MBCT——基于树模型的特征可感知的分箱校准

引言

预估校准（or不确定度校准，Uncertanty Calibration）在广告系统中的重要性越来越被大家所认可，在出价稳定性、混排公平性以及扣费合理性等方面发挥着重要作用，是平台收益和客户体验的重要保障能力。正因如此，我们对校准问题也有着长期的探索和思考。近期，我们借助树结构实现了一种智能化分箱校准算法框架（MBCT），并提出一种全新的校准评价方法（MVCE），取得了预估校准问题的新突破。此外，我们还发布了一个大规模预估校准数据集（CACTRDC），是目前已知最大的CTR校准数据集，为今后的校准研究提供了更丰富的数据基础。相关内容已发表在TheWebConf 2022上，欢迎感兴趣的同学阅读交流。

附相关资料链接：

论文：MBCT: Tree-Based Feature-Aware Binning for Individual Uncertainty Calibration

数据集：https://github.com/huangsg1/Tree-Based-Feature-Aware-Binning-for-Individual-Uncertainty-Calibration

WWW 22 Presentation：https://www.youtube.com/watch?v=Tz5_y8370a0

MBCT工作的解读已首发于阿里妈妈技术公众号，以下内容转载自阿里妈妈技术公众号：
https://mp.weixin.qq.com/s/lE6gt2m780dJvwbVEhCBLg

背景

以点击率预估为代表的许多机器学习应用，不同于图像或文本分类等拥有确定性答案的分类任务，其标签值（Label）通常是某个概率分布下的一次观测结果。在这类预估任务上，我们不仅希望能够预测事件发生的倾向性（如消费者更可能点击哪个商品；患者更可能是患有什么疾病），往往还希望能产出事件发生的真实概率。以广告系统中的点击率为例，预估点击率参与RTB广告系统中的排序和计费逻辑，其大小准确性会显著影响广告的分配效率和计费准确性。

在这类基于不确定事件观测样本的预估任务中，人们通常对问题进行抽象和简化，假设样本特征和事件的联合分布服从某种函数形式，并以Data-Driven的方式学习该分布。但是绝大多数方法（如逻辑回归，神经网络等）只能学习到较好的序关系，其预估值往往会偏离其真实概率。此外，在实际应用中我们只能得到事件的观测结果，而不能得到事件发生的真实概率，因此也难以度量预估值和真实概率的偏差。

预估不确定度校准正是为了解决这些问题，其主要研究方向包括校准误差度量（如何度量预估结果和真实概率值的偏差）和校准算法（如何减小模型预估校准的误差）。本文将介绍我们在这两个方向上的新

工作。

校准误差的度量

为了衡量预估值和真实概率的误差，我们可以直观地定义一个 p 范数的误差函数——TCE（True Calibration Error）：

$$\mathrm{TCE}_p(f) = \left(\mathbb{E}_X\left[|\mathbb{E}[Y|X] - f(X)|^p\right]\right)^{\frac{1}{p}} \qquad (1)$$

其中 X 表示特征空间， Y 表示观测标签的空间， f 表示原始预估模型。

然而在点击率预估等任务中，其真实概率是无法得到的，我们无法直接计算TCE。因此人们通常使用 p 范数的ECE[6,10,11]（Expected Calibration Error）来近似地表征TCE：

$$\mathrm{ECE}_n = \Big(\frac{1}{n}\sum_{i=1}^{n}\mathrm{PCE}(D_i)^p\Big)^{\frac{1}{p}} \qquad (2)$$

其中 n 为分桶数量， D 为数据集， D_i 为数据集第 i 个分桶对应的子集。在计算ECE时，首先要将测试集 D 中的样本按照原始预估值大小排序，然后等频或等距地进行分桶，将其切分为 n 个子集，最后在每个子集 D_i 上计算PCE，并按照公式2得到ECE。其中PCE为Partition Calibration Error，其定义如公式3所示：

$$\mathrm{PCE}(D_b) = |\hat{y}_b^{\mathrm{pred}} - \hat{y}_b| \qquad (3)$$

在过去的研究工作中，ECE一直作为不确定度校准的主要评价指标。Google的Roelofs等人[1]针对ECE的不足，提出了其变体 ECE_{sweep} ，在将样本按照预估值大小排序后，应用一种新的分桶策略（等频分桶下使得每个桶的正样本数保序的最大分桶数量），如公式（4）所示，其中 $\hat{y}_i$ 表示第 i 个分桶中正样本的数量：

$$ECE_{sweep} = max_n((\frac{1}{n'}\sum_{i=1}^{n'}PCE(D_i)^p)^{\frac{1}{p}},\ s.t.\ \hat{y}_1 \le \hat{y}_2\ ... \le \hat{y}_{n'},\ \forall n' \le n) \qquad (4)$$

但是这些指标都只从某些特定的维度去衡量校准误差，对于一个绝对好的校准结果，其任意维度下（样本足够置信的子集合）的PCE都应该接近于0。因此我们提出了一个多维度的校准误差评估指标来缓解现有评价指标的问题，记作MVCE（Multi-View Calibration Error），其计算方式可形式化为：

$$\mathrm{MVCE}_{f,h,\{\mathrm{div}_i\}_{i=1}^{r}}(D) = \left(\frac{1}{r}\sum_{i=1}^{r}\left(\frac{1}{t_i}\sum_{j=1}^{t_i}\mathrm{PCE}(D_{i,j})\right)^p\right)^{\frac{1}{p}} \qquad (5)$$

其中 f 和 h 分别为预估函数和校准函数， div_i 表示对数据集 D 的第 i 种划分， $D_{i,j}$ 表示由第 i 种划分方法得到的第 j 个划分子集。在计算MVCE时，我们采取 r 种不同的划分（分桶）方式，计

算每种划分下的平均ECE作为MVCE的结果。为了构建不同的划分方式，我们首先对数据集进行随机打乱，然后进行等频划分计算ECE，这样迭代多次后MVCE将逐步收敛。

校准评价指标的误差对比

对于任意评价指标 μ ，其与真实校准误差TCE的偏差可以定义为：

$$E_{bias} = |\mathbb{E}[\mu(X,h)] - \mathrm{TCE}(X,h)| \qquad (6)$$

其中 X 为样本特征， h 为预估函数。但是在真实概率分布未知的情况下，公式（6）是无法计算的。因此我们进行了模拟实验，直接假设某个概率分布，并采样生成观测样本，这样就能计算其真实TCE，并可以进一步比较不同校准评价指标与TCE的偏差。采用Monte Carlo方法来估计评价指标 μ 与TCE的期望偏差：

$$\hat{E}_{bias}(n) = \frac{1}{m}\sum_{i=1}^{m} |\mu(X_i^{(n)},h) - \mathrm{TCE}\,(X,h)| \qquad (7)$$

其中 m 为仿真实验次数， n 为每次仿真实验下的采样数。

实验结果如图1和图2所示，其中图1展示了固定分桶数量下各指标误差随样本数量的变化。可以看到在样本量足够时，MVCE显著优于 ECE 和 ECE_{sweep} ，即使在样本量很匮乏时，MVCE也明显优于ECE。

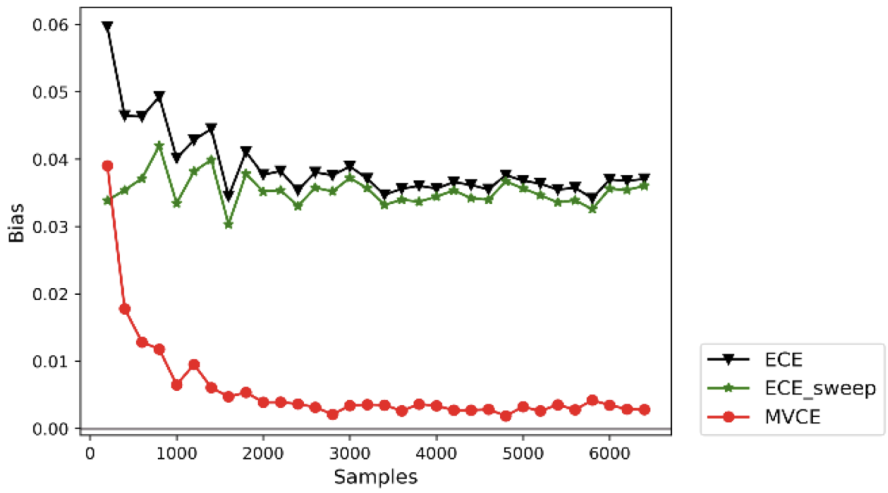

图1. 各指标随样本数量的变化（分桶数固定）

由于ECE和MVCE都有分桶数量的超参，图2则展示了在不同分桶数和样本数下的误差结果。可以看到MVCE在各类超参下都显著优于ECE。因此**模拟实验充分表明MVCE是一个更先进的校准误差评价指标。**

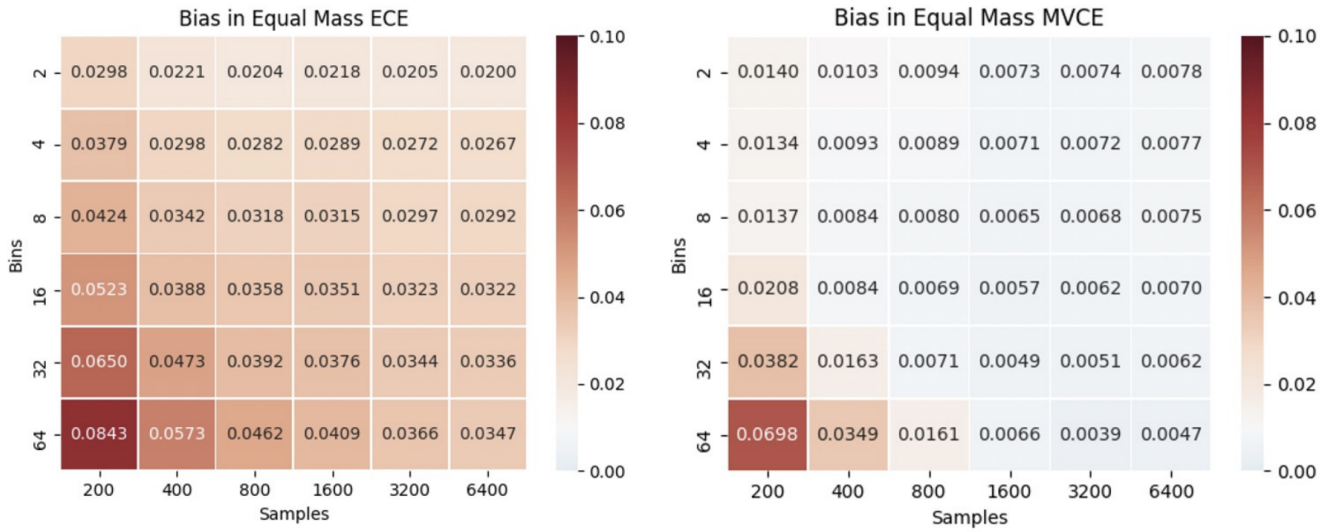

图2. 不同分桶和样本数下ECE和MVCE的校准误差

校准方法

除了校准误差的度量外，我们对现有的校准方法也进行了分析和总结，并提出了不确定性校准的一种新思路。

现有校准方法通常可分为三类：参数化方法（Parametric Methods），非参数化方法（Non-parametric Methods），以及它们的混合方法（Hybrid Methods）。其中参数化方法通常假设学习器的输出具有某种特定的分布，从而对应提出一个参数化的函数来对预估值进行修正，代表工作有Platt scaling[4]，Dirichlet scaling[5]，Temperature scaling[6,7]，Beta Calibration[12]等。这类方法通常对数据量需求较少，但是却缺乏理论保证[2,3]。非参数化方法通常应用于数据充足的校准场景，典型的代表是分桶（Binning）校准。以Histogram Binning为例[9]，首先将样本按照预估值排序，然后将数据等频分到多个桶中，每个桶的标签统计均值作为对应样本的校准结果。与参数化方法相比，非参数方法提供了一些分布无关的误差理论保证，但是却对需训练样本量有着更高的需求。因此，Kumar等人[3]提出了Scaling-Binning方法，将参数化与非参数化方法相结合（即混合方法），以平衡校准方法的理论误差保证和数据应用效率（更多相关文献详见：https://github.com/huangsg1/uncertainty-calibration）。

我们的工作也是一种混合方法，包括一种可学习分桶的思想——**特征可感知的分桶方法**（Feature-aware Binning），其大幅度提升了朴素Binning方法的分桶效率，也使得参数化与非参数化方法的结合更加自然；以及通过简单的参数化设计实现了样本粒度**个性化和非保序校准**。由于主要采用**树模型**进行校准方法设计，因此我们将该工作命名为**Multiple Boosting Calibration Trees（MBCT）**。

特征可感知的分箱方法

传统Binning方法通常将样本按照预估值排序，然后等频（uniform–mass）或等距（equal–width）地进行分桶，并用桶内样本标签均值作为其各样本的校准结果。假设被校准的数据集为 D ， D_b 为任意分桶子集。可以证明，Binning方法有如下理论误差保证[2]：

$$|\mathbb{E}[Y_b] - \hat{y}_b| \leq \sqrt{\frac{2\hat{V}_b \ln(3B/\alpha)}{c_b}} + \frac{3\ln(3B/\alpha)}{c_b} \qquad (8)$$

其中 $\hat{y}_b$ 为 D_b 上的校准结果（也是标签均值）， $\hat{V}_b$ 为样本标签的方差， c_b 为 D_b 的样本数量， B 为分桶数量，不等式会以不小于 $1-\alpha$ 的概率成立，即Binning能做到分桶上的渐进无偏校准。但是对于分桶内更小的样本子集是没有理论保证的，即我们将 D_b 再划分为两个子集，他们很可能仍是被高估或低估的。

Kumar等人提出的Scaling–Binning方法直接将Platt Scaling嵌入了Histogram Binning中，即每个分桶中采用Platt Scaling校准，平衡了Scaling和Binning方法的数据效率和校准效果；同时他们也证明了存在更好的分桶方案，但并没有提供找到它们的方法[3]。

那在Binning思想下有没有可能做到更完美的校准呢？我们进一步对Binning方法进行分析：任何基于Binning的方法都可以看作是先分桶再在桶内应用某种Scaling的方法。其中Histogram Binning采用的是将预估值设定为 $\hat{y}_b$ 的特殊Scaling；Scaling–Binning则使用的是Platt scaling。对于朴素的分桶方法，如果我们恰好能把真实概率相同的样本放到了一个桶里，那校准结果就是完美的（即TCE=0）。但是“真实概率相同”这个条件在大部分应用场景中均难以做到。对于Scaling方法的选择，如果桶内数据分布是符合Scaling分布假设的，那就有可能做到更好的校准。于是，**对于Binning思想下的校准方法，可以考虑采用更好的分桶方法让分桶后的数据分布恰好满足桶内Scaling函数的分布假设，从而得到更完美的校准结果。**

原始预估模型的偏差与其训练样本的特征空间往往是存在特定的对应模式的（如在某些子空间偏向于高估，而在另一些空间偏向于低估），而且这些偏差模式可以通过学习的方式被进一步捕捉。因此我们提出了特征可感知的分桶（Feature–aware Binning）框架，即通过机器学习模型捕捉原始被校准模型在其特征空间上的预估偏差模式，并将具有相同模式的样本分到一个分桶中，从而提升其校准效果。高低估是最直观的偏差模式，因此不失一般性，本文直接采用 $g_b(f(\cdot)) = k_b f(\cdot)$ 作为Scaling函数（其中 f 为被校准模型）。而且这种Scaling方式也会使我们的方法具备个性化和非保序校准的特点，其意义将在下一节详细阐述。

那么在Feature–aware Binning框架下，如何指导模型优化呢？直觉上，如果一个方法在数据集 D 上做到了“完美”校准，那么对于任意具有统计意义的子集，其PCE均应接近0。这与上面提出的MVCE思想具备一致性，因此我们采用MVCE作为损失函数来指导模型迭代。当然，理论上任意的校准误差函数都可以应用在我们的方法中，我们也在实验中比较了应用不同校准误差函数的效果。

校准与序关系的讨论

现有校准方法中绝大多数都直接或间接的保证了校准后的结果与原始打分序关系的一致性，但是根据校准方法是否作用在单样本粒度上，其序关系一致性保证的强弱不同，如保序回归是一种强序关系保证的算法，但Histogram Binning因为每个分桶内所有样本校准结果都一样，可以看做是一种弱保序的算法。在计算广告中，**单样本粒度序关系的优化及其重要**，所以一方面我们也需要将单样本粒度校准的设计思路考虑进来，**本文中我们将这种作用在单样本粒度的校准能力称为个性化校准**；另一方面也思考如何打破校准结果保序的限制，站在原始预估模型“巨人的肩膀”上进一步提升预估值的排序能力。这两方面的设计在下文中会详细阐述。

Multiple Boosting Calibration Tree

基于对特征可感知的分箱方法和样本粒度排序能力优化的思考，我们提出了如图3所示的校准算法：Multiple Boosting Calibration Tree（MBCT）。具体来讲，**考虑到树模型具备可解释、节点可自学习分裂方式、便于集成、易于上线等特点，我们采用集成树结构模型来实现特征可感知的智能分箱能力。**MBCT以原始预估值及相关特征作为输入，以MVCE损失指导节点分裂，每棵树从根节点到任意叶子结点的路径都唯一确定一种分桶模式。

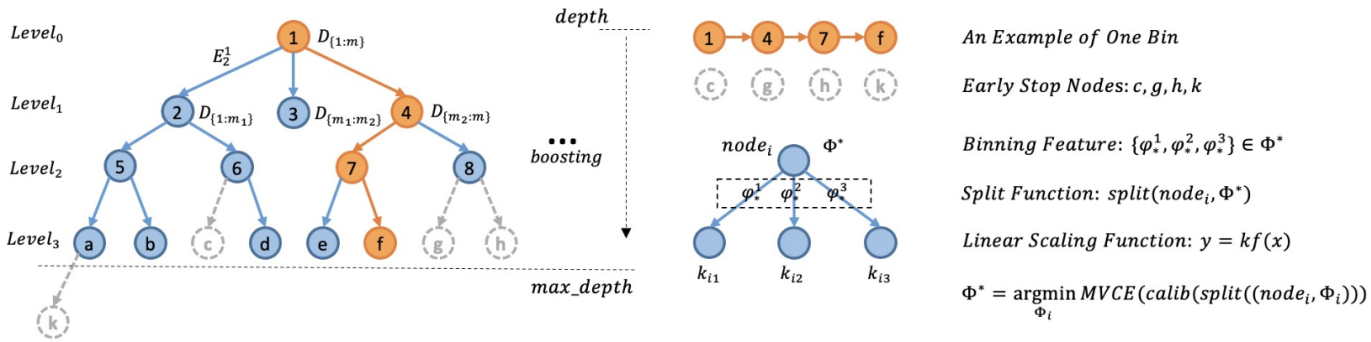

图3. MBCT模型概览

MBCT方法的超参数包括树深度，集成树棵数，百分比误差限制 e ，错误率 α （置信度为 $1-\alpha$ ）等。每个节点分裂时选择使该节点MVCE最优的特征进行分裂，形式化表述为：

$$\Phi_* = \text{argmin}_{\Phi_i \in \mathcal{F}} \text{MVCE}(\text{calib}(\text{split}(D', \Phi_i))) \qquad (9)$$

其中 D' 为节点对应的样本集合，split(·)表示节点分裂函数，calib(·)表示分裂后对每个子节点的校准函数。节点分裂时有三个停止条件：1）树深度已达到预设值；2）节点内样本数量少于 β （ β 为使得公式10成立的c的最小值）；3）节点对应的局部损失函数在分裂后变大。

$$\hat{y}_D \leq \frac{1}{e}(\sqrt{\frac{2\hat{V}_D \ln(3B|D|/c\alpha)}{c}} + \frac{3\ln(3B|D|/c\alpha)}{c}) \qquad (10)$$

此外，我们使用Boosting集成方法来提升模型学习能力，这能使得基学习器校准误差大的样本在后续受到更多关注：单棵校准树中随着层数增大，每个节点数据量将越来越少，直到无法继续分裂，这个过程中会使得我们为了得到精细化的分桶而损失了置信度的保证，而Boosting则能起到一定的优化作用。

MBCT算法框架能够很好的实现特征可感知的智能分箱功能，而且结合这种分箱能力，**在单样本粒度的序关系问题上，则可以直接在公式（9）中的calib(·)中应用朴素的 $y = k * f(x)$ （注：不局限于此种函数）来实现。**直观来讲，这种Scaling方式针对任意样本 x 均能直接计算得到对应的校准结果 y （单样本粒度）；而且每层节点、每条路径及每棵树之间组成复杂的分桶模式，使得原始预估模型误差模式的识别更加精细，各种误差模式之间是不保序的，从而使得在MBCT学习能力足够时，校准后的排序效果是有进一步提升的。

实验与分析

实验设置

我们进行了离线和在线实验，以验证我们方法的有效性以及实际应用价值。 离线实验中，我们使用了从阿里妈妈展示广告日志系统中抽取并脱敏出的一个数据集：CACTRDC（Computation Advertising Click-Through Prediction Dataset for Calibration ），以及两个公开数据集 Porto Seguro和 Avazu。具体如表 1所示，其中Predictor Train为用来训练被校准模型的数据，Calibration Train为用来训练校准模型的数据，Calibrattion Test为校准的测试集。

表1. 离线数据集信息

Dataset	Predictor Train	Calibration Train	Calibration Test
CACTRDC	48M	11M	1M
Porto Seguro	357K	208K	30K
Avazu	24M	12M	4M

对比算法包括Platt Scaling[4], Beta Calibration[12], Histogram Binning[9], Isotonic Regression[8], Scaling-Binning[3]。评价指标采用MVCE和AUC分别评价校准误差和排序水平。更多实验设置详见论文。

离线实验

离线实验结果如表2所示，从结果上看，MBCT在各数据集上均显著优于对照算法（注：MVCE越小越好），并且在boosting的作用下其效果有进一步提升。

表2. 离线实验主要结果

Method	CACTRDC		Porto Seguro		Avazu	
	MVCE ↓	AUC ↑	MVCE ↓	AUC ↑	MVCE ↓	AUC ↑
Original Predictions	0.00394	0.77902	0.00619	0.62869	0.00976	0.71880
Platt Scaling	0.00374	0.77902	0.00604	0.62869	0.00792	0.71880
Beta Calibration	0.00371	0.77902	0.00601	0.62869	0.00789	0.71880
Histogram Bining	0.00372	0.77895	0.00597	0.62998	0.00787	0.72381
Isotonic Regression	0.00371	0.77915	0.00598	0.62936	0.00787	0.72030
Scaling-Binning	0.00373	0.77892	0.00605	0.62880	0.00792	0.71871
Our full MBCT	**0.00368**	**0.78693**	**0.00586**	**0.63097**	**0.00780**	**0.74177**
Our model w/o boosting	0.00374	0.78373	0.00595	0.62999	0.00784	0.73797

我们还做了一些更细致的分析，限于篇幅只介绍CACTRDC上的结果。1）在校准误差评价时，分桶数量对置信度有直接影响，为衡量模型结果的置信度，按不同分桶数对比各算法校准效果，如图4所示，MBCT稳定优于各类基准算法。

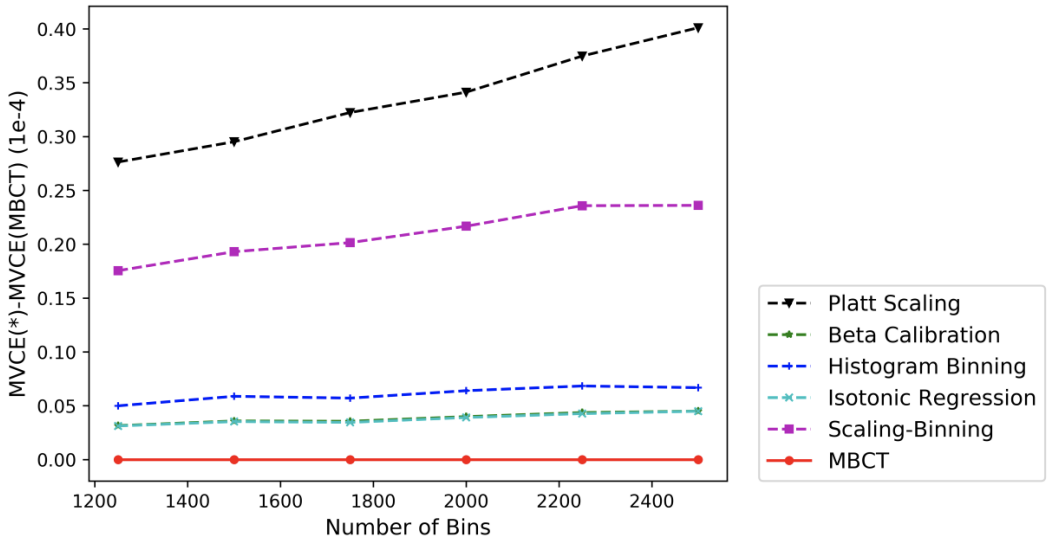

图4. 不同分桶数量下基准方法与MBCT的MVCE差值

2）使用多种度量指标作为损失函数来训练校准模型，并对比校准误差，如图5所示，采用MVCE作为损失函数取得了最好的效果，这进一步验证MVCE指标的优越性。

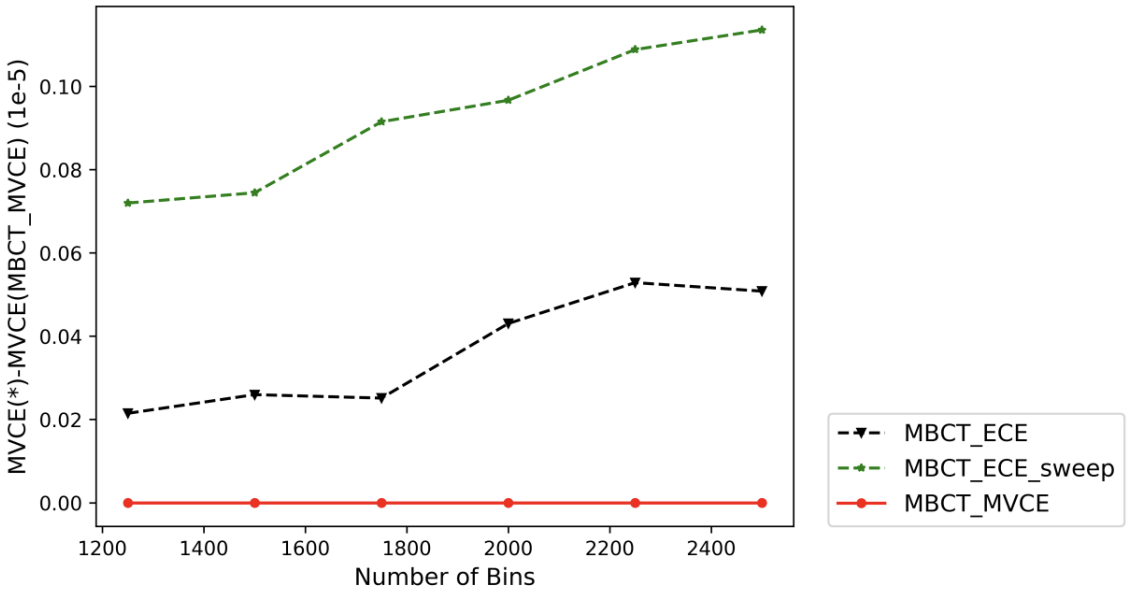

图5. 训练Loss比较结果

在线实验

在阿里妈妈线上广告系统中进行了15天的AB实验，主流量校准基线为Isotonic Regression。实验结果如表3所示，可以看到MBCT带来了相当可观的点击率和广告收入提升，表明了其在广告业务中的重要价值。

表3. 在线实验结果

Metric	No Calibration	Isotonic Regression	MBCT
AUC	0.0%	+0.3%	+5.7%
CTR	0.0%	+8.8%	+22.1%
eCPM	0.0%	+4.5%	+8.4%

结论与展望

本文介绍了我们在不确定度校准上的新工作，一种智能化分箱校准算法框架（MBCT）。从学术角度看，我们从更广义的视角将传统Binning与Scaling方法纳入同一体系，并通过贪心方法在模型特征空间上学习更好的误差模式，突破了传统启发式方案的局限性，也使得Binning和Scaling的结合相比已有工作[3]更加自然和高效。此外，我们打破了绝大部分校准算法对保序的固有认知，取得了校准误差和排序能力的双重突破。

但是MBCT仍有较多优化空间：1）MBCT以近似贪心的方法进行学习，如何实现更高效的全局优化方法？2）Scaling分布的假设是直观且朴素的，如何给出更准确的分布先验并结合智能分箱框架取得更好

的效果？此外，在计算广告实际应用中，召回阶段我们真正关心的是Bid*CTR（ECPM）排序能力的优化，此时校准算法的优化空间及预估误差容忍度是和精排阶段不同的，如何针对这一业务属性设计校准方法？这些问题我们将持续探索，而且我们论文挂出后也陆续收到业界朋友的关注和反馈，我们非常欢迎大家持续交流，共同推动校准技术的发展。

参考文献

[1] Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. 2020. Mitigating bias in calibration error estimation. arXiv preprint arXiv:2012.08668 (2020).

[2] Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. 2020. Distributionfree binary classification: prediction sets, confidence intervals and calibration. In Advances in Neural Information Processing Systems.

[3] Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In Advances in Neural Information Processing Systems. 3787—3798.

[4] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers 10, 3 (1999), 61—74.

[5] Meelis Kull, Miquel Perello–Nieto, Markus Kängsepp, Hao Song, Peter Flach, et al. 2019. Beyond temperature scaling: obtaining well–calibrated multiclass probabilities with dirichlet calibration. arXiv preprint arXiv:1910.12656 (2019).

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In International Conference on Machine Learning. 1321—1330.

[7] Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. 2018. Attended temperature scaling: a practical approach for calibrating deep neural networks. arXiv preprint arXiv:1810.11586 (2018).

[8] Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 694—699.

[9] Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In International Conference on Machine Learning. 609—616

[10] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In AAAI Conference on Artificial Intelligence, Vol. 29.

[11] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. 2014. Binary classifier calibration: non–parametric approach. arXiv preprint arXiv:1401.3390 (2014).

[12] Meelis Kull, Telmo Silva Filho, and Peter Flach. 2017. Beta calibration: a wellfounded and easily implemented improvement on logistic calibration for binary classifiers. In Artificial Intelligence and Statistics. 623—631.