

鱼遇雨欲语与余

Contents

目录

- | | |
|--------|-----------|
| 1 团队介绍 | 4 模型介绍 |
| 2 赛题理解 | 5 规则与模型融合 |
| 3 特征工程 | 6 总结与思考 |



团队介绍



刘育源

现就读于哈尔滨工业大学

集成电路工程

硕士二年级



郭达雅

中山大学与微软亚洲研究院

联合培养

计算机科学与技术

博士一年级



王贺

现就职于京东

算法工程师



赛题理解

数据

- 历史日志数据：广告请求时间、用户id、广告位id、竞价广告信息等
- 用户信息数据：包含用户id、年龄、性别、地域、行为兴趣等
- 广告数据：广告操作信息、广告静态信息

目标

预测广告的日曝光量

评估指标

$$\text{SMAPE} = \frac{1}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{(F_t + A_t)/2}$$

准确性指标

$$\text{score} = \frac{1}{n} \sum_{k=1}^n \frac{(imp_0 - imp_k)(bid_0 - bid_k)}{|(imp_0 - imp_k)(bid_0 - bid_k)|}$$

出价单调相关性指标

训练目标

预测

初始训练目标

$$\frac{|f(x) - imp|}{(f(x) + imp)^2}$$

$$y = f(x)$$

梯度平滑，且与原训练目标相似

$$|f(x) - \log(1 + imp)|$$

$$y_{base} = e^{f(x)} - 1$$

单调性考虑，保证训练出来结果符合单调性。

$$\left| f(x) - \frac{\log(1 + imp)}{\log(1 + bid)} \right|$$

$$y_{bid} = e^{f(x) \times \log(1 + bid)} - 1$$

基本曝光与单调性的结合

$$y = w_1 y_{base} + (1 - w_2) y_{bid}$$

数据集划分

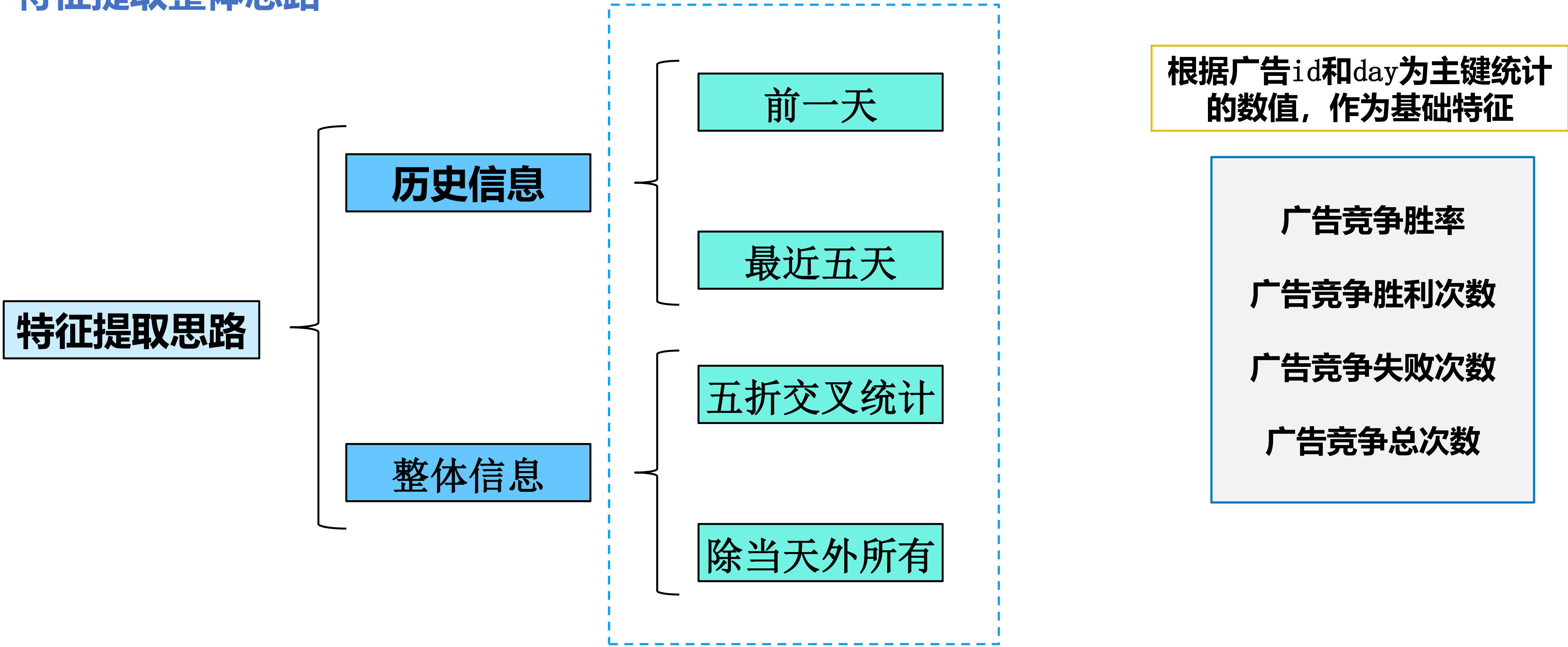


复赛A榜训练集和测试集是连续的，复赛B榜则是不连续的，面对“**跨天**”预测，难度是非常大的。所以我们利用“**远程监督**”的方式，利用现有的标注数据，训练一个模型，给未标注数据进行标注，然后拿过来再训练。



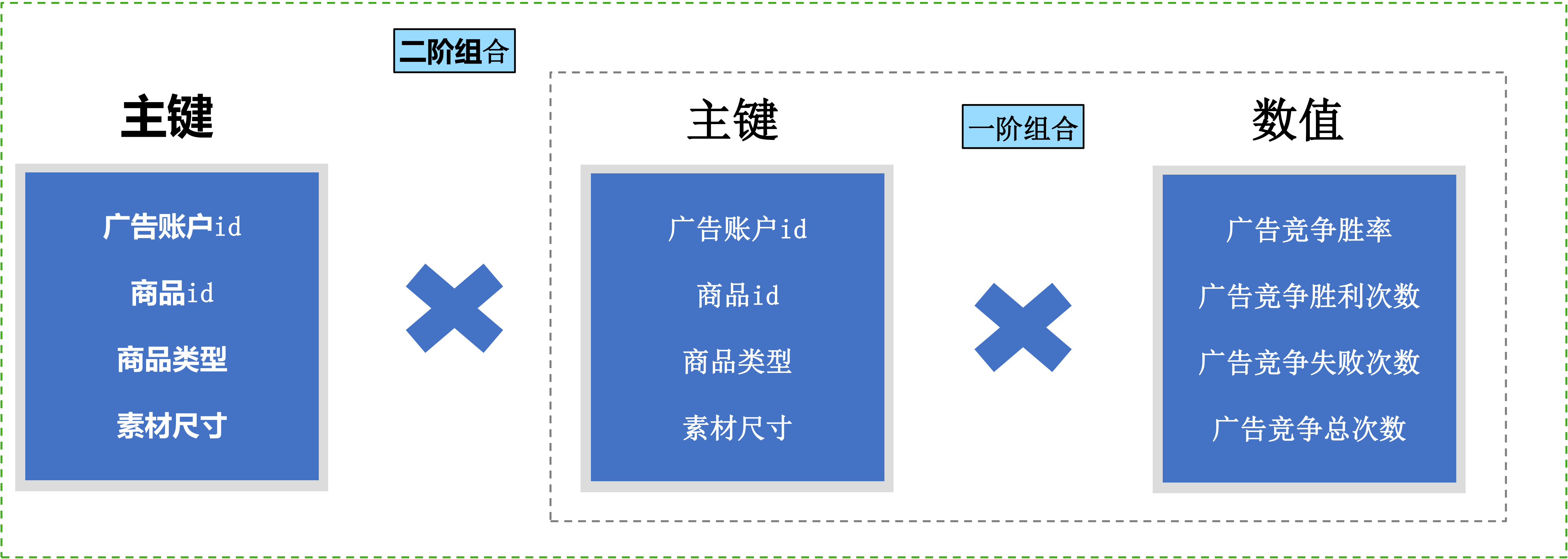
特征工程

特征提取整体思路



多维度下提取特征，覆盖更多可能性

如何构造新广告的特征



通过组合构造主键下数值的均值、中位数可以很好的覆盖新广告，同时进行一阶、二阶构造，更细粒度挖掘。

更进一步扩展

假数值

广告竞争胜率

广告竞争胜利次数

广告竞争失败次数



真数值

广告竞争总次数



接近真实

广告竞争胜率 * 广告竞争总次数

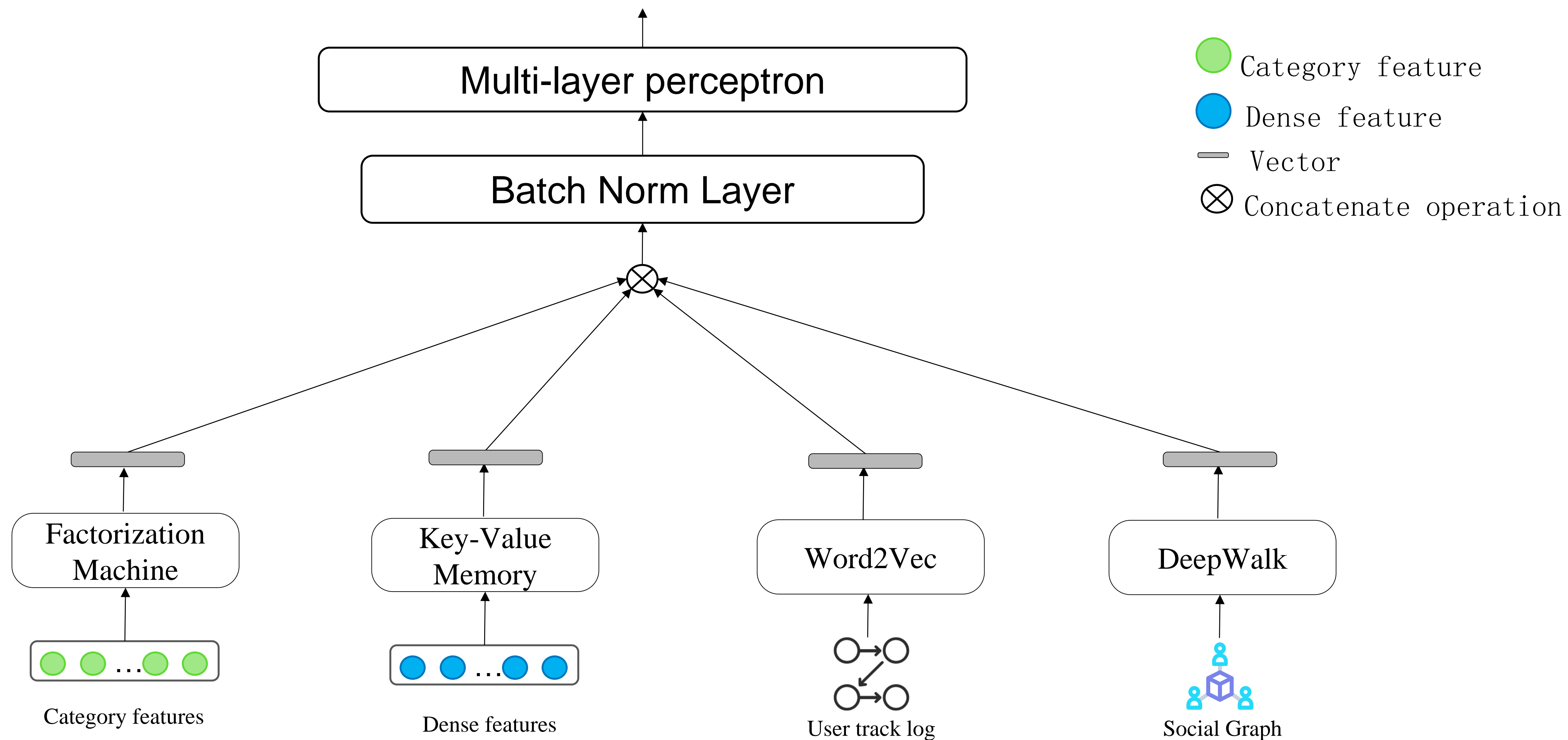
广告竞争总次数 - 广告竞争胜利次数

广告竞争总次数 - 广告竞争失败次数

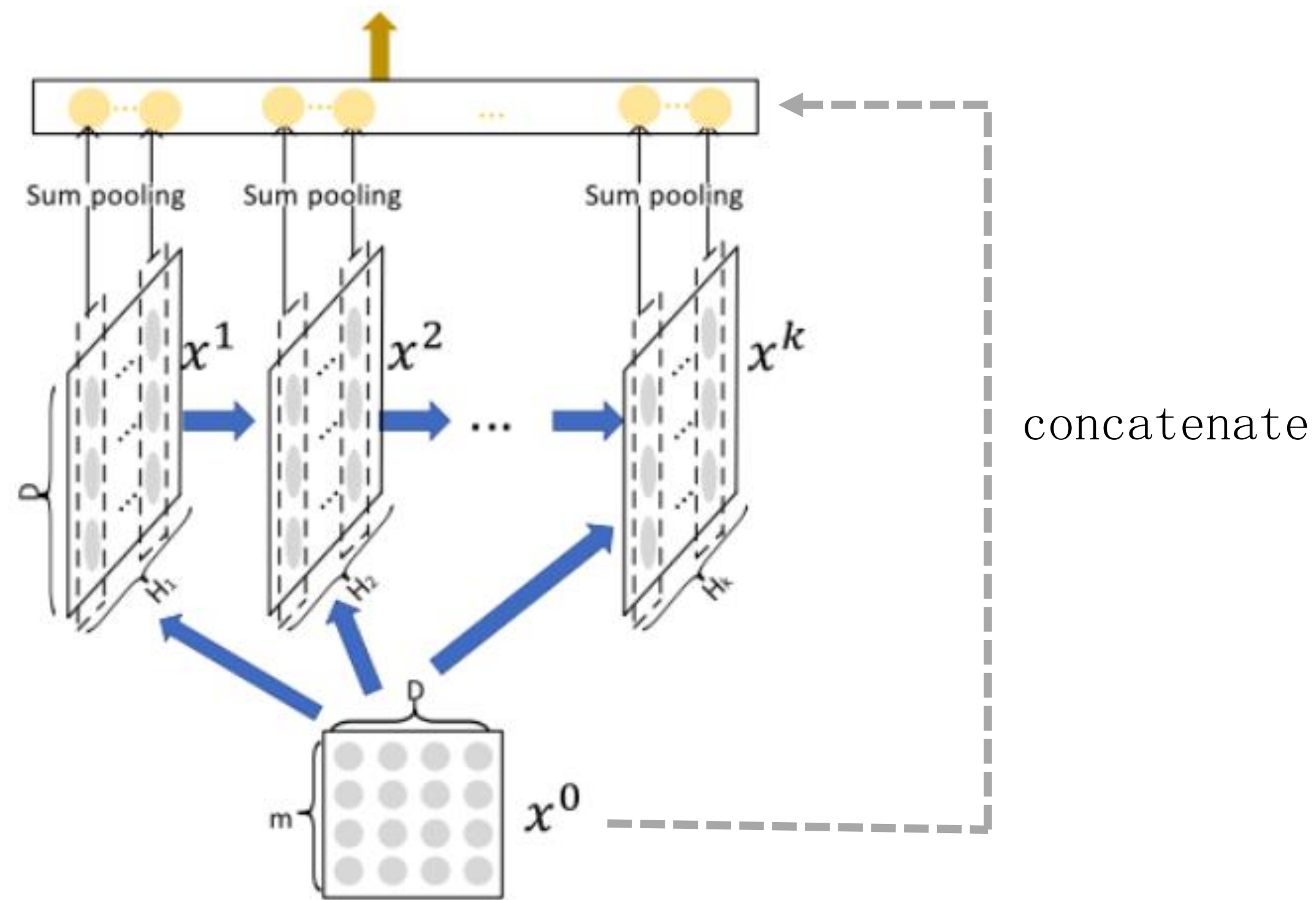
由之前特征提取方式能够得到当天新的广告竞争胜率、广告竞争胜利次数、广告竞争失败次数，我们称其为“假数值”，然后和当天真正广告竞争总次数进行交叉。



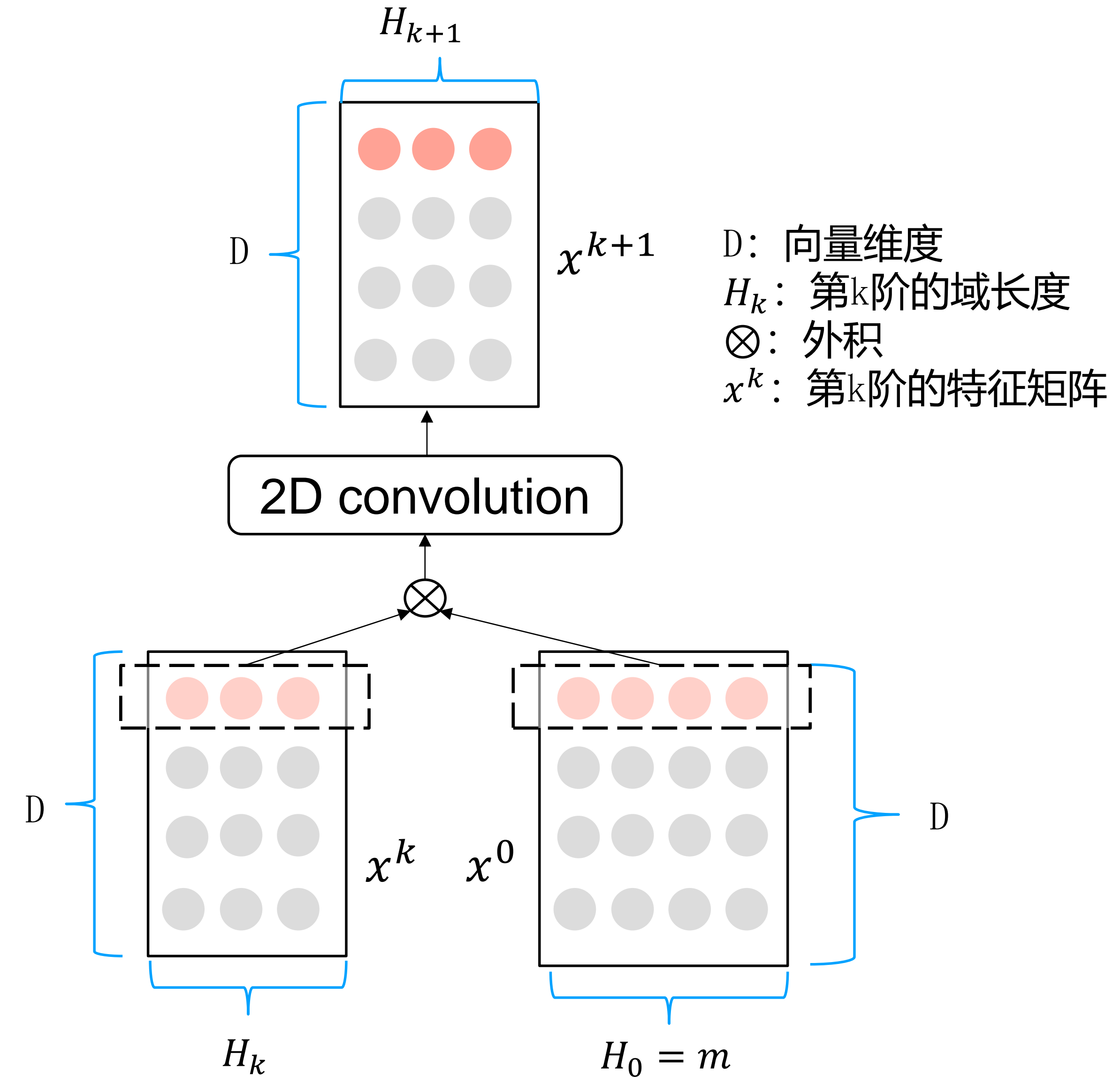
模型介绍



压缩交互网络 (Compressed Interaction Network, 简称CIN) 的神经模型 实现自动学习显式的高阶特征交互

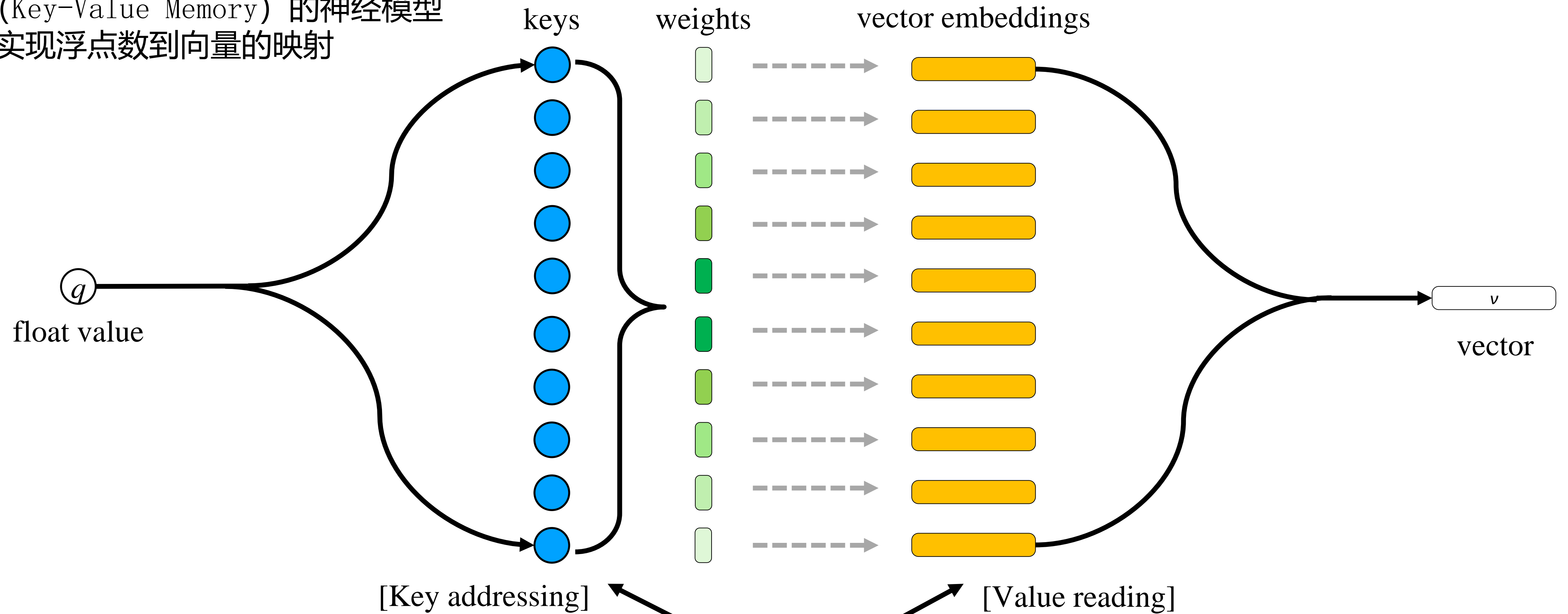


An overview of the CIN architecture



Key-Value Memory

键值存储 (Key-Value Memory) 的神经模型
实现浮点数到向量的映射



Parameter: $N=20$ $k_i = \frac{i}{N}$

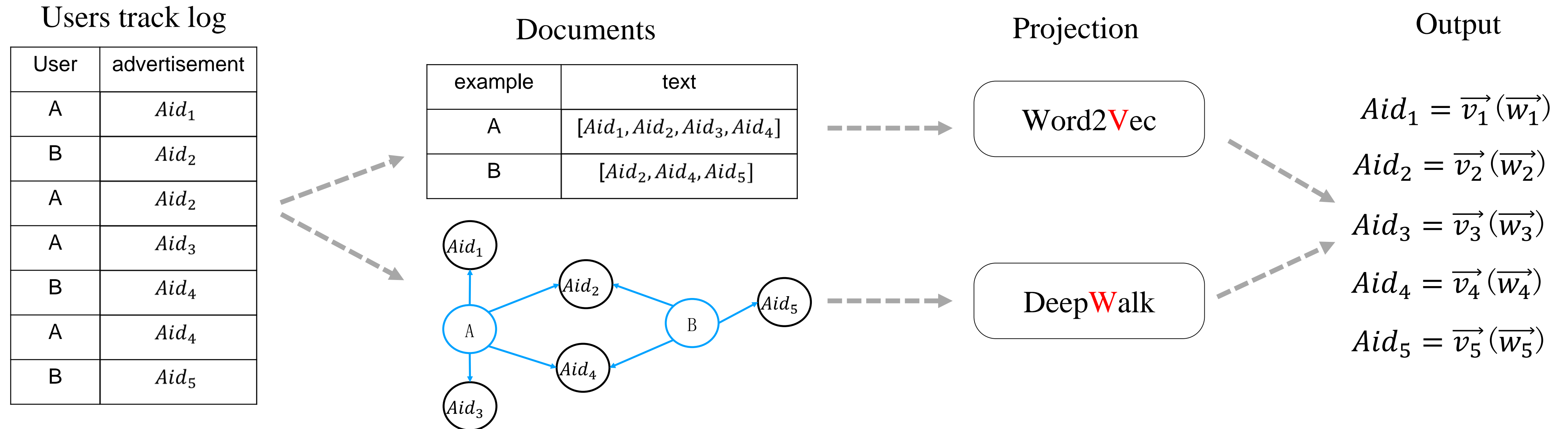
Key addressing: $w_i = \text{softmax}(\frac{1}{|q - k_i| + e^{-15}})$

Value reading: $v = \sum_{i=1}^N w_i v_i$

(k_0, v_0) (k_1, v_1) (k_2, v_2) (k_3, v_3) ... (k_N, v_N)

Key-Value Memory

用户的曝光记录转化为文本和关系图,使用word2vec和DeepWalk算法对广告进行嵌入, 例如:



两个问题:

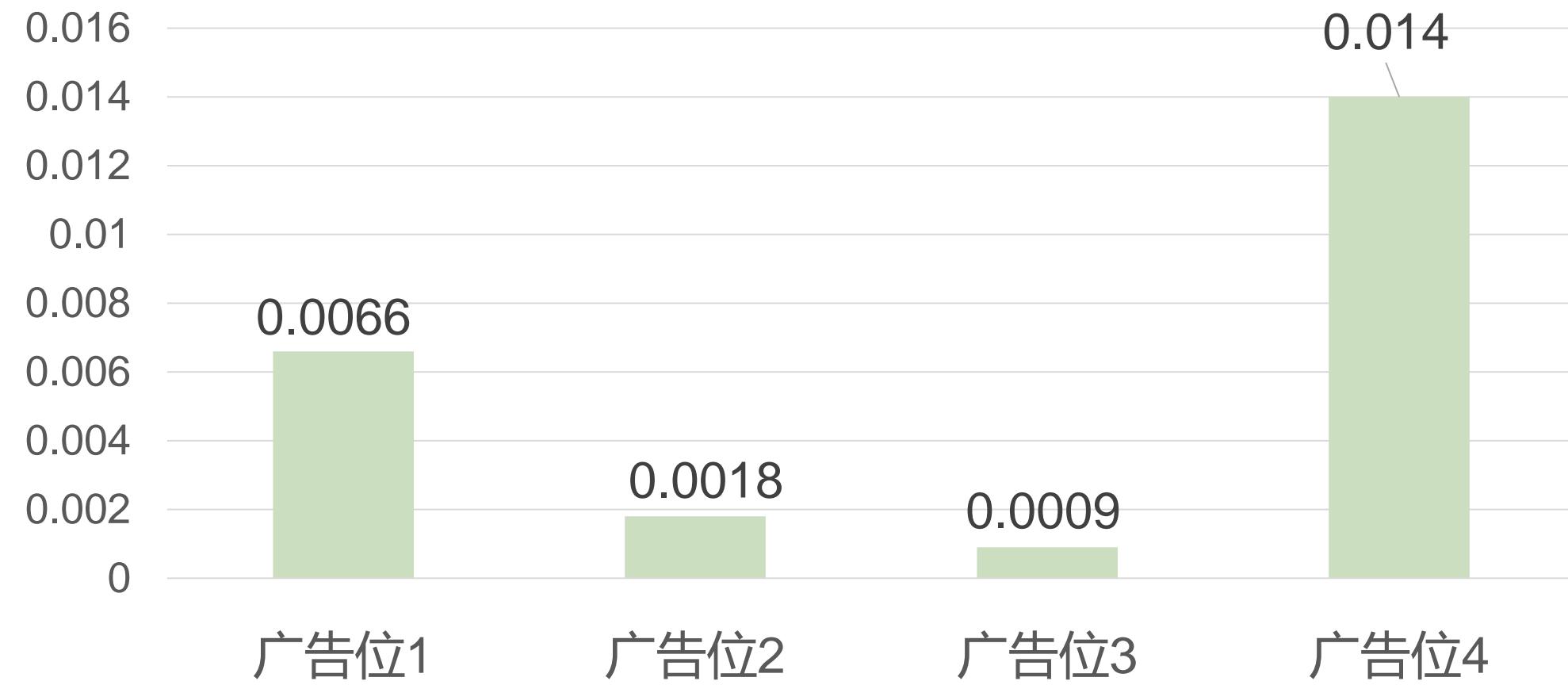
1. 只有在日志中曝光过的广告才会有相应的嵌入向量, 通过广告有无嵌入向量, 会泄露了无曝光广告的标签
2. 测试数据中存在曝光非0但无嵌入向量的广告, 这在训练集中是不存在的, 导致训练测试不一致

解决方法:

随机掩盖掉5%广告的嵌入向量

05 规则与模型融合

历史曝光数据分析



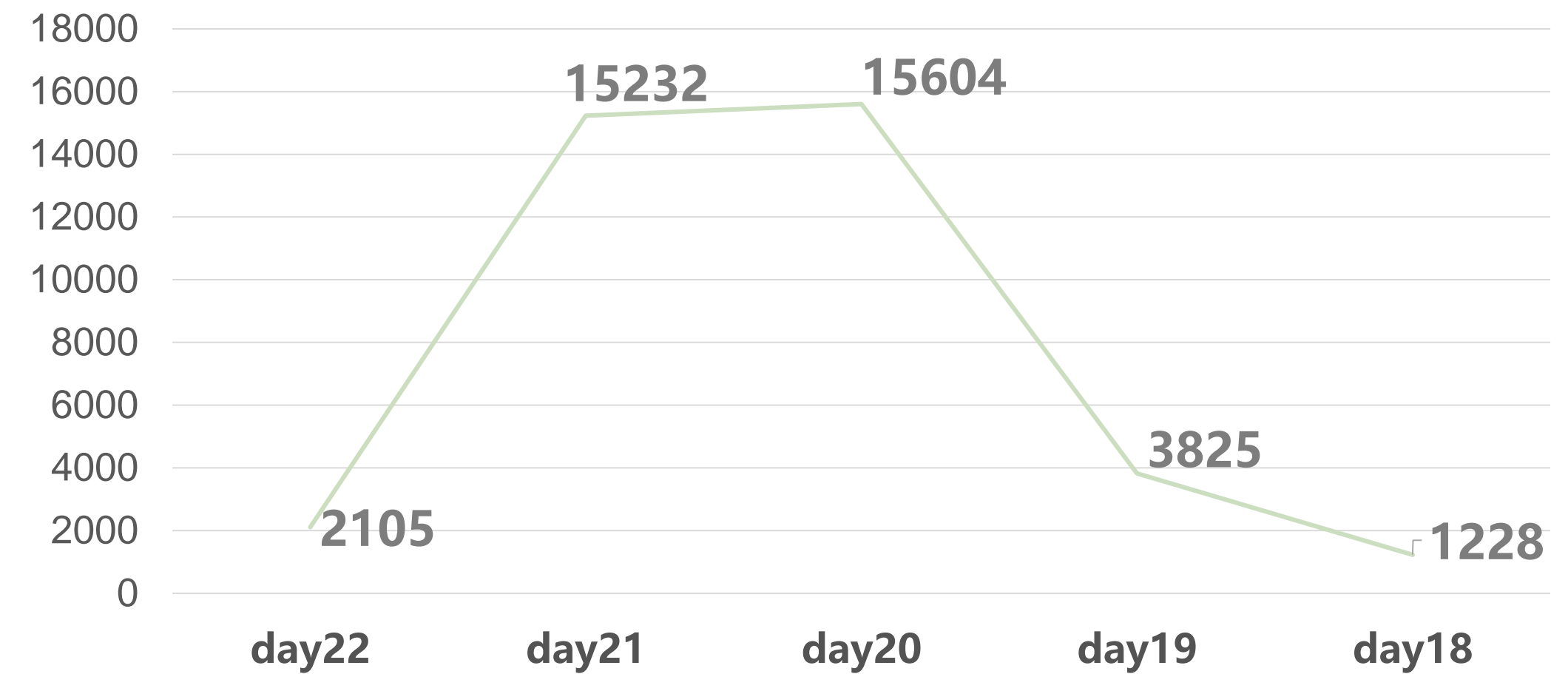
同广告不同广告位有不同的胜率

规则计算方式

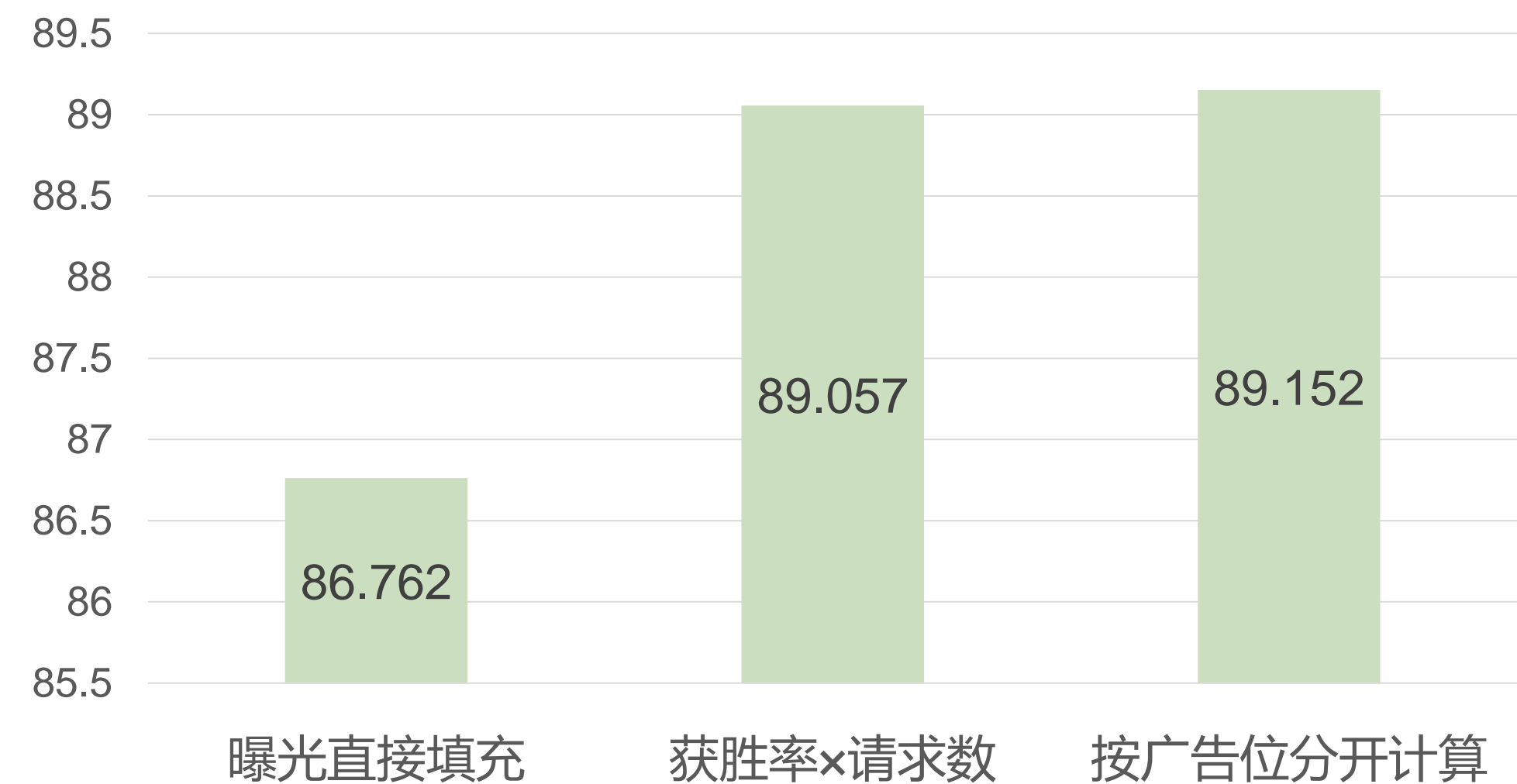
$$pred = \sum_{i=1}^4 history_rate_i \times request_i$$

$history_rate_i$: 广告在广告位 i 上的历史胜率。

$request_i$: 广告在广告位 i 上发出的请求总数。



同广告请求数在不同的日期存在差异



历史胜率history_rate计算方式

$$\text{history_rate} = \frac{\sum_{day=1}^{12} w_{day} \times \text{expose}_{day}}{\sum_{day=1}^{12} w_{day} \times \text{request}_{day}}$$

离预测当天越近的数据准确度越高，相应的权重也应该越大。

权重w计算方式

方式1: $w_{day} = \alpha * day + \beta$

方式2: $w_{day} = \alpha^{-day} + \beta$

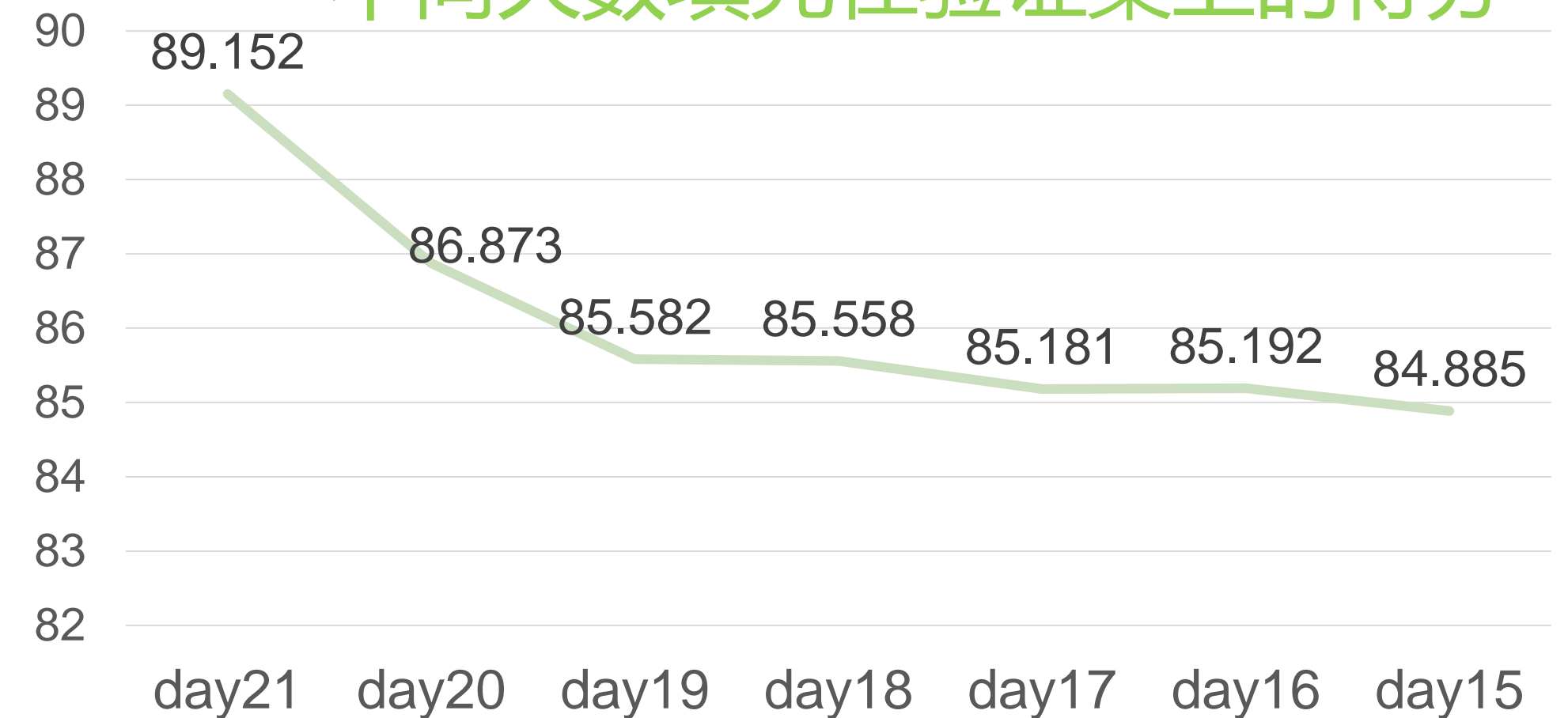
方式3: $w_{day} = \frac{1}{\alpha * day + \beta} + \gamma$

使用线性搜索寻找最优参数

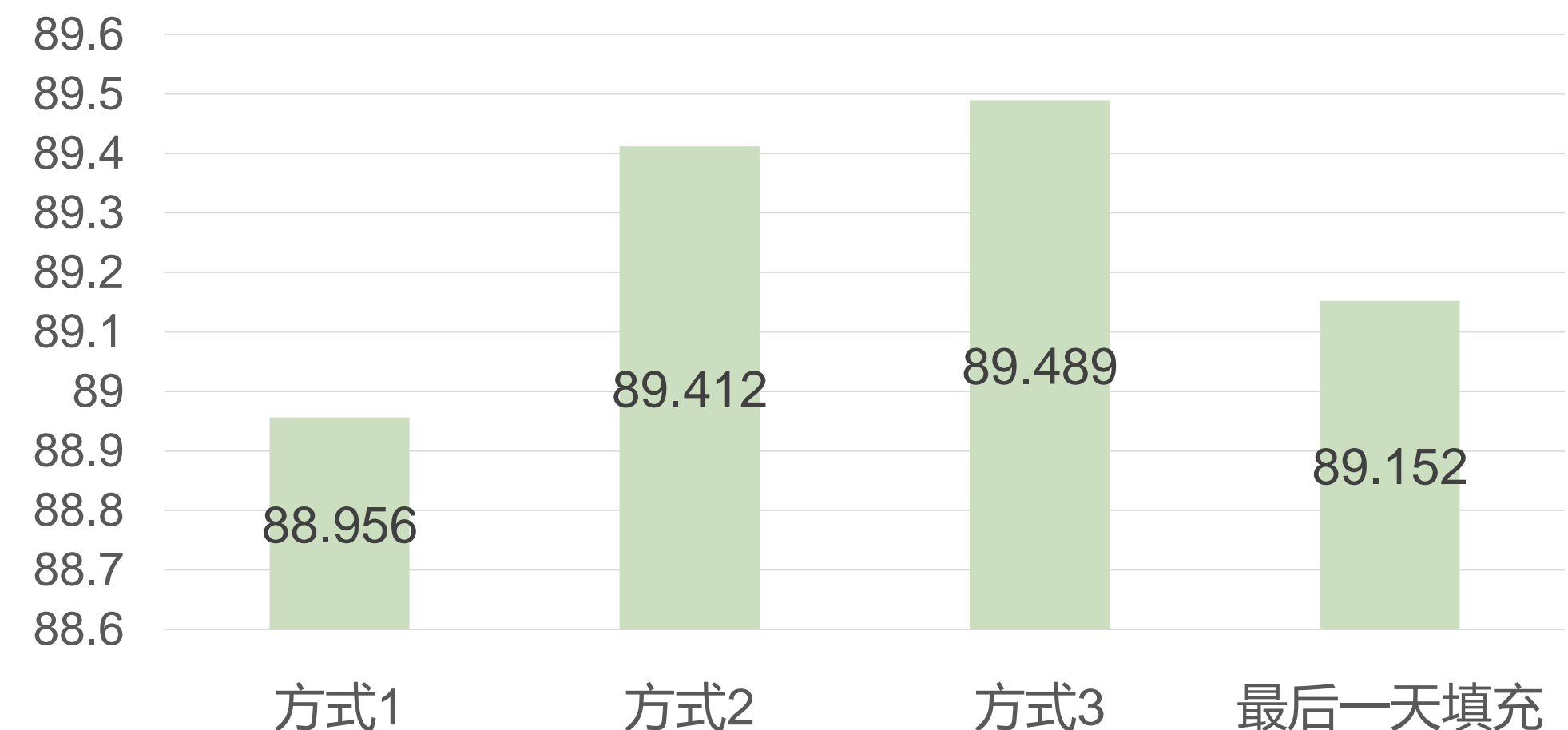
$$\bar{\theta} = \arg \max_{\theta} \sum_{i=1}^N \text{score}(y_i, \text{pred}_i)$$

右图表示了在最优参数下三总方式在验证集上的得分对比

不同天数填充在验证集上的得分



不同天数填充在验证集上的得分



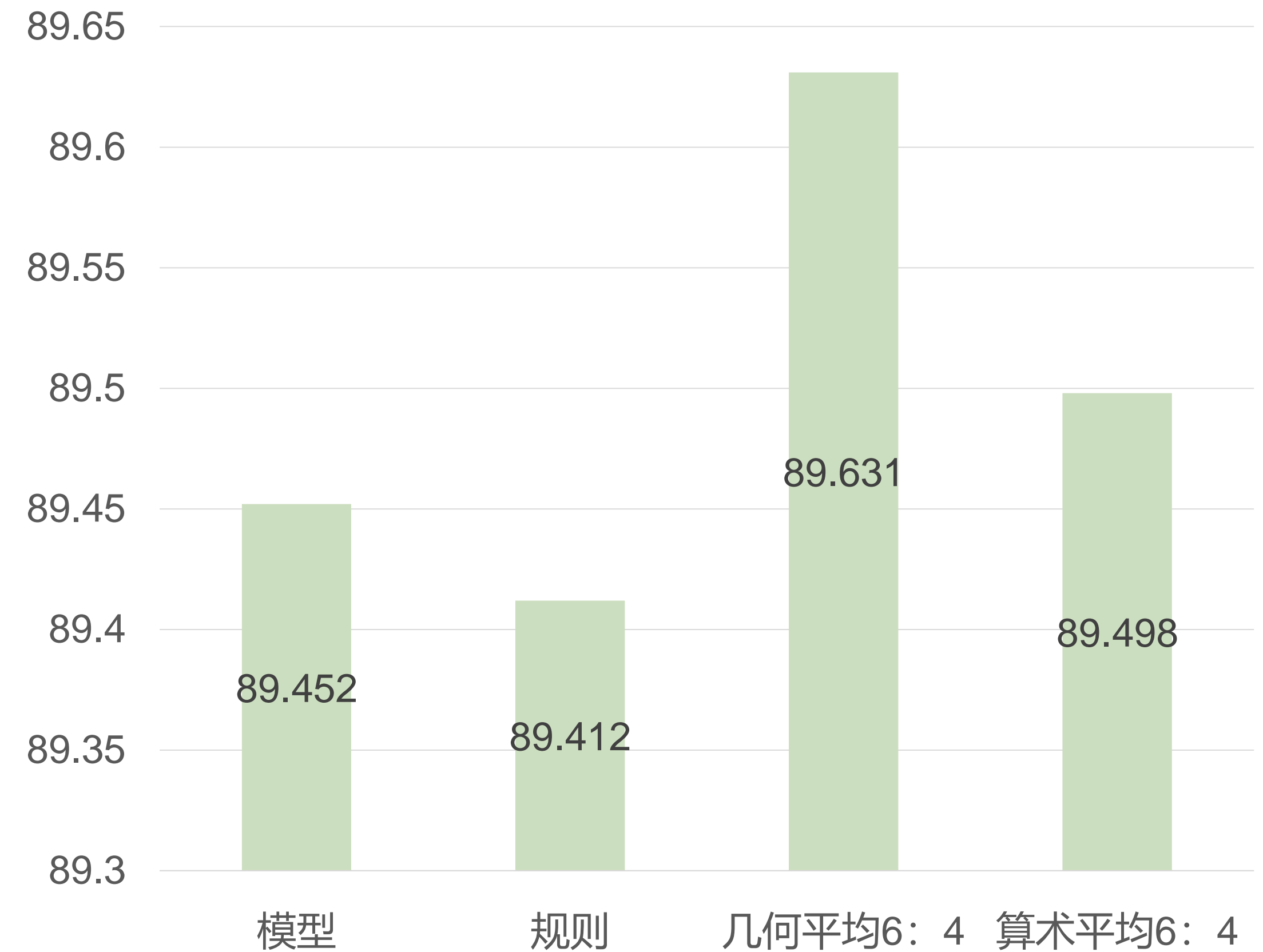
模型的两种融合方式：

算术平均： $pred = \alpha \times pred_a + (1 - \alpha) \times pred_b$

几何平均： $pred = pred_a^\beta \times pred_b^{(1-\beta)}$

由于SMPAE的评分规则，算术平均会使融合的结果偏大

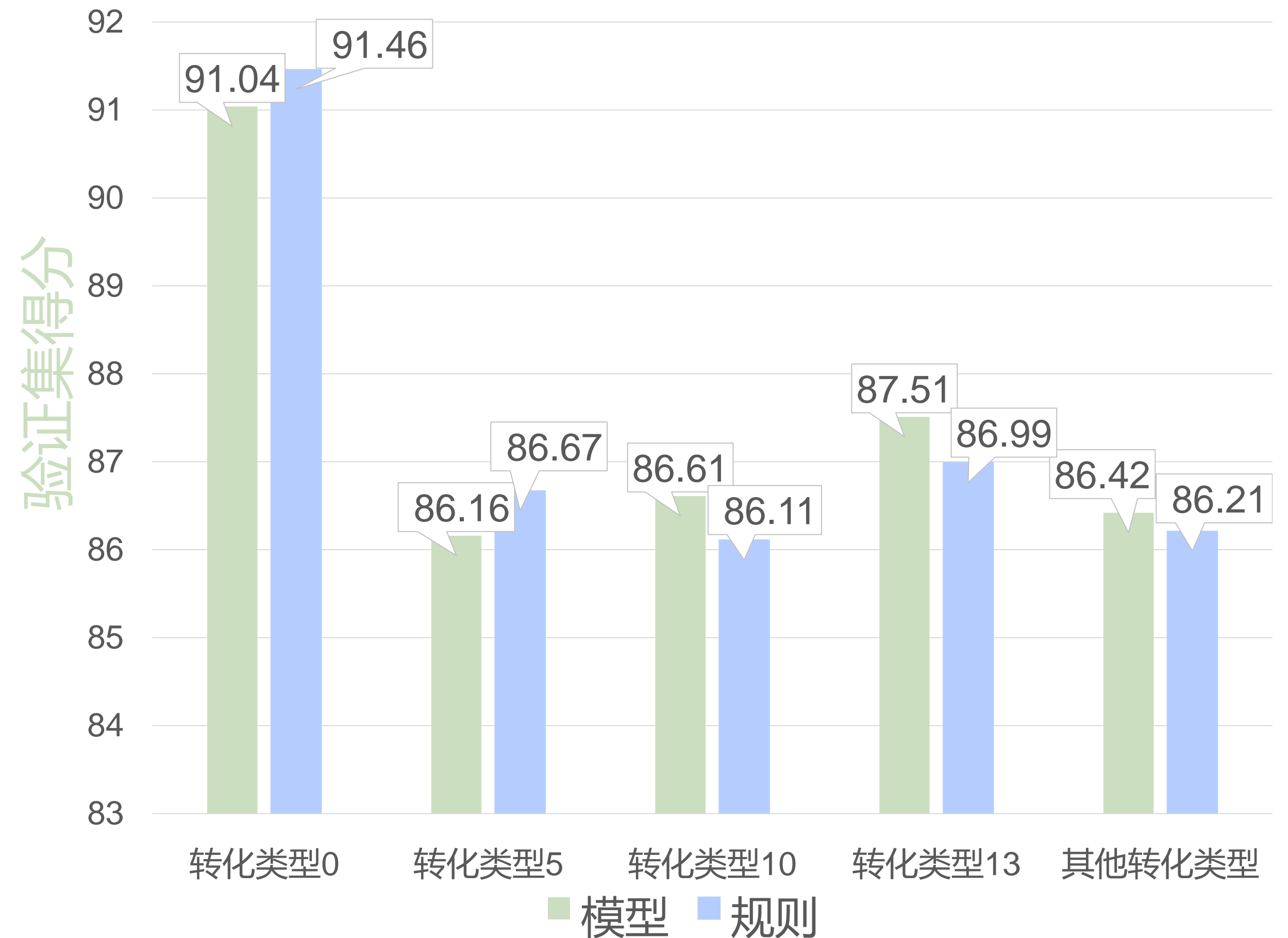
模型、规则以及不同融合方式验证集得分对比



更细致的融合方式

- 无论是模型还是规则，预测结果在不同的转化类型上得分差异都很大
- 模型和规则在不同的转化类型上得分也存在差异，右图表示了模型和规则在不同转化类型上的表现。

根据模型和规则在不同转化类型上的得分现，调整权重值，线上可以获得0.5个千的提升



分析与结果

- LightGBM**单模型**即可取得**top1**的成绩，领先第二名0.6个千。
- 模型模型融合，保证特征差异性，模型差异性，进一步扩大优势，领先第二名1.9个千。
- 模型和规则融合，优化旧广告的结果，锁定胜局，领先第二名2.5个千。

模型	线上分数	排名
LightGBM（1）	87.7789	1
LightGBM（2）	87.8	1
NN	87.46	9
LightGBM（1）+ LightGBM（2） +NN	87.9037	1
LightGBM（1）+ LightGBM（2） +NN + 规则	87.9683	1

队伍排名	队伍名称	最佳成绩
1	鱼遇雨欲语与余	87.9683
2	levy	87.7152
3	小迷弟	87.6654
4	人工智障	87.6017
5	慌呀哩	87.5916
6	长河落日圆	87.5218
7	ddw	87.521
8	DataAI	87.512
9	MindRank.ai	87.4539
10	小人国的蜗牛上分队	87.3457



总结与思考

主要创新

- 提出了一种基于Key-Value Memory的浮点数映射成向量的方法
 - 相较于直接使用浮点数，该方法保留更多的语义信息
 - 相较于数值 \times 向量的方法，该方法具有非线性特点
 - 相较于分桶并作为类别特征的方法，该方法的相邻向量具有相关性
- 解决Word2Vec和DeepWalk等无监督学习造成的数据泄露问题
 - 充分利用了曝光日志记录，基于用户行为对广告进行聚类

问题思考

- 本次比赛虽然使用到出价，但并没有将出价作为特征输入模型中。不同的出价其广告的竞争力会有所不同，将直接影响了曝光量，因此出价是非常重要的特征
- 加入约束条件保证模型的单调性
- 设计出价单调递增的模型，如输出为 $f(x) + |g(x)| \times \log(1 + \text{出价})$
- 本次比赛并没有用到用户属性相关数据，根据广告投放人群信息，或许可以获得更多有用的内容

总结

历时两个半月的腾讯广告大赛，非常感谢工作人员辛苦的答疑。感谢主办方提供真实的业务场景与数据，让我们能在比赛中学习到更多知识，在广告业务中做更多尝试。

THANKS

