KDD'24 | 腾讯广告:深入理解推荐模型Ranking Loss

原创 州懂学习笔记 州懂学习笔记 2024年09月23日 08:15 广东


州懂学习笔记
分享大模型推荐系统相关知识和学习笔记
37篇原创内容

公众号

KDD'24 | 腾讯广告:深入理解推荐模型Ranking Loss

> 标题: Understanding the Ranking Loss for Recommendation with Sparse User Feedback
> 地址: https://arxiv.org/pdf/2403.14144
> 公司: 腾讯广告
> 会议: KDD'24
> 代码: https://github.com/SkylerLinn/Understanding-the-Ranking-Loss

1. 问题背景

近年来, 业界有不少工作是在CTR模型中额外引入排序损失函数, 即:

$$L_{\text{final}} = \alpha L_{\text{classification}} + (1-\alpha) L_{\text{rank}}$$

比如早年Twitter的工作《Click-through Prediction for Advertising in Twitter Timeline》, 去年KDD'23阿里的JRC方法《Joint Optimization of Ranking and Calibration with Contextualized Hybrid Model》,这些都取得了不错的业务收益。

已有的工作主要将效果归功于Ranking Loss的引入提升了模型的排序能力, 但是, 并没有分析Ranking Loss的引入对CTR模型主要的分类能力的影响。这里, 作者的研究重点在于探索这种组合损失对模型分类能力的影响。

2. 梯度消失视角分析

2.1 组合损失代表: Combined-Pair Loss

这里作者DCN-V2作为Backbone模型, 并使用Combined-Pair Loss的组合损失作为不失一般性的研究对象, 其分类损失使用Binary Cross Entropy(简称BCE), 而排序损失使用RankNet loss, 具体地:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\sigma(z_i)) + (1-y_i)\log(1-\sigma(z_i))]$$

$$\mathcal{L}_{\text{RankNet}} = -\frac{1}{N_+ N_-}\sum_{i=1}^{N_+}\sum_{j=1}^{N_-}\log(\sigma(z_i^{(+)} - z_j^{(-)}))$$

同样的, 也是做加权融合

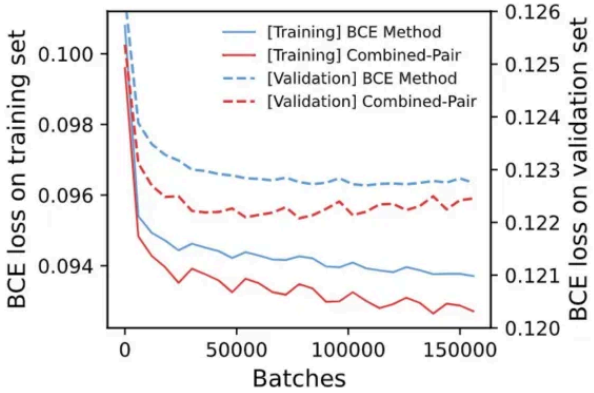
$$\mathcal{L}^{\text{CP}} = \alpha\mathcal{L}_{\text{BCE}} + (1-\alpha)\mathcal{L}_{\text{RankNet}}$$

2.2 分类能力探究

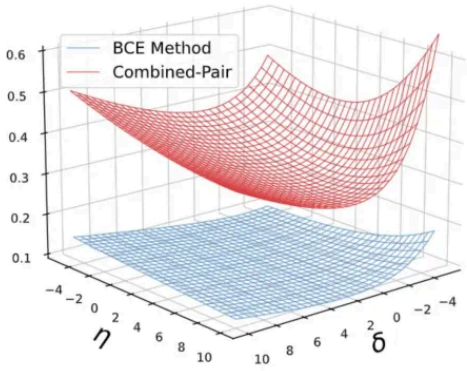
现有的研究一般认为, 在BCE损失之外引入Ranking Loss是可以提高排序能力, 但对于CTR模型的分类能力的影响并没有探究。

为了探索引入Ranking Loss对CTR模型分类能力的影响, 作者在公开的Criteo数据集上, 分别基于BCE Loss和Combined-Pair Loss进行训练。这里,需要注意的是, 由于Criteo数据集本身对负样本做了下采样, CTR很高, 为了模拟真实业务中CTR正样本稀疏的情况, 作者在训练和评估时都引入了小于1的正样本权重。

作者分别绘制了BCE方法和ComBined-Pair方法在验证集和测试集上的BCE损失, 如下图(a)所示:



(a) BCE Loss Dynamics along epochs.



(b) Loss Landscapes.

基于上述实验, 作者发现:

> **发现1:** Combined-Pair方法在验证集的BCE Loss比BCE方法更低, 说明了组合损失同时也提高了CTR模型的分类能力, 而不仅仅是排序能力。
>
> **发现2:** Combined-Pair方法在训练集的BCE Loss比BCE方法也低, 说明引入辅助排序损失有助于BCE损失的优化。

此外, 作者也分析了Combined-Pair方法和BCE方法的损失曲面的差异。如上图(b)所示, 可以看到, BCE方法的损失曲面比Combined-Pair更加平缓。

2.3 梯度分析

为了深入分析上述现象产生原因, 作者对BCE和Combined-Pair方法的梯度进行了详细分析。根据链式法则，DNN中每层参数的梯度与logit的梯度成正比, 因此, 作者分析了Logit梯度。

2.3.1 负样本BCE Loss的梯度

对于负样本的logit，其BCE Loss的梯度可以推导为:

$$\begin{aligned}\nabla_{z_j^{(-)}}\mathcal{L}_{\mathrm{BCE}} =& \frac{1}{1-\sigma(z_j^{(-)})}\cdot\sigma(z_j^{(-)})(1-\sigma(z_j^{(-)})) \\ =& \sigma(z_j^{(-)}) = \hat{p}_j.\end{aligned}$$

这表明负样本的梯度与它的pCTR值成正比, 而当正反馈稀疏时，pCTR值会很小。这样,负样本的梯度也会倾向于相对较小。作者将这种现象称为在**稀疏正反馈下的负样本梯度消失**。基于此, 作者得出如下结论:

> **发现3:** 当正反馈稀疏时，负样本会发生梯度消失，因为它们正比于估计的正率成正比，而这个正率在无偏估计器中是很小的。

2.3.2 正样本BCE Loss的梯度

对于正样本的Logit, 作者也推导了其BCE Loss的梯度:

$$\begin{aligned}\nabla_{z_i^{(+)}}\mathcal{L}_{\mathrm{BCE}} =& -\frac{1}{\sigma(z_i^{(+)})}\cdot\sigma(z_i^{(+)})(1-\sigma(z_i^{(+)})) \\ =& -(1-\sigma(z_i^{(+)})) = -(1-\hat{p}_i).\end{aligned}$$

可以发现, 正样本Logit的梯度$\nabla_{z_i^{(+)}}\mathcal{L}_{\mathrm{BCE}} \propto 1-\hat{p_i}$, 在稀疏正反馈下, 1 - pCTR是接近于1, 因此不会出现像负样本那样梯度消失的问题。

2.3.3 负样本Combined-Pair Loss的梯度

Combined-Pair Loss包含两部分, 先看RankNet损失中负样本logit的梯度，从上面RankNet Loss中, 可以很容易推导出:

$$\nabla_{z_j^{(-)}}\mathcal{L}_{\mathrm{Rank}}^{\mathrm{CP}} = \frac{1}{N_+}\sum_{i=1}^{N_+}\sigma(z_j^{(-)} - z_i^{(+)}).$$

在广告场景, 由于正反馈非常稀疏, 作者观察到即使是正样本的预估值也是远远低于0.5的, 也就是正样本对应的logit $z_i^{(+)}$是小于0的, 这会使得RankNet Loss中, 负样本的梯度会比BCE Loss的更大, 即有:

$$\begin{aligned}\nabla_{z_j^{(-)}}\mathcal{L}_{\text{Rank}}^{\text{CP}} =& \frac{1}{N_+}\sum_{i=1}^{N_+}\sigma(z_j^{(-)} - z_i^{(+)}) \\ >& \frac{1}{N_+}\cdot N_+ \cdot \sigma(z_j^{(-)}) \\ =& \sigma(z_j^{(-)}) = \nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}}.\end{aligned}$$

上述不等式说明了在稀疏正反馈的情况下, 对于相同的负样本logit, RankNet Loss可能会比BCE Loss有更大的梯度。这样, 对于Combined-Pair方法, 下面的不等式成立:

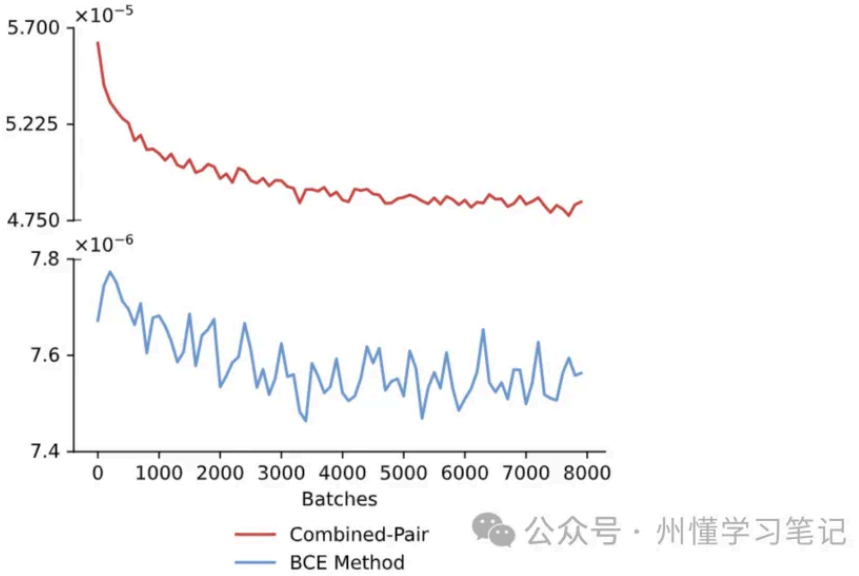
$$\begin{aligned}\nabla_{z_j^{(-)}}\mathcal{L}^{\text{CP}} =& \alpha\nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}} + (1-\alpha)\nabla_{z_j^{(-)}}\mathcal{L}_{\text{Rank}}^{\text{CP}} \\ >& \alpha\nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}} + (1-\alpha)\nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}} \\ =& \nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}}.\end{aligned}$$

基于此, 可以得到以下结论:

> **发现4:** 当正反馈稀疏时, 对于负样本, Combined-Pair方法有比BCE方法更大的梯度

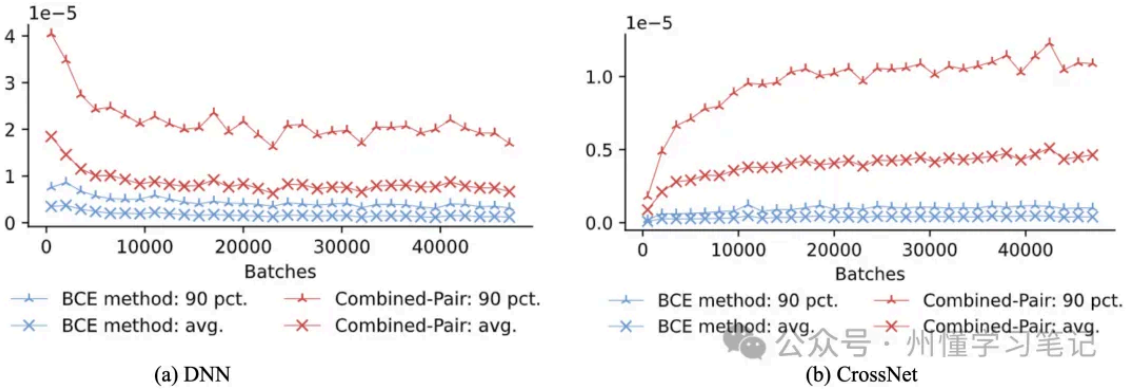
2.3.4 实证分析验证梯度消失问题

为了实证分析验证前面的梯度消失问题，作者在训练的第一个周期检查了负样本logits的梯度范数的动态变化过程。基于Criteo数据, 这里为了模拟稀疏正反馈情况, 作者调整了正样本权重, 并设置融合系数$\alpha = 0.5$, 实验结果如下图所示:



可以看出, BCE方法对负样本的梯度范数非常小, 而相应的Combined-Pair方法就要大很多。

此外, 作者还近一步的比较了DCN-V2里的DNN和CrossNet的底层梯度范数, 输出了训练过程中梯度范数的第90百分位数和平均值的动态过程。



(a) DNN (b) CrossNet

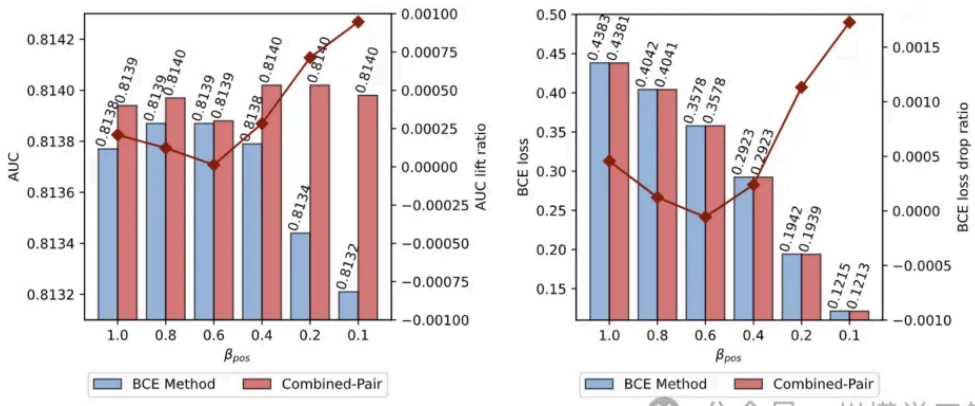
可以看出, 上述负样本梯度差异在训练过程中是一直存在的, 这也进一步验证了Combined-Pair方法能有效地缓解可学习参数的梯度消失问题。

3. 实验部分

同样还是基于Criteo数据集, 并对正样本权重一个小于1的权重β_{pos}, 实验主要围绕分析以下问题

3.1 RQ1: 不同正样本稀疏率下的效果评估

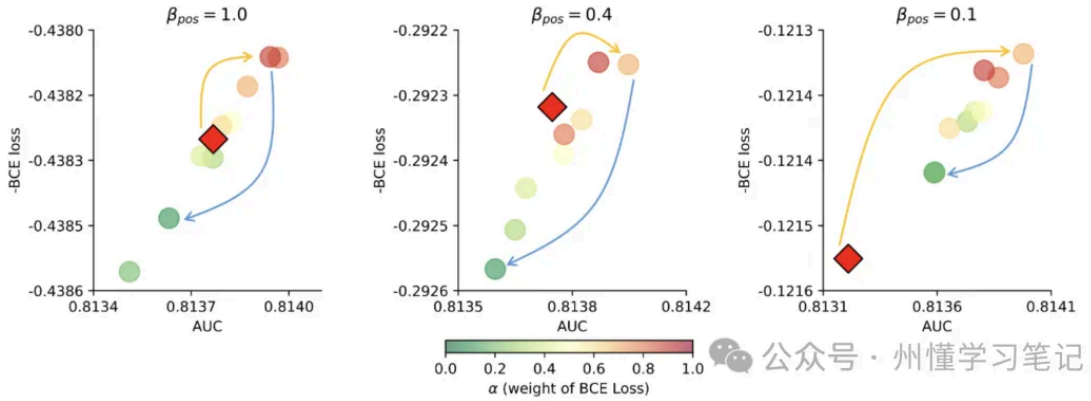
当正样本稀疏率小于某个阈值($\beta = 0.6$) 时, Combined-Pair方法比BCE方法在AUC和BCE Loss两个指标上, 都实现了更大的效果提升。



(a) AUC (b) BCE loss

3.2 RQ2：分类和排序损失间的权衡

组合损失中是通过α来平衡分类(α)和排序($1-\alpha$)损失的。下图作者给出了不同正样本稀疏率下, 通过调整α的值, 在AUC和BCE Loss两个指标上的效果(越右上表示效果越好)。特别留意下图中剪头的几个颜色的点, 分别是: 红色($\alpha = 1.0$完全由分类损失控制)、橙色($\alpha = 0.7$平衡了两个Loss), 绿色($\alpha = 0.1$由排序损失主导)。

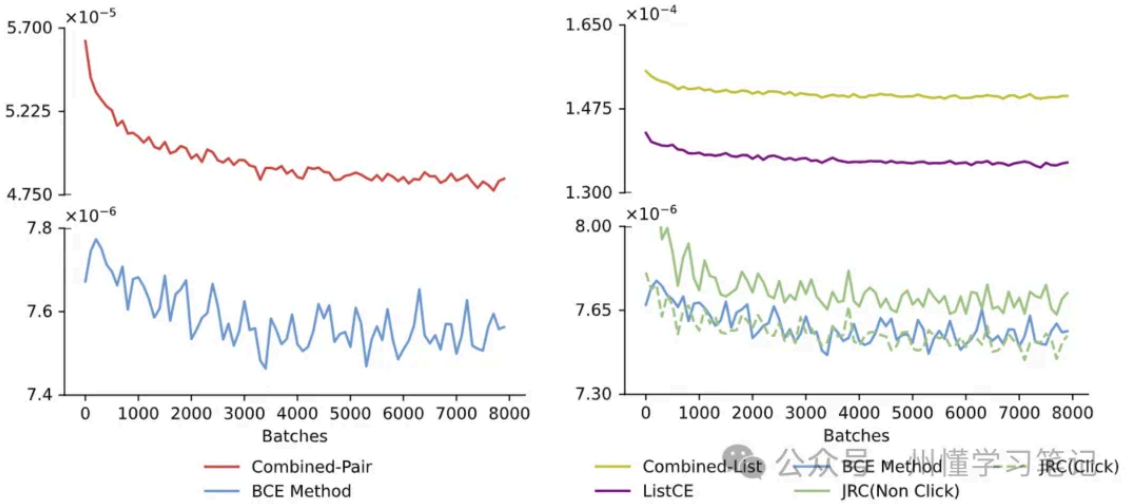


可以发现:

- 橙色的剪头表明: 通过减少α直到某个阈值，可以单调地提高分类和排序能力
- 蓝色的剪头表明:当排序损失在组合损失中占据主导地位超过某个阈值时，分类和排序能力都会单调地恶化。
- 当正反馈非常稀疏时, 即使排序损失占据主导，模型的效果仍优于BCE方法。

3.3 RQ3: 其他排序损失的评估

作者也分析了Combined-Pair之外的其他分类-排序组合方法(如Combined-List, RCR, JRC等)的梯度，如下图右侧所示:



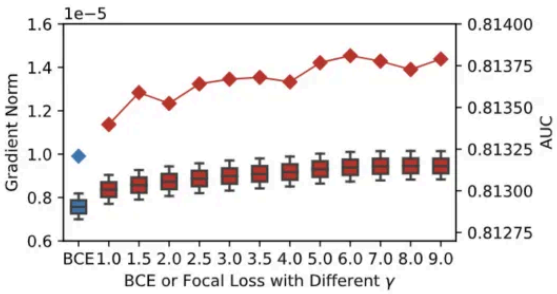
可以发现, 和BCE方法相比，这些组合方法在负样本的梯度范数都有不同程度的提高, 表明梯度消失问题在这些方法中也得到了缓解。

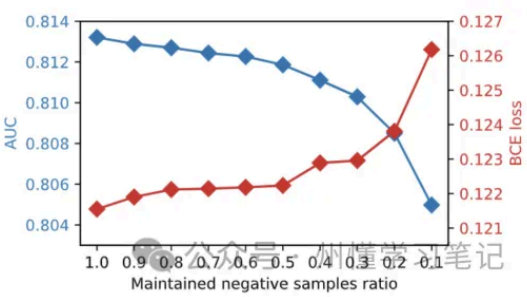
此外, 作者还给出了$\beta_{\text{pos}} = 0.1$下, 不同损失组合方法的效果, 说明了其它损失组合方法通过引入Ranking Loss缓解了梯度消失问题, 进而带来性能提升。

Metric	BCE	BCE+Pairwise	BCE+Listwise		
		Combined-Pair	JRC	Combined-List	RCR
AUC↑	0.81321	**0.81398**↑	0.81355↑	0.81351↑	0.81349↑
BCE loss↓	0.12152	**0.12131**↓	0.12146↓	0.12152	0.12141↓

3.4 RQ4: 可以扩展到分类排序组合损失之外的方法

3.4.1 Focal Loss和负采样方法





3.4.2 与对比学习Loss融合的方法

Stage	Metrics	BCE Method	Combined-Contrastive
Training	Gradient Norm	4.9×10^{-6}	7.5×10^{-6}
	BCE loss ↓	0.09667	**0.09428**↓
Testing	AUC↑	0.81321	**0.81340**↑
	BCE loss↓	0.12152	**0.12147**↓

3.5 线上实验

3.5.1 整体效果

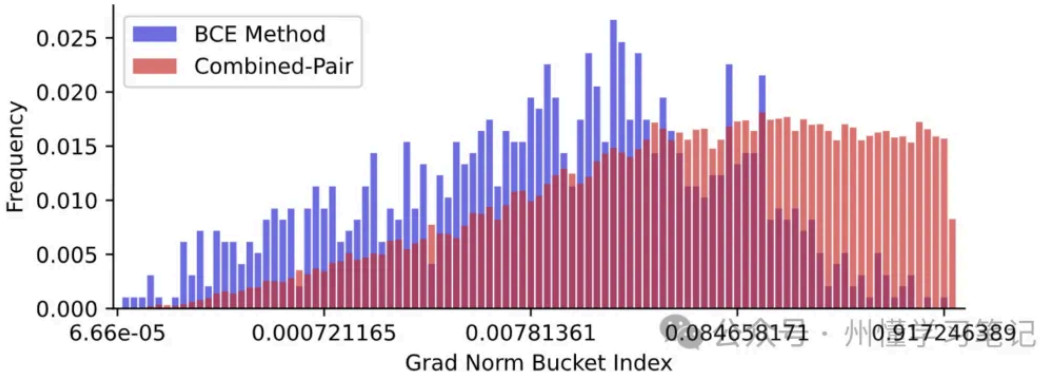
在腾讯广告的多个广告CTR预估场景下的整体效果

Table 4: Online A/B Testing Results.

Ad Scenario	CTR	GMV	Cost
WeChat Channels	+0.91%	+1.08%	+0.29%
WeChat Moment	+0.16%	+0.70%	+0.59%
DSP	-0.04%	+0.55%	+0.15%

3.5.2 负样本梯度分布

与BCE方法相比，Combined-Pair方法的负样本梯度分布显著右偏，表明Combined-Pair方法在负样本上获得了更大的梯度。



3.5.3 新广告冷启

越新的新广告(比如在当天创建的新广告),GMV效果提升越大, 表明对冷启友好

Table 5: Online A/B Testing Results for New Ads.

Launch Date	GMV	Cost
T	+1.04%	+0.27%
T-1	+1.04%	+0.27%
T-2	+0.83%	+0.47%
T-3	+0.81%	+0.17%
Total	+1.26%	+0.34%