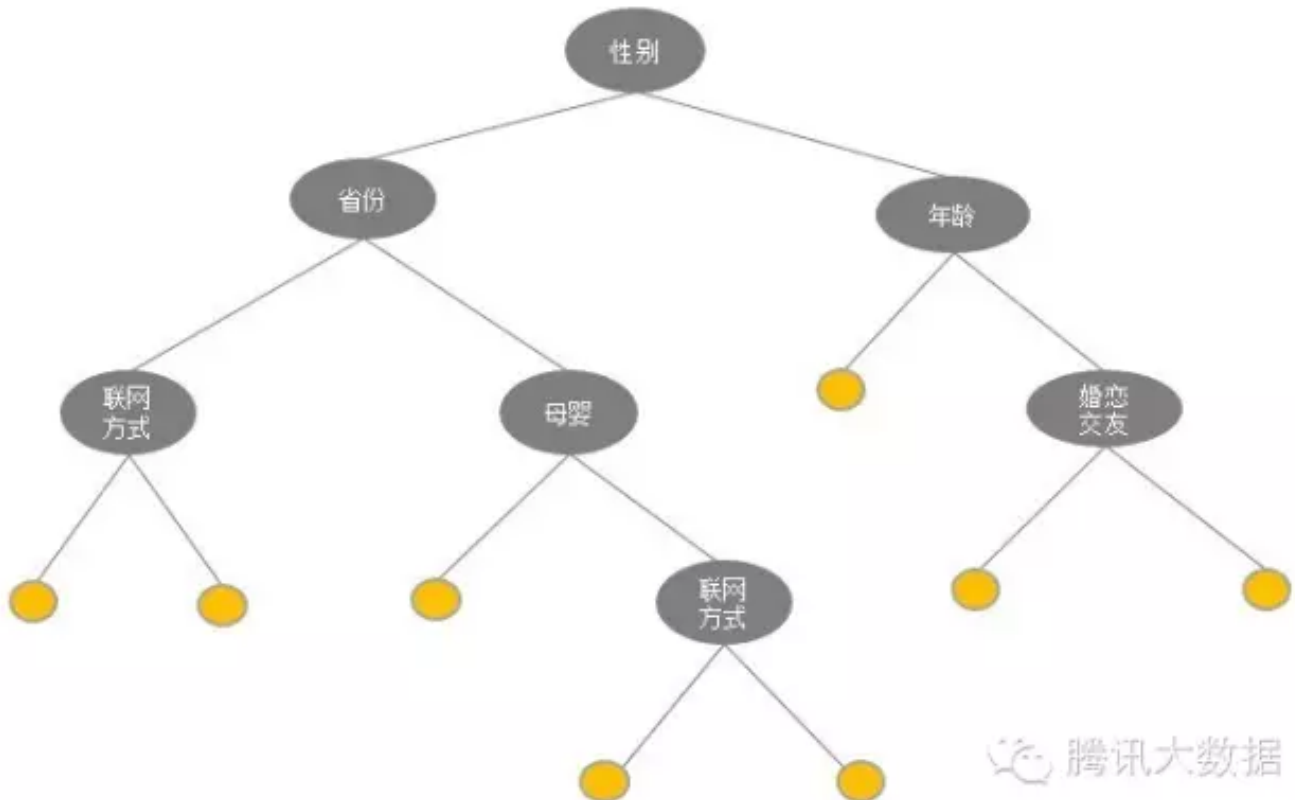


图4

### 如何使用GBDT 映射得到的特征？

通过GBDT生成的特征，可直接作为LR的特征使用，省去人工处理分析特征的环节，LR的输入特征完全依赖于通过GBDT得到的特征。此思路已尝试，通过实验发现GBDT+LR

在曝光充分的广告上确实有效果，但整体效果需要权衡优化各类树的使用。同时，也可考虑将GBDT生成特征与LR原有特征结合起来使用，待尝试。

## 4 总结与展望

点击率预估模型涉及的训练样本一般是上亿级别，样本量大，模型常采用速度较快的LR。但LR是线性模型，学习能力有限，此时特征工程尤其重要。现有的特征工程实验，主要集中在寻找到有区分度的特征、特征组合，折腾一圈未必会带来效果提升。GBDT算法的特点正好可以用来发掘有区分度的特征、特征组合，减少特征工程中人力成本，且业界现在已有实践，GBDT+LR、GBDT+FM等都是值得尝试的思路。不同场景，GBDT融合LR/FM的思路可能会略有不同，可以多种角度尝试。

参考文献：

- [1].Chapelle O, Manavoglu E, Rosales R. Simple and scalable response prediction for display advertising[J]. ACM
- [2].[http://blog.csdn.net/lilyth\\_lilyth/article/details/10032993](http://blog.csdn.net/lilyth_lilyth/article/details/10032993)
- [3].He X, Pan J, Jin O, et al. Practical lessons from predicting clicks on ads at facebook[C]. Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2014: 1-9.
- [4].<http://www.csie.ntu.edu.tw/~r01922136/Kaggle-2014-criteo.pdf>
- [5].<https://github.com/guestwalk/Kaggle-2014-criteo>
- [6].<http://www.cnblogs.com/leftnoteasy/archive/2011/03/07/random-forest-and-gbdt.html>
- [7].<https://github.com/dmlc/xgboost>
- [8].<http://cos.name/2015/03/xgboost/?replytocom=6610>
- [9].<http://vdisk.weibo.com/s/vlQWp3erG2yo/1431658679>
- [10].Ensemble Methods: Foundations and Algorithms (Chapman & Hall/Crc Machine Learning & Pattern Recognition): Zhi-Hua Zhou: 9781439830031
- [11].[http://blog.csdn.net/hero\\_fantao/article/details/42747281](http://blog.csdn.net/hero_fantao/article/details/42747281)
- [12]. Richardson M, Dominowska E, Ragno R. Predicting clicks: estimating the click-through rate for new ads[C]. Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 521-530.