中国科学: 信息科学 2019 年 第 49 卷 第 1 期: 17–41
SCIENTIA SINICA Informationis
论文

# 基于异构信息融合的广告响应预测方法

单丽莉*, 林磊*, 孙承杰

哈尔滨工业大学计算机科学与技术学院, 哈尔滨 150001
* 通信作者. E-mail: shanll@insun.hit.edu.cn, linl@insun.hit.edu.cn

收稿日期: 2017–08–21; 接受日期: 2018–01–31; 网络出版日期: 2019–01–02

国家自然科学基金青年科学基金项目 (批准号: 61602131)、国家自然科学基金面上项目 (批准号: 61572151, 61672192) 和国家高技术研究发展计划 (批准号: 2015AA015405) 资助项目

**摘要** 为了更有效地挖掘用户、上下文和广告之间的三维交互关系, 张量分解模型开始被用于解决实时竞价广告响应预测问题. 然而实时竞价广告响应预测面临严峻的数据稀疏和冷启动问题, 尤其是广告转化率预测, 单纯地依靠某类或某些信息很难有效地解决这些问题, 只有同时综合利用各种各样的异质、异构信息才能有效地应对这些问题. 本文面向张量分解模型, 提出了基于异构信息融合的综合解决方案来解决数据稀疏问题. 该方案针对不同信息的性能、类型、结构、存在方式和作用特点等, 提出了不同的融合策略和不同的实现方法, 提升了基于张量分解模型的广告响应预测方法的可靠性和准确性, 有效地缓解了需求方平台进行广告响应预测时面临的严峻数据稀疏问题. 在选定数据集上基于异构信息融合的模型预测性能与基准方法相比取得了显著的提升.

**关键词** 实时竞价广告, 广告响应预测, 张量分解, 异构信息融合, 数据稀疏, 冷启动, 预测方法

## 1 引言

由于实时竞价具有传统交易方式无法比拟的优势, 自 2012 年以来, 以开放的实时竞价交易 (real-time bidding, RTB) 为核心的程序化交易的崛起获得了业界的广泛关注, 被众多专业人士及专业机构预测为数字广告未来时代发展的大趋势. 由于实时竞价的程序化交易的实施涉及到信息获取、文本分析、信息检索、统计学、机器学习、自动分类, 以及优化算法等专业技术. 所以分别代表广告供求双方利益的独立技术平台和支持实时竞价的交易平台逐渐发展兴盛, 主要包括需求方平台 (demand side platform, DSP)、供应方平台 (supply side platform, SSP) 和广告交易平台 (Ad exchange, ADX). 在实时竞价交易市场中, 需求方对于流量的评价、选择和自主控制能力达到了极致, 因此需求方平台所面临的技术和算法的难度挑战也是前所未有的. 为了通过实时竞价优化广告主的投资回报率 (return on

**引用格式:** 单丽莉, 林磊, 孙承杰. 基于异构信息融合的广告响应预测方法. 中国科学: 信息科学, 2019, 49: 17–41, doi: 10.1360/N112017-00157
Shan L L, Lin L, Sun C J. Response prediction via integration of heterogeneous information (in Chinese). Sci Sin Inform, 2019, 49: 17–41, doi: 10.1360/N112017-00157

 

investment, ROI), 要求需求方平台具备更强的计算能力, 包括更精准的广告点击率和转化率预测 (统称为广告响应预测) 能力.

为了更准确地预测广告的点击率和转化率, 基于特征的机器学习模型被引入以便于综合各类特征, 并根据广告历史投放效果自动学习预测模型. 广义线性模型例如逻辑回归模型, 由于其实现简单、易于并行化、适用于大数据计算需求, 成为应用于广告响应预测的常用模型 [1~6]. 但由于线性模型在挖掘非线性关系上的限制, 随着特征的不断精细化, 数量的不断增多, 特征工程越来越繁琐而预测性能的提升也趋近于瓶颈, 能自动进行特征变换的模型开始逐步取代线性模型, 如因子分解模型 [7~9] 和决策树模型 [10~12], 甚至是能够做特征挖掘和发现的深度学习的方法 [13]. 基于因子分解的协同过滤模型, 例如矩阵分解模型或分解机模型, 在推荐系统中取得了令人瞩目的成绩 [14~18], 基于广告响应预测与推荐问题的相似之处, 更因为矩阵分解模型在处理稀疏数据、特征发现上的巨大潜力, 基于矩阵分解的协同过滤算法也被广泛地用于广告点击率预测 [7~9]. 传统的推荐问题是基于 〈 用户, 物品 〉 的二元交互关系挖掘, 解决的是用户对物品的历史反馈矩阵缺损数据填补问题. 然而与传统的推荐问题不同, 广告响应是在给定 〈 用户, 广告和上下文 〉 情况下, 3 个对象之间交互作用的结果. 因此已经有研究者 [19~22] 使用张量模型描述广告响应问题, 并将其转化为立方分解问题, 使用三阶张量分解算法进行求解. 虽然张量分解模型在挖掘用户、广告和上下文三者之间的交互关系方面具有明显的性能优势, 然而在解决面向需求方平台的实时竞价广告响应预测问题时, 仍然面临着严峻的数据稀疏和冷启动问题. 由于需求方只能获得竞价成功的广告投放反馈数据, 因而无法拿到所有的流量, 更不可能尝试投放所有的用户来获得所有用户的反馈数据. 所以可以获得的历史投放展示与所有可能的投放展示的比例是非常稀疏的. 用户、上下文环境和广告数量庞大, 导致在大量的对象上无法获得足够数量的真实投放反馈或者反馈稀少, 我们称这种现象为数据稀疏. 数据稀疏会导致训练样本分布偏离真实样本分布, 造成模型训练过拟合和模型预估不准确不可靠. 最糟糕的情况是没有任何历史投放数据可以借鉴, 尤其对于刚刚加入系统的新广告、无法识别的新用户和新的上下文环境, 没有任何关于它们的投放历史数据可以借鉴, 被称为冷启动问题, 冷启动问题更是增加了广告响应的预测难度.

为了缓解数据稀疏带来的影响, 大量的数据平滑技术被用来解决这个问题 [4, 7, 11, 23, 24], 然而, 首先, 很多方法都是在模型级别的融合, 不能直接被用于张量分解模型. 特征的融合可以在特征级别进行, 也可在模型级别进行, 到底什么样的特征适合在特征级别融合, 什么样的特征适合在模型级别融合, 现有文献并没有进行深入地研究和探讨. 其次, 如果能够同时利用点击和转化反馈信息, 充分发挥它们之间的互补互促作用, 则有望提升广告响应预测的性能, 然后目前并没有这方面的研究工作可以借鉴. 点击数据已经是很稀疏了, 而转化数据的数量更是微乎其微, 因此转化率的预测面临的稀疏挑战更大. 与未点击样本相比, 点击样本具有更高的概率被转化, 如果能有效地利用点击信息来辅助转化的预测, 则能在一定程度上缓解数据稀疏, 有望提高转化预测的性能. 在进行点击率预测时, 历史点击信息有可能是误点击或点击欺诈, 而历史转化信息可以为点击预测提供更加稳定可靠的用户兴趣倾向性信息, 既点击又转化了的样本比只被点击而未转化的样本应该具有更高的概率被点击, 因此有效利用转化信息将有望提升点击率的预测性能. 最后, 很多方法只是提出了某类或某些信息的解决方案, 没有给出融合各类异构信息的完整解决框架, 要将各类异构信息协调统一地融合起来缓解数据稀疏问题, 必须针对各类信息的特点采用不同的融合方法, 既能有效地利用信息又要考虑与其他信息的协调性, 还要考虑实践可行性. 为了解决使用张量分解模型进行实时竞价广告响应预测时面临的严峻的数据稀疏问题, 本文对各种不同的信息进行深入分析, 针对不同类型、结构和形式的数据, 提出综合实现方案和具体实践方法, 综合利用异质、异构信息有效缓解需求方平台进行广告响应预测面临的数据稀疏问题.

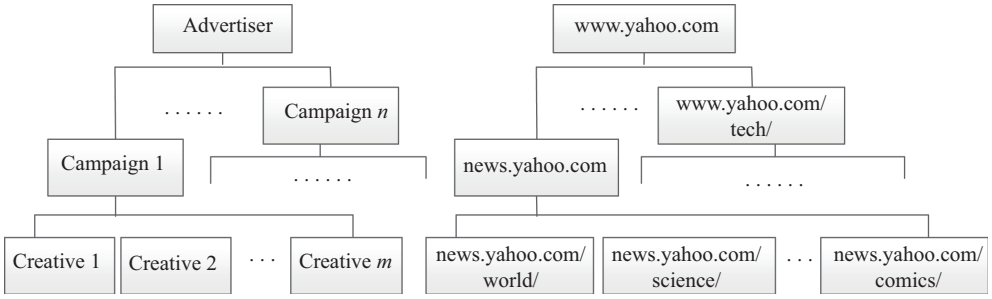


**图 1 广告和网页固有层次结构示意图**

**Figure 1** An illustrative example of data hierarchies

## 2 相关工作

为了缓解数据稀疏给广告响应预测带来的影响, 研究者们尝试了大量的数据平滑技术. 最常用的就是利用对象之间固有的层次信息 [4,7,11,25~27], 逐步提高特征的粒度, 以缓解数据稀疏. 广告和网页通常都存在天然的层次关系, 如图 1 所示. 例如, 一个广告活动 (campaign) 通常设计不同的广告创意 (creative), 一个广告主通常会发起不同的广告活动; 网页根据 URL 地址具有天然的层次结构. 如果一个广告的历史投放数据非常少, 或者没有任何数据, 基于它与兄弟节点或父子节点应该具有相似的点击率的假设, 利用层次结构中的 "兄弟" 关系或 "父子" 关系, 从历史观测数据相对丰富的兄弟节点或父节点 "借" 信息, 这种方法常见于基于极大似然估计的概率模型和基于特征的机器学习模型 [4,7,11,23~26]. 例如 Lee 等 [23] 利用广告与网站固有的层次结构, 基于假设同处在一个层次节点里的所有对象具有相似的转化率, 按 3 类对象的层次组合来估计转化率, 最后用逻辑回归模型组合一次展示中涉及的所有转化率. Wang 等 [26] 也利用层次关系网页和广告的层次关系构建基于 $\beta$ – 二项分布的两级复合层次 Bayes 模型, 对点击率进行平滑处理. Oentaryo 等 [24] 和 Menon 等 [7] 都采用基于子节点与父节点的点击率相似的假设, 将层次信息组合到分解机模型的正则化项中, 来优化模型.

对于没有天然层次关系的对象, 例如用户, 通常利用聚类技术构造类别甚至是层次关系, 提高特征粒度 [23,26,28], 或者是提取高层次的较粗粒度特征, 这些特征常常被称为附加特征 [7,9], 利用相同特征投放效果的相似性来相互借鉴辅助预测. Menon 等 [7] 使用附加信息的线性组合再结合矩阵分解模型来提高特征粒度, 为发布商解决响应预测. 还可以通过设计相似度度量办法, 找出相似对象集, 借助相似集的对象的信息来辅助预测 [29], 这种方法与聚类方法有相似之处.

除了将细粒度特征提升为较粗粒度的特征之外, 很多研究者利用投放效果在时间上的连续性使用时间特征进行平滑 [9,26,27,30], Wang 等 [26] 利用短期时间内发生的事件具有一定的连续性和相关性的特点, 利用当天之前短期内的展示数量和点击数量对当天的数量进行指数平滑. 然后, 利用网页和广告固有的层次关系构建基于 $\beta$ – 二项分布的两级复合层次 Bayes 模型, 对点击率进行平滑处理. Kota 等 [27] 也是利用历史时间段的投放效果对现在时间段投放的衰减影响来进行平滑. Agarwal 等 [30] 设计了一个时空模型 (spatio-temporal model) 为内容推荐问题估计点击率, 利用动态的 Γ–Poisson 模型 (gamma-Poisson model) 追踪特定位置的 CTR (click-through rate) 随时间的变化并通过动态线性回归组合相关位置的信息.

对于新广告的预测, 常常使用与上述类似的方法 [31~33], 例如: Regelson 等 [31] 使用广告关键字的文本相似度信息对广告进行层次聚类, 利用相似广告辅助预测. 除此之外, 基于广告内容挖掘相似广告也是常用的方法 [33,34]. Cheng 等 [34] 从显示广告中提取图片和 Flash 动画的多媒体信息, 例如, 亮

度、饱和度、色调、纹理等特征, 提高特征粒度, 最终提升新广告的点击率预测质量.

虽然直接应用张量分解模型来解决广告响应预测的方法还相当的稀少, 但由于矩阵分解模型在推荐系统中的广泛应用, 以矩阵分解为代表的协同过滤模型融合附加信息缓解数据稀疏的来解决广告响应预测问题的技术应用却比较广泛. 除了用户和物品等对象的特征之外, 研究者们也同样尝试使用粗粒度的类别信息、层次信息或聚类信息等来进一步缓解数据稀疏. 具体到融合方法上, 研究者们针对协同过滤模型针对性地提出了各种不同的方法, 总体来说可以归纳为 3 种. 第 1 种方法是特征级别的集成. 先将附加信息按对象分类, 然后将特征隐因子向量按类别分别与对象隐因子向量线性组合来表示对象. Koren 等 [14,35] 为了将类型各异的附加信息集成到矩阵分解模型, 使用了类似的方法. 第 2 种常用的方法是模型级别的集成, 即将擅于挖掘交互特征的矩阵分解模型与擅于处理线性组合特征的线性模型集成在一起. Menon 等 [7,36] 和 Yang 等 [37] 使用了类似的方法. 第 3 种方法是综合使用以上两种方法, 按照附加特征的性质和类别, 或在特征级或在模型级做集成 [8,38]. Chen 等 [16,38] 组合了以上两种集成附加信息的方法 [7,14,36,37], 一方面采用与 Koren 等 [14] 相似的办法将用户和物品各自的所有特征组合起来表示用户和物品, 利用矩阵分解进行预测, 另一方面, 又使用与 Menon 等 [7,36] 和 Yang 等 [37] 相似的方法将所有的特征输入线性回归模型进行预测, 最后将两个预测值相加. 另外, 除了用户特征和物品特征, 此模型还同时考虑了全局特征, 即与用户和物品对相关的特征产生的偏置. 作者将此模型命名为基于特征融合的矩阵分解 (feature-based matrix factorization). 这种基于特征融合的矩阵分解模型既考虑了特征之间的线性关系, 也考虑了非线性关系, 实现了附加信息的集成, 有助于缓解数据稀疏和冷启动问题. 因此, 本文将这种特征组合方法做适当地改进后引入张量分解模型作为融合方案的一部分, 以有效地缓解数据稀疏问题.

虽然基于普通对象特征的融合方法在以上基于协同过滤方法解决广告响应预测问题相关的文献中都有涉及, 但是, 针对单值特征和多值特征的不同处理方法, 以及对于类别特征和数值特征的具体处理细节并没有给出详细的方案. 而且很多融合方法都是模型级别与模型相关的, 所以不能被直接应用于张量分解模型, 本文提出的方法将普通对象特征与张量分解模型融合, 同时又能保证不同对象特征之间固有的交互关系. 另外, 现存的层次聚类特征的融合方法往往都比较复杂, 本文提出的策略是利用层次聚类特征之间的相似性, 使用与普通多值对象特征类似的方法来利用这类信息, 方法简单而有效. 最后, 本文还实现了点击信息和转化信息的同时融合, 以缓解转化率预测时面临的更严重的数据稀疏和点击预测时遇到的点击噪声问题.

## 3 基于张量分解的广告点击率预测问题的形式化描述

给定广告历史投放日志集合 $T=\{(d_i,r_i)|i=1,\ldots,n\}$, 其中, $d_i$ 是形如 $(u,p,a)$ 的三元组, 表示广告 $a$ 在上下文 $p$ 下, 被展示给了用户 $u$, 被称为一次展示或曝光 (impression). 设 $S=U\times P\times A$ 是所有可能曝光的集合, 其中, $U$ 是所有用户的集合, $P$ 是所有上下文的集合, $A$ 是所有广告的集合. 设 $D$ 是历史曝光的集合, 则有 $d_i\in D$ 且有 $D\subseteq S$. $r_i\in\{0,1\}$, 当 $r_i=1$ 时, 表示曝光产生了点击或转化即为正反馈, 而当 $r_i=0$ 时, 表示曝光未产生点击或转化即为负反馈. 此处的上下文的定义依赖于具体的应用环境, 例如, 在搜索广告中, 上下文可以定义为用户输入的查询; 在上下文定向广告中, 上下文可以定义为用户浏览的网页; 在显示广告中, 上下文可以定义为广告位的大小、位置等; 在移动广告中, 上下文可以定义为用户的地理位置等.

广告点击率预测的任务是, 学习一个函数或模型 $\hat{R}:U\times P\times A\to\mathbb{R}$, 对于给定的广告曝光 $(u,p,a)$, 预测此次广告曝光的用户响应, 函数值表示广告 $a$ 在上下文环境 $p$ 下, 展示给用户 $u$ 时, 产生点击或

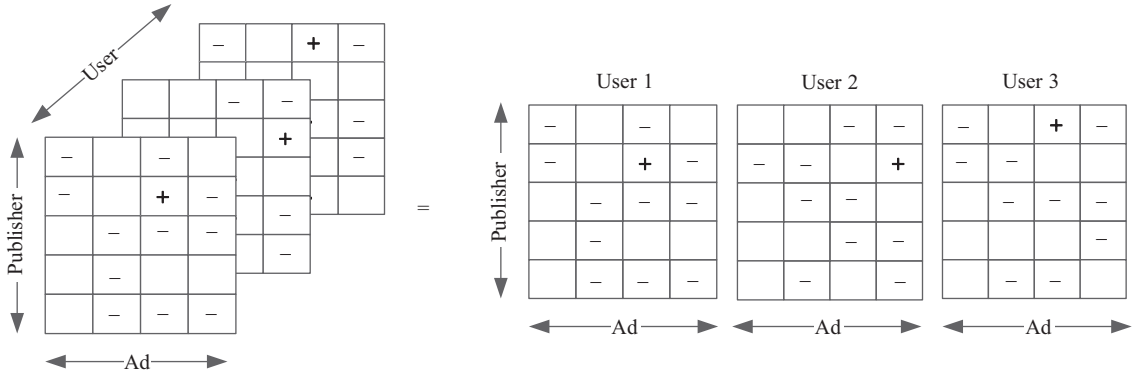


**图 2 广告响应预测张量表示图**

**Figure 2** The third-order response tensor

转化的可能性大小或概率.

为了解决这个问题, 本文根据历史广告投放数据构建有数据缺失的三阶张量 $\mathcal{R} \in \mathbb{R}^{|U|\times|P|\times|A|}$ 来表示 ⟨ 用户, 广告, 上下文 ⟩ 三元组中三者之间的关系, 如图 2 所示. 张量中坐标 $(u,p,a)$ 上的值有 3 种情况, 如果 $(u,p,a)$ 在训练集中出现过, 并且是正反馈则用 "+" 表示, 若是负反馈则用 "−" 表示, 若未出现过, 即 $(u,p,a) \notin D$, 其值为缺失, 用空白表示. 广告响应预测的任务就是补充这些缺失的数值, 这些值对应广告曝光 $(u,p,a)$ 的响应预估值, 与点击率或转化率成正比. 本文将应用张量分解技术来实现张量的重构, 实现缺失数据的估计.

由于基于历史广告投放反馈构建的原始三阶张量 $\mathcal{R}$ 是有严重的数据缺失的三维数组, 所以无法直接使用张量分解算法, 只能采用近似算法求解. 数据缺失张量低秩逼近的基本模型可以描述如下: 假设三阶张量 $\mathcal{R}$ 近似为一个秩 $k$ 张量, 其已知真值的元素子集为 $D$, 则求一个秩不超过 $k$ 的张量 $\hat{\mathcal{R}}$, 使其与 $\mathcal{R}$ 在已知历史投放响应数据集 $D$ 上具有最小的平方逼近误差:

$$\min_{\mathrm{rank}(\hat{\mathcal{R}})\leqslant k} \sum_{(u,p,a)\in D} (r_{u,p,a} - \hat{r}_{u,p,a})^2 \triangleq \|\mathcal{R} - \hat{\mathcal{R}}\|_D^2, \tag{1}$$

注意式 $\|\cdot\|_D^2$ 只表示一个度量, 而不是范数. 本文称 $\hat{r}_{u,p,a}$ 为真值 $r_{u,p,a}$ 的估计, 将采用张量分解算法来计算 $\hat{r}_{u,p,a}$.

## 4 面向张量分解模型的异构信息融合方法

### 4.1 信息的异构性

影响广告响应行为的信息的异构性体现在信息的类型多样、结构复杂和形式多变. 具体体现在以下几个方面.

• 既有类别特征也有数值特征. 类别和数值特征是最常用的两类特征, 例如用户的性别和地区是类别特征, 而广告位的高度和宽度是数值特征. 类别特征和数值特征通常要采用不同的方式处理和利用, 才能充分发挥特征的有效性.

• 既有单值特征也有多值特征. 通常特征都只有一个值, 但有些特征有多个值, 例如用户标签就是多值特征, 有的用户有多个标签, 有的用户只有很少的标签, 而有的用户甚至没有标签. 很多研究工作

将多值特征转换为多个单值特征来处理, 但这样会增加特征空间的维度使计算更复杂. 既能够不增加计算代价又能够有效利用多值特征的融合方法很值得我们研究.

• 既有结构化的层次信息和半结构化的聚类信息也有无结构的独立特征. 通常利用的网页和广告之间的固有层次信息属于结构化的信息, 而用户或其他的聚类信息属于半结构化的, 另外还有其他的无结构信息, 这些信息的结构不尽相同, 无法采用简单统一的融合方法加以利用, 为综合利用不同结构的信息提出了新的挑战.

• 既有预测时可获得的实时信息, 如用户画像和广告信息, 又有预测时不可获得的滞后可得的点击和转化反馈信息. 历史点击样本很稀疏, 而转化样本更稀疏, 如果能有效地利用历史点击信息来辅助未来的转化预测, 则有望提高转化预测的性能. 反之, 历史点击信息通常都包含点击欺诈和误点击等点击噪声, 而历史转化信息几乎是没有噪声的, 比历史点击信息表现出了更强的用户兴趣倾向性, 如果能有效地利用历史转化信息, 则有望缓解点击噪声带来的影响, 提升点击预测的性能. 然而点击信息和转化信息与其他实时可得的信息不同, 这些信息是在预测时不可实时获得的, 否则也就不用预测了, 所以想同时利用两种历史信息辅助预测就要使用与其他特征不同的融合策略.

总之, 一方面, 不同类型、结构和形式的信息具有不同的性质和特点注定要采用不同的融合方法, 另一方面, 又要将它们协调统一地纳入一个框架中集成来缓解数据稀疏问题, 提升系统广告响应预测性能.

本文根据可利用信息的类型、性质、结构、形式及其在解决数据稀疏问题中所起的作用等多方面的因素将所有信息分成 4 个类别: 普通对象特征、层次和聚类信息、全局特征及点击反馈信息和转化反馈信息. 不同类别的信息采用不同的融合方法. 下面按不同的特征类型分别进行详细的阐述.

### 4.2 不同异构信息的融合策略

#### 4.2.1 普通对象特征的融合

普通对象特征是指与用户、广告或上下文对象三者之一有关的特征. 普通对象特征的特点是预测时实时可获得且无结构. 例如: 用户的性别、个人关注、位置等; 广告位的大小和样式等; 广告的主题和创意等. 如果在张量分解模型中仅仅使用各个对象的标识符作为特征, 可能会遭遇严重的数据稀疏问题. 例如: 用户在历史上只被投放了一次广告, 而且未点击, 因此按照历史数据预估, 得到的结论将都是不点击. 显然只根据该用户的一次广告投放反馈来预测今后用户对于所有广告的响应都是不点击, 是片面且不可靠的. 那么如果使用用户的性别、年龄、兴趣爱好、职业甚至是购买历史等用户特征做参考, 就可以借鉴其他具有相同特征甚至是相同购买历史的用户的历史反馈作参考来预测当前用户未来对广告的响应行为. 这些特征对应的对象集合大了很多, 所以这些特征比用户的标识符特征 (只对应一个用户) 粒度大得多, 历史投放反馈数据也就更丰富得多, 也能得到更可靠的预测, 因此本文融合普通对象特征的隐因子向量来表示对象, 而不使用对象的标识符.

普通对象特征的融合面临两个问题. 第一, 普通对象特征可能是单值特征也可能是多值特征, 如用户的性别是单值特征而个人关注可以是多方面的多值特征. 对于多值特征, 如果简单地将其转换为单值特征处理, 由于多值特征数量不固定, 当数量较多时, 可能远远超过所有其他单值特征的数量, 简单线性组合的结果会严重影响其他单值特征发挥应用的作用. 为了解决这个问题, 将单值特征和多值特征分开处理, 将多值特征作为一类特征, 经过归一化处理后再与其他的单值特征进行组合. 第二, 普通对象特征可能是类别特征也可能是数值特征, 例如用户性别是类别特征, 而广告位的大小是数值特征. 因此, 在融合时需要全面考虑不同类型的异构特征并给出统一的融合方法. 为了实现类别特征和数值

特征的统一处理, 本文采用 “特征的隐因子向量乘以特征值” 的统一形式处理类别特征和数值型数据, 实现了类别特征和数值特征隐因子向量学习过程的形式统一.

这里以用户 $u$ 的表示为例, 假设用户 $u$ 只有一个多值特征即标签, 用户拥有的标签数量是可变的, 设 $T(u)$ 是用户 $u$ 的所有标签的集合. 除此之外该用户还拥有多个单值特征, 设 $C(u)$ 是该用户所有其他单值特征的集合. 于是用户 $u$ 表示为

$$u_u = |T(u)|^{-0.5} \sum_{i \in T(u)} t_i + \sum_{c \in C(u)} \alpha_c u_c, \tag{2}$$

其中, $T(u)$ 是用户 $u$ 的所有标签的集合, $i$ 表示用户 $u$ 的某个标签, $t_i \in \mathbb{R}^f$ 表示标签 $i$ 对应的隐因子向量, $|T(u)|^{-0.5}$ 是多值特征的归一化系数. $C(u)$ 是该用户所有单值特征的集合, $c$ 表示用户 $u$ 的某个单值特征, $\alpha_c$ 是特征 $c$ 的特征值, $u_c \in \mathbb{R}^f$ 是对应特征 $c$ 的隐因子向量. 首先, 对于多值特征即标签, 采用独立的融合方法, 即式 (2) 中的第 1 个累加项, 将多值特征作为一类特征, 经过归一化处理后再与其他单值特征进行组合. 这样就能有效地避免多值特征数量不固定对模型训练造成的扰动. 其次, 无论是类别特征还是数值型特征, 都采用统一的表示形式: 特征的隐因子向量乘以特征值, 即 $u_c \cdot \alpha_c$. 对于 $\forall c \in C(u)$, 特征 $c$ 可以是类别特征也可以是数值特征, 当特征 $c$ 是类别特征时, 例如用户的性别, 其特征值 $\alpha_c$ 为 Boole 值, 如果该用户拥有特征 $c$, 那么 $\alpha_c = 1$, 否则 $\alpha_c = 0$. 当特征 $c$ 是数值特征时, 例如用户的历史网购次数, 其特征值 $\alpha_c$ 为实数 (通常是归一化处理为 0 到 1 之间的实数). 无论是类别特征还是数值特征, 它的隐因子向量都是 $u_c$, 特征的融合形式都是 $u_c\alpha_c$. 上下文 $p$ 和广告 $a$ 也进行类似的表示, $C(p)$ 定义为上下文 $p$ 的单值特征集合; $C(a)$ 定义为广告 $a$ 的单值特征集合.

#### 4.2.2 层次聚类信息的融合

层次关系信息是指对象之间具有的层次结构关系信息. 例如广告商 – 广告活动 – 广告创意之间存在着天然的祖 – 父 – 子层次关系; 网站与网页之间也存在着天然的类似的层次关系, 如图 1 所示. 层次关系信息的特点是预测时实时可获得且有层次结构的类别特征. 对于没有天然层次关系的用户来说, 可以使用聚类技术对用户进行聚类, 聚类结果可以是层次聚类信息, 也可以是半结构化的类别特征信息. 对于不属于相同广告商的广告同样也可以采用聚类技术进行聚类, 获得类别特征以提升特征的粒度. 鉴于层次关系信息与聚类结果信息的相似性, 本文统称为层次聚类信息. 基于层次关系中父节点与子节点中对象应该具有 (获得) 相似的反馈行为, 或者同属于一个类别的对象也应该具有 (获得) 相似的反馈行为的思想, 缺乏历史反馈数据的节点对象就可以借鉴其父节点对象甚至祖父节点对象或者聚类结果中相同类别中其他对象的历史反馈数据或反馈行为规律, 来进行响应行为预测. 因此, 有效地融合有结构的层次聚类信息到张量分解模型, 利用父子之间或者同类对象之间的相似性, 对抗数据稀疏, 能够提升预测准确度和可靠性.

层次结构和聚类信息用于平滑预估值的技术在早期的基于概率模型的广告响应预测中应用非常广泛, 通常采用对预估值按照层次关系或聚类关系进行后处理的方法, 然而随着影响广告投放效果的因素越来越多, 越来越复杂, 这些方法不再能够满足日益增长的对于预估效率和准确率的要求, 这些方法也无法直接被融合到基于特征的模型训练中. 已有的在因子分解模型中集成层次关系信息的方法主要是基于层次关系中子节点对象与父节点对象应该具有相似的隐因子向量的思想, 在优化函数中加入具有父子关系的对象的隐因子向量的差的 L2 正则化项. 这种方法的缺点是只能考虑父子关系, 当某一对父子节点都没有历史投放数据可以参考时, 这种方法就无法发挥作用了. 实际上, 位于不同层次中的对象面临的数据稀疏性是不同的, 以广告对象之间的广告商 (祖) – 广告活动 (父) – 广告创意

(子) 构成的 3 层层次关系为例, 层次越低的对象, 投放历史越少, 数据稀疏现象越严重, 反之越靠近顶层的对象的历史投放数据越多. 因此, 本文将融合信息的层数进行扩充, 处于不同层的对象利用不同的层次关系来缓解稀疏. 对于数据稀疏性最强的最下层子节点, 利用祖 – 父 – 子 3 层关系来缓解稀疏, 具体方法是线性组合 3 层节点的隐因子向量来表示最下层子节点. 对于中间父节点, 采用祖 – 父两层关系来缓解稀疏即线性组合, 祖 – 父两层节点的隐因子向量来表示中间父节点. 具体实践中, 将父辈节点对象看作子节点对象的类别特征, 这样一来, 就可以采用与对象的类别特征一致的融合方法实现了, 值得注意的是, 由于表示每个节点的隐因子向量数量可能不同, 所以需要做归一化处理.

为了实现聚类信息的集成, 前提是完成类别的聚类, 例如将用户或者网页完成聚类, 得到每个对象的聚类类别 $c$. 然后实现聚类信息的集成策略相对就简单一些了. 只需要为每个聚类类别 $c$ 学习一个隐因子向量 $u_c$. 为了将对象与对象所属的聚类类别隐因子向量进行线性组合, 同样采用 “特征的隐因子向量乘以特征值” 的形式即 $u_c\alpha_c$ 来组合聚类信息. 这样一来, 聚类信息的融合形式便与普通的对象特征的融合形式达成了一致, 也可以统一表示为式 (2) 中的第 2 个累加项, 只需注意这时聚类类别特征也被看作一种单值特征来处理了. 因此, 层次聚类信息的融合也是集成在对象表示中, 我们的融合策略实现了在形式上与普通对象特征中的类别特征集成的一致性, 降低了融合策略在实践中的复杂性.

#### 4.2.3 全局特征的融合

全局特征指实时竞价广告响应预测中涉及的与用户、广告和上下文都相关, 或者都不相关的特征. 例如, 发起竞价请求的时间和网络交易平台, 这两个特征并不能明显地被分为是归属于用户、广告和上下文中的哪个对象. 例如时间特征与用户行为有关, 用户行为可能在时间上有一定的规律, 有些用户可能在周末休息时上网多一些, 而有些用户上班时工作轻松上网购物会多一些, 而回家后需要照顾孩子和家庭反而没有时间上网购物. 时间特征与广告也有关, 很多广告的点击率/转化率显示出明显的时间特性, 例如美食类广告可能在临近周末或者饭口时间之前的点击率/转化率会比其他时间略高一些; 而旅游类广告则在法定假期之前的点击率/转化率会比其他时间略高一些. 究竟是哪个网络交易平台发起的展示竞价请求与用户、广告和上下文几乎没有什么关系, 但不同网络交易平台的资质、信誉、实力、业务范围等方面的差异会影响媒体提供的展示的质量和产品信誉, 间接影响点击率/转化率. 因此, 这些因素在预测中都要被考虑, 却又区别于与对象有明确归属关系的普通对象特征, 需要采用不同的融合方法. 本文将通过模型级的线性组合来融合全局特征, 设全局特征的集合为 $C(t)$, 那么基于线性回归的全局特征组合可以表示为

$$g(x) = \sum_{c \in C(t)} w_c x_c, \tag{3}$$

其中, $x_c$ 为对应特征 $c$ 的特征值, $w_c$ 为对应特征 $c$ 的权重, 需要训练学习得到. 式 (3) 将矩阵分解或张量分解等预测模型进行线性组合完成以上各类信息的最终融合, 具体形式见 4.3 小节式 (16) 和 4.4 小节式 (20).

#### 4.2.4 点击反馈信息和转化反馈信息的融合

只利用点击信息或者只利用转化信息进行预测时 [4, 6∼8, 20, 23, 39, 40], 至少面临两个困难, 点击噪声和转化稀疏. 点击噪声主要是由点击欺诈和意外点击构成. 无论是点击欺诈还是意外点击, 都不是在用户对广告的真实兴趣驱动下产生的, 因此, 这些点击噪声给模型自动挖掘用户对广告的真实兴趣造

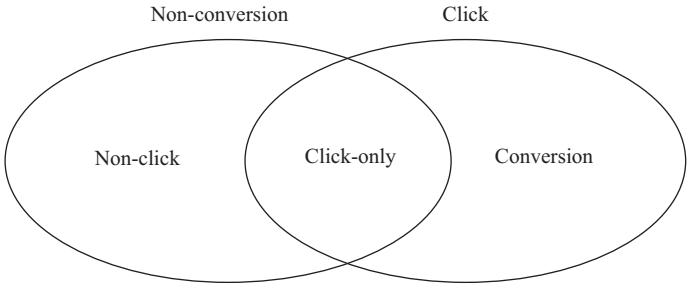


**图 3 广告响应分类法**

**Figure 3** The taxonomy of ad responses

成了严重的干扰, 降低了模型预测的准确性和可靠性. 幸运的是, 历史反馈数据中还有转化信息可用, 而且由于转化行为操作成本比点击行为操作成本大得多, 因此, 转化信息几乎没有噪声和意外. 更重要的是, 转化与点击之间具有强烈的正相关关系, 曝光的转化倾向性越高, 其点击倾向性也越高; 反之亦然. 换句话说, 产生转化的曝光应该比仅产生点击而未产生转化的曝光具有更高的被点击的可能性. 因此, 如果能有效地利用转化信息辅助点击预测, 对点击预测的准确性提升必然是有益的. 另一方面, 转化预测的最大困难就是转化数据的高稀疏性. 同样幸运的是, 历史反馈数据中的点击数据比转化数据丰富得多, 更重要的是, 与没有产生点击的曝光相比, 产生点击的曝光显然被转化的可能性更大, 因此, 如果能有效地利用点击信息辅助转化预测, 对转化预测性能的提升也会是有益的. 然而, 点击反馈信息和转化反馈信息与其他 3 类特征不同, 属于预测时不可获知的非实时特征, 因此融合方法也是大相径庭, 相对复杂.

在线广告曝光之后, 如果用户没有对广告曝光做出任何后续操作就像页面上没有广告一样, 称这种响应为 “未点击 (non-click)”; 如果用户点击了广告打开了广告落地页, 但是后续没有任何其他操作, 称这种响应为 “仅点击 (click-only)”; 如果用户在广告落地页面上进行了预定义的操作, 例如注册用户、下载软件、购买商品等, 称这种响应为 “转化 (conversion)”. 根据用户对广告曝光的响应, 所有的广告曝光可以分为互斥的 3 类, 即未点击、仅点击和转化. 另一方面, 从探讨广告点击问题的角度, 所有的广告曝光又可以分为互斥的两类, 即未点击和点击 (click), “仅点击” 和 “转化” 反馈都属于 “点击” 事件; 从探讨广告转化问题的角度, 所有的广告曝光又可以分为另外互斥的两类即未转化 (non-conversion) 和转化, “仅点击” 和 “未点击” 反馈都属于 “未转化” 事件; 不同响应的分类方法和类别之间的关系如图 3 所示.

为了能够在广告点击预测或转化预测中同时利用历史点击和历史转化信息之间的互补性, 取长补短, 互相促进和提升, 本文将转化、仅点击和未点击展示的正确排序作为模型训练的优化目标, 将广告响应预测问题看作三分类排序问题来求解, 提出了基于三元组排序优化的学习策略来实现张量分解模型的训练. 具体地说, 把一定时间段内的广告流量看作一个曝光列表, 给定固定的预算, 为了优化投资回报率, 我们的目标是将产生转化的曝光排在最前面, 将仅产生点击的曝光排在中间, 而将不会产生点击的曝光排在最后面. 于是, 无论是进行点击反馈预测还是转化反馈预测, 曝光在排序中的位置越高, 预测值也就越高.

现存的大部分广告响应预测方法 [4, 7, 20, 23, 39] 都是将广告响应预测问题看作回归问题来求解. 由于历史反馈数据中存在类别极度不平衡问题, 通常点击反馈只占桌面广告反馈数据的 1% 左右, 而转化反馈就更少了. 因此在这样的训练集上训练得到的分类和回归模型倾向于将样本预测为数量较多的负例类别或者趋向于预测 0, 从而严重影响预测质量. 本文针对广告响应预测问题的真实特点, 将其看

作未点击、仅点击和转化的三分类排序问题, 三分类排序问题以样本的正确排序为优化目标, 不再是逐点学习, 能够有效地避免类别不平衡对模型训练造成的影响, 从而实现了点击信息和转化信息的同时融合. 具体来说, 根据不同的曝光响应类型, 将历史曝光集合 $D$ 分为 3 个互斥的子集 $N^{++}$, $N^{+}$ 和 $N^{-}$. $N^{++}$ 表示所有产生 "转化" 的曝光集合; $N^{+}$ 表示所有产生 "仅点击" 的曝光集合; $N^{-}$ 表示所有 "未点击" 的曝光集合. 然后将广告响应预测看作三分类排序问题, 对于随机选择的曝光 $x_i \in N^{++}$, $x_j \in N^{+}$, $x_k \in N^{-}$, 学习参数为 $\Theta$ 的预测函数 $Y$ 使其满足这样的条件, $y(x_i) > y(x_j)$ 且有 $y(x_j) > y(x_k)$. 满足此条件的最佳排序是所有曝光按照其函数值从大到小排序时, 所有 $N^{++}$ 集合中的曝光都能排在最前面, 而所有 $N^{-}$ 集合中的曝光都能排在最后面. 否则都不是最佳排序. 为了求得满足条件的函数参数 $\Theta$, 需要优化随机选择的曝光三元组 $(x_i, x_j, x_k)$ 的排序, 因此, 需要构建新的曝光集合作为训练集 $D_t \subset D \times D \times D$, 具体形式为

$$D_t = \left\{(x_i, x_j, x_k) \,|\, x_i \in N^{++},\ x_j \in N^{+}, x_k \in N^{-}\right\}. \tag{4}$$

为了方便, 简记为

$$y_{x_i} = y(x_i), \quad \Delta y_{ij} = y_{x_i} - y_{x_j}. \tag{5}$$

于是, 本文定义实现点击和转化信息同时融合时采用的损失函数见式 (6) 所示. 其他情况下, 都采用最小二乘法来定义损失函数, 具体定义见式 (23) 所示.

$$L(\Theta) = \sum_{(x_i, x_j, x_k) \in D_t} -\ln \sigma\left(\Delta y_{ij}\right) - \ln \sigma\left(\Delta y_{jk}\right) + \lambda_{\Theta} \|\Theta\|^2, \tag{6}$$

式中, $\sigma(x)$ 定义为 S 形 Logistic 函数:

$$\sigma(x) := \frac{1}{1 + \mathrm{e}^{-x}}. \tag{7}$$

因为目标损失函数可导, 可以采用随机梯度下降算法来求解. 由于 $D_t$ 集合中的三元组数量大得惊人, 因此, 采用 Rendle 等 [41] 的建议, 使用有放回重复采样的随机梯度下降算法求解. 据 Rendle 研究表明, 这种有放回重复采样的随机梯度下降算法不仅比训练集全遍历的算法收敛速度快, 而且性能更稳定. 函数的梯度如下:

$$\begin{aligned} \frac{\partial L}{\partial \Theta} &= \sum_{(x_i, x_j, x_k) \in D_t} -\frac{\partial}{\partial \Theta} \ln \sigma\left(\Delta y_{ij}\right) - \frac{\partial}{\partial \Theta} \ln \sigma\left(\Delta y_{jk}\right) + \lambda_{\Theta} \frac{\partial}{\partial \Theta} \|\Theta\|^2 \\ &\propto \sum_{(x_i, x_j, x_k) \in D_t} \frac{-\exp(-\Delta y_{ij})}{1 + \exp(-\Delta y_{ij})} \cdot \frac{\partial\left(\Delta y_{ij}\right)}{\partial \Theta} + \frac{-\exp(-\Delta y_{jk})}{1 + \exp(-\Delta y_{jk})} \cdot \frac{\partial(\Delta y_{jk})}{\partial \Theta} + \lambda_{\Theta} \Theta. \end{aligned} \tag{8}$$

因此参数 $\Theta$ 更新规则如下:

$$\Theta \leftarrow \Theta - \eta \cdot \left(\frac{-\exp\left(-\Delta y_{ij}\right)}{1 + \exp\left(-\Delta y_{ij}\right)} \cdot \frac{\partial\left(\Delta y_{ij}\right)}{\partial \Theta}\right) + \eta \cdot \left(\frac{-\exp\left(-\Delta y_{jk}\right)}{1 + \exp\left(-\Delta y_{jk}\right)} \cdot \frac{\partial\left(\Delta y_{jk}\right)}{\partial \Theta}\right) + \eta \cdot \lambda_{\Theta} \Theta. \tag{9}$$

在工程实现中, 为了节省存储空间, 随机有放回重复采样的过程并没有事先完成构造并存储三元组集合 $D_t$, 而只是分别存储集合 $N^{++}$, $N^{+}$ 和 $N^{-}$, 在训练过程实时随机有放回地采样 3 个样本即 $x_i \in N^{++}$, $x_j \in N^{+}$ 和 $x_k \in N^{-}$, 这样就得到了一个三元组 $(x_i, x_j, x_k) \in D_t$. 具体算法如算法 1 所示, 有关此方法的更详细阐述见文献 [22].

**Algorithm 1** Triplet-wise learning algorithm

Input: Weighting coefficient $\alpha$, learning rate $\eta$, regularization coefficient $\lambda_\Theta$, training dataset $D$.

Output: Learned parameter $\Theta$.

Initialize parameter $\Theta$;
Construct $N^{++}$, $N^{+}$ and $N^{-}$ according to $D$;
**repeat**
  Draw uniformly $x_i \in N^{++}$;
  Draw uniformly $x_j \in N^{+}$;
  Draw uniformly $x_k \in N^{-}$;
  // Then we have a tuple $(x_i, x_j, x_k) \in D_t$;
  Calculate $\Delta y_{ij}$ and $\Delta y_{jk}$;
  **for** each $\theta \in \Theta$ **do**
    Calculate $\frac{\partial(\Delta y_{ij})}{\partial\theta}$ and $\frac{\partial(\Delta y_{jk})}{\partial\theta}$;
    Calculate gradients $\frac{\partial L}{\partial\theta}$;
    Update $\theta$;
  **end for**
**until** convergence;
**return** $\Theta$.

#### 4.2.5 异构信息综合融合方案

面向张量分解模型的广告响应预测异构信息综合融合方案如图 4 所示. 当预测样本 $i$ 到达时, 预测系统首先进行特征提取, 分为对象特征和全局特征. 全局特征送入线性回归模型预测出预估值 $y_i'$. 而对象特征根据所属对象继续分为普通特征和层次聚类特征, 根据两种特征的处理方法, 分别进行特征隐因子向量的组合, 然后根据各对象的特征组合表示 $f_u$, $f_p$, $f_a$, 使用张量分解模型进行广告响应预测得到预估值 $y_i$, 最后再将 $y_i'$ 和 $y_i$ 进行一次线性组合, 得到最终的预估值 $r_i$. 其中, 图中第 6 步的张量分解模型是使用融合了点击和转化信息的基于三元组排序优化的学习策略学习的参数. 本文将集成了异构信息的张量分解模型称为 “基于异构信息融合的张量分解模型”, 下文中将针对具体的张量分解模型给出具体的融合公式.

### 4.3 基于异构信息融合的 CP 分解模型

三阶 CP 分解将三阶张量 $\mathcal{R} \in \mathbb{R}^{I\times J\times K}$ 分解为一组秩为 1 的张量之和. 将 CP 分解用于计算广告响应的估值, 分解过程还可以表示为 $n$ 模积的形式, 如图 5 所示. 根据 CP 分解, 三阶张量 $\mathcal{R}$ 可以被分解为

$$\mathcal{R} \approx \sum_{f=1}^{m} u_f \circ p_f \circ a_f, \tag{10}$$

其中, $m$ 是一个正整数, 表示分解后秩为 1 的张量的数量; $u_f \in \mathbb{R}^I$, $p_f \in \mathbb{R}^J$, $a_f \in \mathbb{R}^K$.

相应的, CP 分解的估值公式为

$$\hat{r}_{i,j,k} = \sum_{f=1}^{m} u_{f,i} p_{f,j} a_{f,k}, \tag{11}$$

其中, $i$, $j$, $k$ 分别表示三阶张量中的元素 $r$ 所在 3 个维度的下标; $m$ 是一个正整数, 表示分解后秩为 1 的张量的数量.

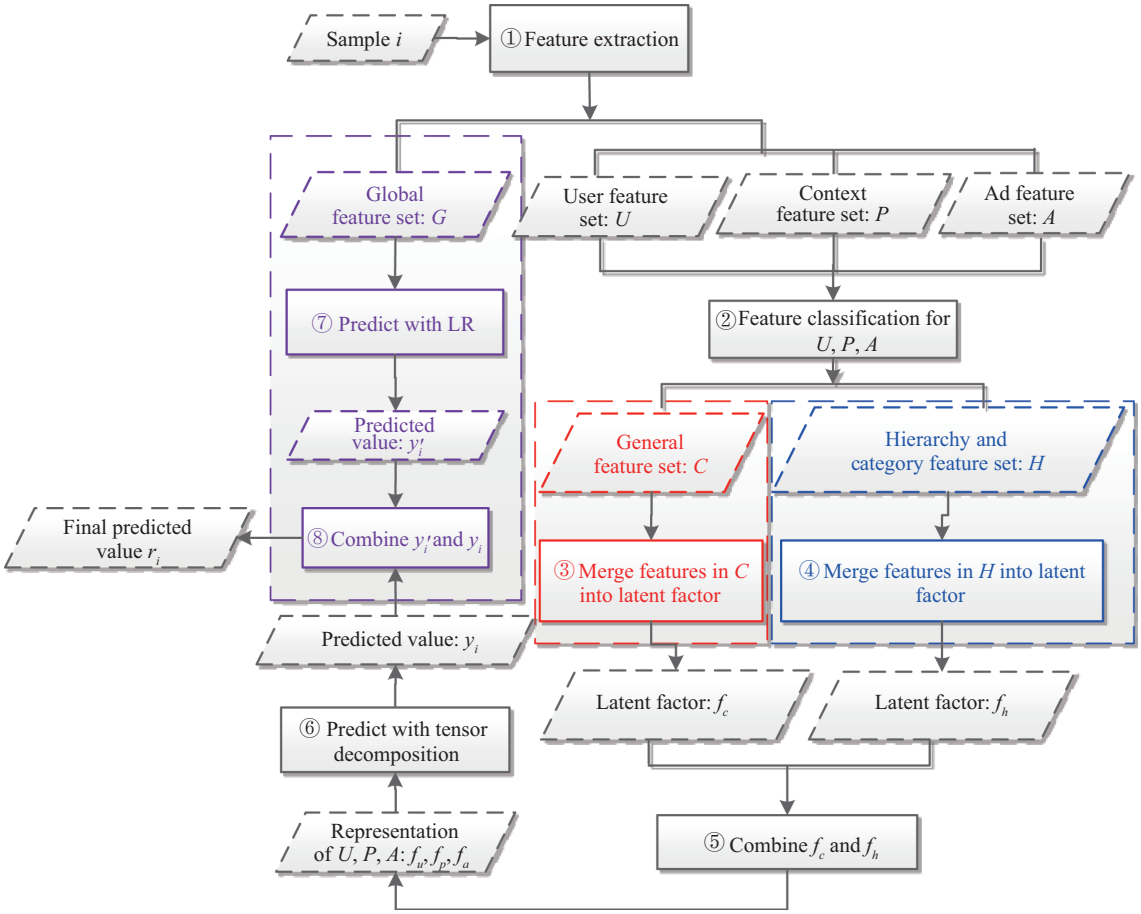


**图 4　(网络版彩图) 异构信息融合方案示意图**

**Figure 4**　(Color online) The integrating framework for heterogeneous information

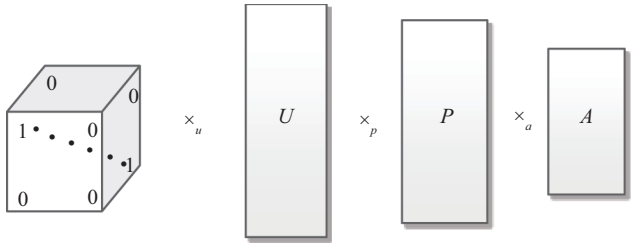


**图 5　广告响应预测 CP 分解图**

**Figure 5**　CP decomposition for ad CTR prediction

当使用标准的 CP 分解来进行广告响应预估时, 按照上文中给出的集成对象特征和层次聚类信息的 CP 分解模型的估值公式如下:

$$\hat{r}_{u,p,a} = \sum_{f=1}^{m} \left( \sum_{c \in C(u)} u_{c,f} \alpha_c \cdot \sum_{c \in C(p)} p_{c,f} \beta_c \cdot \sum_{c \in C(a)} a_{c,f} \gamma_c \right), \tag{12}$$

其中, $C(u)$, $C(p)$, $C(a)$ 分别对应用户、上下文、广告的对象特征或者层次聚类特征集合; $\alpha_c$, $\beta_c$, $\gamma_c$ 分别对应用户、广告、上下文的对象特征值, 如果是类别特征, 其值为 Boole 值, 数值特征是实数;

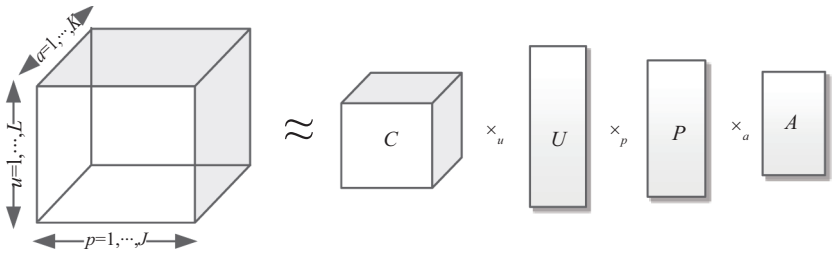


**图 6 广告响应预测问题的 Tucker 分解图**

**Figure 6** Tucker decomposition for ad CTR prediction

$u_c \in \mathbb{R}^K$, $p_c \in \mathbb{R}^K$, $a_c \in \mathbb{R}^K$ 分别对应各个特征的隐因子向量; $K \in \mathbb{N}$ 对应隐因子向量维度.

虽然因子分解模型能够灵活地处理各种各样的数据, 然而, 除了对象之间的交互关系, 对象本身对点击事件也会有影响. 例如, 有的用户明显比其他用户活跃, 点击广告的平均次数比其他用户多很多, 还有的用户几乎从来不点击广告. 再比如, 有些广告本身的创意新颖, 或者是当前的流行或热门主题, 那么这个广告会比其他广告获得更多的点击. 因此, 加入偏置变量来刻画这些影响. 加入偏置后, 估值公式如下:

$$\hat{r}_{u,p,a} = \sum_{f=1}^{m} \left( \sum_{c \in C(u)} u_{c,f}\alpha_c \cdot \sum_{c \in C(p)} p_{c,f}\beta_c \cdot \sum_{c \in C(a)} a_{c,f}\gamma_c \right) + b + \mu, \tag{13}$$

其中, $\mu \in \mathbb{R}$ 为统计均值; 偏置 $b$ 定义如下:

$$b = \sum_{c \in C(u)} b_c^{(u)}\alpha_c + \sum_{c \in C(p)} b_c{}^{(p)}\beta_c + \sum_{c \in C(a)} b_c{}^{(a)}\gamma_c. \tag{14}$$

再集成全局特征的基于异构信息融合的 CP 分解模型的估值公式如下:

$$\hat{r}_{u,p,a} = \varphi \left( \sum_{f=1}^{K} \left( \sum_{c \in C(u)} u_{c,f}\alpha_c \cdot \sum_{c \in C(p)} p_{c,f}\beta_c \cdot \sum_{c \in C(a)} a_{c,f}\gamma_c \right) + b \right) + (1-\varphi) \left( \sum_{c \in C(t)} w_c g_c \right) + \mu, \tag{15}$$

其中, $C(t)$ 为全局特征集合, $g_c$ 全局特征值, $w_c$ 表示对应 $g_c$ 的权重. 设 $C = C(u) \cup C(p) \cup C(a)$, 则基于异构信息融合的 CP 张量分解模型的预测时间复杂度是 $O(K \cdot |C|)$.

为了简化训练过程, 减少参数, 假设 $\varphi = 0.5$, 相当于两部分的系数都是 1, 因此最终的基于异构信息融合的 CP 分解模型的估值公式如下:

$$\hat{r}_{u,p,a} = \sum_{f=1}^{K} \left( \sum_{c \in C(u)} u_{c,f}\alpha_c \cdot \sum_{c \in C(p)} p_{c,f}\beta_c \cdot \sum_{c \in C(a)} a_{c,f}\gamma_c \right) + \left( \sum_{c \in C(t)} w_c g_c \right) + b + \mu. \tag{16}$$

### 4.4 基于异构信息融合的 Tucker 分解模型

Tucker 分解将三阶张量 $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ 分解为一个核心张量与 3 个矩阵的模积, 广告响应预测问题的 Tucker 分解过程如图 6 所示.

$$\mathcal{X} \approx \mathcal{G} \times_1 A \times_2 B \times_3 C = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_p \circ b_q \circ c_r, \tag{17}$$

其中, $A \in \mathbb{R}^{I\times P}$, $B \in \mathbb{R}^{J\times Q}$, $C \in \mathbb{R}^{K\times R}$ 是分解后的隐因子矩阵, $\mathcal{G} \in \mathbb{R}^{P\times Q\times R}$ 是分解后的核心张量. 根据 Tucker 分解, 三阶张量 $\mathcal{R} \in \mathbb{R}^{I\times J\times K}$ 可以被分解为

$$\mathcal{R} \approx \mathcal{C}\times_u U\times_p P\times_a A = \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N} c_{lmn} u_l \circ p_m \circ a_n, \tag{18}$$

其中, $U \in \mathbb{R}^{I\times L}$, $P \in \mathbb{R}^{J\times M}$, $A \in \mathbb{R}^{K\times N}$ 是分解后的隐因子矩阵, $\mathcal{C} \in \mathbb{R}^{L\times M\times N}$ 是分解后的核心张量. 相应的, Tucker 分解的估值公式为

$$\hat{r}_{i,j,k} \approx \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N} c_{l,m,n} u_{i,l} p_{j,m} a_{k,n}. \tag{19}$$

同样, 为了缓解数据稀疏带来的影响同时解决冷启动问题, 给出基于异构信息融合的 Tucker 分解模型如公下:

$$\begin{aligned}\hat{r}_{u,p,a} = &\sum_{l=1}^{L}\left(\sum_{c\in C(u)} u_{c,l}\alpha_c\right)\sum_{m=1}^{M}\left(\sum_{c\in C(p)} p_{c,m}\beta_c\right)\sum_{n=1}^{N}\left(\sum_{c\in C(a)} a_{c,n}\gamma_c\right)\cdot y_{i,j,k}\\ &+\left(\sum_{c\in C(t)} w_c g_c\right) + b + \mu,\end{aligned} \tag{20}$$

其中, $y_{i,j,k}$ 为核心张量对应坐标 $(i,j,k)$ 位置上的元素值; $C(u)$, $C(p)$, $C(a)$ 分别对应用户、上下文、广告的对象特征或层次聚类特征集合, $C(t)$ 为全局特征集合; $\alpha_c$, $\beta_c$, $\gamma_c$, $g_c$ 分别应对用户、广告、上下文特征及全局特征值, 如果是类别特征, 其值为 Boole 值, 数值特征是实数; $u_c \in \mathbb{R}^L$, $p_c \in \mathbb{R}^M$, $a_c \in \mathbb{R}^N$ 分别对应特征的隐因子向量; $\mu \in \mathbb{R}$ 为统计均值; $L, M, N \in \mathbb{N}$ 分别对应用户、广告和上下文的隐因子向量维度, 通常取 $L = M = N$. 设 $K = \min\{L, M, N\}$, $C = C(u) \cup C(p) \cup C(a)$, 则基于异构信息融合的 Tucker 张量分解模型的预测时间复杂度是 $O(K^3 \cdot |C|)$.

## 5 实验与分析

### 5.1 实验数据集

本节的实验验证, 选用了 3 个实时竞价广告的数据集, 这 3 个数据集都是知名需求方平台公司品友互动 2014 年公开发布的数据集, 来自 2013 年举办的全球实时竞价算法大赛的三季数据集 [42]. 每一季的数据都有来自若干个广告商, 连续一段时间内的广告展示、点击和转化日志. 每一季的数据分别使用不同的方法被分成了两个部分, 分别用于训练和离线测试. 数据集的收集日期、展示数量、点击数量, 以及点击率统计如表 1 所示. 第 1 和 2 季数据集均包含了连续 10 天的广告投放日志, 前 7 天数据是训练集, 接下来的 3 天数据是测试集, 为了训练学习率和正则化系数等超参数, 实验中将训练集划分为两部分, 前 5 天的数据做训练集, 后 2 天的数据做验证集, 测试集不变; 第 3 季数据集也包含连续 10 天的数据, 但是从第 2 天到第 9 天随机抽取出小部分数据加上第 10 天全天的数据做测试集, 其他数据做训练集, 并采用与前两季同样的方式抽取了验证集. 三季数据共包括来自不同产业领域的 12 个广告商发起的 25 个广告活动, 164 个广告创意, 如表 2 所示. 第 1 季数据中没有明确地给出广告商编号, 但根据品友互动的说明 [42] 登陆页编号可以用来区分不同的广告商.

**表 1 三季数据集采集日期和点击率统计表**

**Table 1** Charateristics for three-quarters datasets

| Quarter | Dataset | Date | Impression No. | Click No. | Click-through rate (%) |
|---|---|---|---|---|---|
| 1 | Training dataset | May 11 ∼ May 17 | 9262861 | 7482 | 0.076 |
| | Test dataset | May 18 ∼ May 20 | 2594386 | 8934 | 0.075 |
| 2 | Training dataset | June 6 ∼ June 12 | 12237229 | 8961 | 0.073 |
| | Test dataset | June 3 ∼ June 15 | 2524630 | 1873 | 0.074 |
| 3 | Training dataset | October 19 ∼ October 27 | 3158171 | 2709 | 0.086 |
| | Test dataset | October 21 ∼ October 28 | 1579086 | 1120 | 0.071 |

**表 2 三季数据集广告商类别及广告数量统计表**

**Table 2** The statistics of advertiser categories and ad numbers

| Advertiser | Quarter | Industry type | Campaign | Creative |
|---|---|---|---|---|
| df6f61b2409f4e2f16b6873a7eb50444 | 1 | Consumer packaged goods (CPG) | 1 | 14 |
| 3a7eb50444df6f61b2409f4e2f16b687 | 1 | Chinese vertical e-commerce | 1 | 12 |
| 9f4e2f16b6873a7eb504df6f61b24044 | 1 | Vertical online media | 1 | 7 |
| 1458 | 2 | Chinese vertical e-commerce | 1 | 8 |
| 3358 | 2 | Software | 3 | 25 |
| 3386 | 2 | International e-commerce | 1 | 19 |
| 3427 | 2 | Oil | 2 | 13 |
| 3476 | 2 | Tire | 11 | 11 |
| 2259 | 3 | Milk powder | 1 | 22 |
| 2261 | 3 | Telecom | 1 | 9 |
| 2821 | 3 | Footwear | 1 | 3 |
| 2997 | 3 | Mobile e-commerce app install | 1 | 23 |

训练集和测试集中对应广告商的历史广告展示数量、点击数量及 CTR 统计如表 3 和 4 所示. 从表中可以看出, 除了广告商 2997 的点击率为 0.444% 之外, 无论在单个数据集中, 还是广告商上, 点击率都小于 0.1%. 桌面显示广告的点击率通常都在 0.1% 左右, 远远低于搜索广告的平均点击率 1.2%, 更是远远低于推荐系统中的用户反馈率 (MovieLens 大约为 4.5%, Netflix 为 1.2%). 广告商 2997 属于移动电子商务应用领域, 据统计报告显示移动广告的平均点击率高于桌面显示广告, 通常在 1% 左右, 主要是因为移动设备触屏界面尺寸有限, 而手指比鼠标指针 “肥得多”, 所以常常发生误点击操作, 因而提升了移动广告的点击率.

主要特征的维度如表 5 所示. 通常, 每条记录至少包含 4 类特征, 用户特征 (用户 ID、用户代理、区域和城市等)、广告特征 (广告商 ID、广告活动 ID 和广告创意 ID 等)、上下文特征 (广告位 ID、广告位宽度、广告位高度、网址和域名等) 和竞价特征 (广告交易平台、竞价价格和付费价格等). 竞价相关特征经常被用于实时竞价优化策略的研究 [5, 12, 43, 44], 考虑到本文的工作不涉及出价价格的优化策略, 因此, 放弃了这部分特征.

这 3 个数据集的共同点是都来自于实时竞价广告的需求方平台, 基本数据结构都是一致的, 既包含类别特征也包含数值特征 (广告位宽度和广告位高度). 3 个数据集的不同点主要体现在以下 4 个方面, 第一, 第 1 季数据没有用户标签特征, 其他两季数据都有, 用户标签是多值特征. 第二, 如前所述,

**表 3　训练集广告商历史点击率统计表**

**Table 3**　Training dataset statistics

| Quarter | Advertiser | Impression No. | Click No. | Click-through rate (%) |
|---|---|---|---|---|
| 1 | 9f4e2f16b6873a7eb504df6f61b24044 | 3251782 | 3055 | 0.094 |
| 1 | 3a7eb50444df6f61b2409f4e2f16b687 | 3182633 | 2644 | 0.083 |
| 1 | df6f61b2409f4e2f16b6873a7eb50444 | 2828446 | 1303 | 0.046 |
| 2 | 1458 | 3083056 | 2454 | 0.080 |
| 2 | 3358 | 1742104 | 1358 | 0.078 |
| 2 | 3386 | 2847802 | 2076 | 0.073 |
| 2 | 3427 | 2593765 | 1926 | 0.074 |
| 2 | 3476 | 1970360 | 1027 | 0.052 |
| 3 | 2259 | 835556 | 280 | 0.034 |
| 3 | 2261 | 687617 | 207 | 0.030 |
| 3 | 2821 | 1322561 | 843 | 0.064 |
| 3 | 2997 | 312437 | 1386 | 0.444 |
| Total | 12 | 24658119 | 18559 | 0.075 |

**表 4　测试集广告商历史点击率统计表**

**Table 4**　Test dataset statistics

| Quarter | Advertiser | Impression No. | Click No. | Click-through rate (%) |
|---|---|---|---|---|
| 1 | 9f4e2f16b6873a7eb504df6f61b24044 | 896908 | 850 | 0.095 |
| 1 | 3a7eb50444df6f61b2409f4e2f16b687 | 918846 | 679 | 0.074 |
| 1 | df6f61b2409f4e2f16b6873a7eb50444 | 778632 | 403 | 0.052 |
| 2 | 1458 | 614638 | 543 | 0.088 |
| 2 | 3358 | 300928 | 339 | 0.113 |
| 2 | 3386 | 542421 | 496 | 0.091 |
| 2 | 3427 | 536795 | 395 | 0.074 |
| 2 | 3476 | 523848 | 302 | 0.058 |
| 3 | 2259 | 417179 | 131 | 0.031 |
| 3 | 2261 | 343862 | 97 | 0.028 |
| 3 | 2821 | 661964 | 394 | 0.060 |
| 3 | 2997 | 153063 | 533 | 0.348 |
| Total | 12 | 6689084 | 5162 | 0.077 |

第 3 季的测试集数据采集方式与其他两季不同. 第三, 第 3 季数据还有一个不同就是训练数据的数量明显少于前 2 季数据, 后面通过实验发现由于训练数据不够充分使得第 3 季训练模型的泛化能力不及前 2 季数据集. 第四, 不同季的广告主工业类型不同, 广告活动的数量也不同. 第 2 季广告主、广告活动和广告创意的层次关系更丰富、更完整.

表 5 数据集主要特征维度统计表

**Table 5** Dimensonality of main features for three-quarters datasets

| Quarter | Dataset | Impression No. | User No. | Tag No. | Slot No. | Page No. | Advertiser No. | Campaign No. | Creative No. |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Training dataset | 9262861 | 6799908 | null | 124684 | 2082249 | 3 | 3 | 32 |
| | Test dataset | 2594386 | 2164525 | | 58945 | 811585 | 3 | 3 | 33 |
| 2 | Training dataset | 12237229 | 10146491 | 45 | 141515 | 2362123 | 5 | 18 | 74 |
| | Test dataset | 2524630 | 2310303 | 68 | 48458 | 663218 | 5 | 18 | 74 |
| 3 | Training dataset | 3158171 | 2818424 | 69 | 53518 | 963576 | 4 | 4 | 57 |
| | Test dataset | 1579086 | 1490321 | 58 | 43603 | 552694 | 4 | 4 | 54 |

### 5.2 实验设置与评价指标

我们使用基于异构信息融合的矩阵分解模型 [16,38] 以及逻辑回归模型作为基准方法来评价基于异构信息融合的张量分解模型的性能. 基于异构信息融合的矩阵分解模型的估值公式如下:

$$\hat{r}_{i,j,k} = \mu + \left(\sum_k b_k^{(g)}\gamma_k + \sum_i b_i^{(u)}\alpha_i + \sum_j b_j^{(i)}\beta_j\right) + \left(\sum_i p_i\alpha_i\right)^{\mathrm{T}}\left(\sum_j q_j\beta_j\right), \tag{21}$$

其中, 在广告响应预测问题中, $i$ 和 $j$ 分别对应用户和广告的特征, 而 $k$ 代表其他全局特征, 由于矩阵分解只能建模用户和广告之间的二维交互, 所以只考虑了全局特征的偏置影响. $\alpha_i$, $\beta_j$, $\gamma_k$ 分别为用户 $i$、广告 $j$ 和全局 $k$ 的特征表示; $p_i$, $q_j$ 分别为用户 $i$ 和广告 $j$ 的隐因子向量; 式中的 $b$ 为各个对象对应的偏置变量.

本文使用 ROC 曲线下面积 (area under the ROC curve, AUC) [45~47] 和均方根误差 (root mean squared error, RMSE) 作为评价指标. 由于 AUC 非常适合类别不均衡测试集的性能评价, 被广泛用于各种评测大赛和工业实践中以评价算法的广告响应预测性能或排序性能 [2,5,8,12,24,39], 2012 年 KDD 广告点击率预测大赛就采用 AUC 作为评价指标. 本文采用与 Fawcett [47] 提出的算法 3 相似的算法来计算 AUC. 而 RMSE 是基于回归问题的常用评价指标, 设 $D$ 为训练样本集, 计算公式如下:

$$\mathrm{RMSE} = \sqrt{\frac{1}{|D|}\sum_{(u,p,a)\in D}\left(r_{u,p,a} - \hat{r}_{u,p,a}\right)^2}. \tag{22}$$

本文实验根据不同情况采用了两种损失函数, 一种情况是当同时融合点击和转化信息时, 采用基于三元组排序化的损失函数, 具体定义见式 (6). 另一种情况是只考虑点击信息, 而不考虑转化信息即只融合其他类别信息的情况, 本文采用在观测数据集 $D$ 上直接建模的方法求真实反馈张量 $\mathcal{R}$ 的估计 $\hat{\mathcal{R}}$, 使其与 $\mathcal{R}$ 在已知数据集 $D$ 上具有最小的平方逼近误差. 为了避免过拟合, 采用附加的正则化项后的损失函数定义为

$$L(\Theta) = \frac{1}{2}\sum_{(u,p,a)\in D}\left[\left(r_{u,p,a} - \hat{r}_{u,p,a}\right)^2 + \lambda\|\Theta\|^2\right], \tag{23}$$

其中, $\Theta$ 是需要学习的参数, $\lambda$ 是正则化系数. 由于此函数可微, 采用随机梯度下降算法训练模型, 求解其最小值.

模型的所有隐因子向量的长度均设置为 32. 模型其他超参数的初始值均由 $N(0, 0.1^2)$ 随机抽样生成. 选择在验证集上性能最优的迭代次数、学习率和正则化系数. 具体方法是固定其他超参数, 手

**表 6　模型表示与信息融合关系表**

**Table 6**　Interpretation for model notation

| Notation | Object feature | Hierarchy and category feature | Global feature | Click and conversion feature |
|---|---|---|---|---|
| X_0 | | | | |
| X_1 | √ | | | |
| X_2 | √ | √ | | |
| Featured-based X | √ | √ | √ | |
| X_IHI | √ | √ | √ | √ |

工改变需要设置的参数, 直至在验证集上取得最优性能. 所有因子分解模型的超参数基本是相同的, 学习率为 0.0001, 正则化系数为 0.001; 而逻辑回归模型的学习率和正则化系数都为 0.001. 数值特征采用特征值除以特征最大值的方法进行归一化处理.

### 5.3　不同级别异构信息融合方法的性能对比

为了对比融合不同级别信息的不同模型的性能, 本文采用融合不同级别信息的 MF, Tucker 和 CP 模型进行实验对比. 模型表示符号与信息融合方法之间的对应关系如表 6 所示.

表中的符号 X, 代表模型名称, 下文中会用具体的模型名称 MF, CP 或 Tucker 代替. 后面的数字或字符串分别代表不同的特征融合方法. 例如 “X_0” 所在行的右侧特征列没有任何的勾选, 表示只采用对象标识未融合其他任何信息的 X 模型; “X_IHI” 所在行的右侧所有特征列均勾选, 表示融合了所有特征即对象特征、层次聚类特征、全局特征及点击与转化特征. 再如模型符号 “CP_2” 表示融合了对象特征和层次聚类特征的 CP 子模型. 本文针对 MF, Tucker 和 CP 模型采用不同的特征融合方法, 共实验了 15 种方案的对比.

首先, 比较了相同模型情况下, 不同信息融合方法的性能表现. 图 7∼9 分别给出矩阵分解模型、Tucker 分解模型和 CP 分解模型在 3 个数据集上采用不同级别特征融合时的性能表现. 依图可见, 当采用对象特征组合表示对象时, 各个模型 (X_1 与 X_0 比较) 的性能提升都是很大的, 平均提升达到 4%∼5%. 而在此基础上再增加层次聚类信息时 (X_2 与 X_1 比较), 预测性能的提升幅度就非常有限了, 最高提升刚刚达到 1%, 而最小提升在 0.1% 左右. 通过不同数据集上的对比发现, 由于只有第 2 季数据集具有完整的广告主 – 广告活动 – 广告创意 3 层结构, 而第 1 和 3 季数据集都只有两层结构即广告主 – 广告创意, 因此, 各个模型融合层次聚类信息后, 在第 2 季数据集上的性能提升平均值大于在第 1 和 3 季数据集上的结果. 融合 4 类特征的 MF, CP 和 Tucker 3 种模型性能平均提升 5% 以上. 其中, MF 模型的平均提升最大达到了 7.83%.

图 7∼9 显示, 从 AUC 性能指标上看, 同时考虑了点击和转化信息的三元组排序优化学习比只考虑点击信息的独点学习策略性能要好. 同时融合了历史点击和历史转化信息的三元组排序优化学习将所有的点击样本分为了两类即转化和仅点击, 但是, 它的优化目标仍然是将这两类标本排在未点击样本的前面. 由于它通过排序优化策略组合了转化信息, 所以在转化信息丰富而且点击噪声较多的第 3 季数据集上, 转化信息的利用能够缓解点击噪声带来的影响, 使得三元组排序优化学习在第 3 季数据集上解决点击率预测的 AUC 性能提升更多. 然而, 在第 1 或 2 季数据集上, 由于没有第 3 季数据集中那么多的点击噪声或者转化信息也没有第 3 季中的丰富, 因此三元组排序优化学习的表现没有第 3 季数据集上出色, 但也有显著的性能提升.

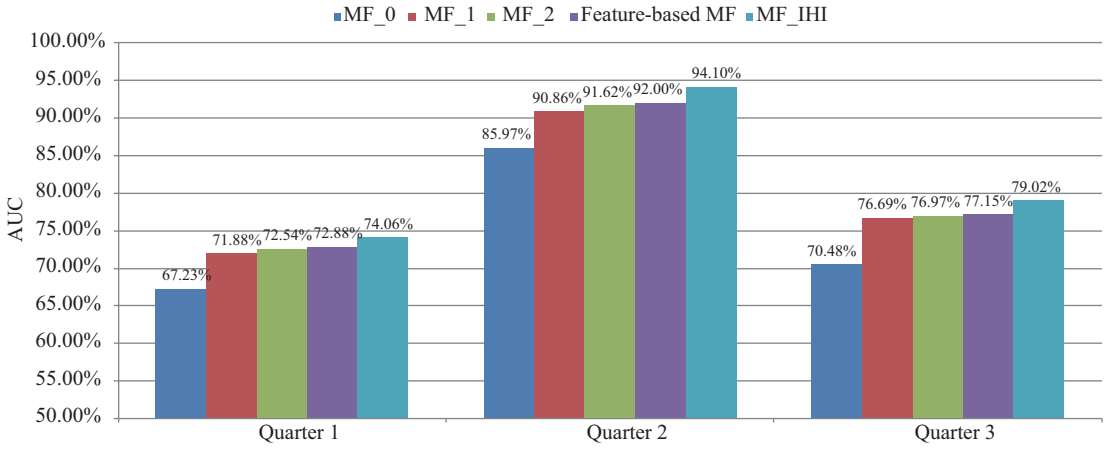


**图 7** (网络版彩图) 矩阵分解模型融合不同级别特征的性能

**Figure 7** (Color online) Performance of matrix factorization integrated different level features

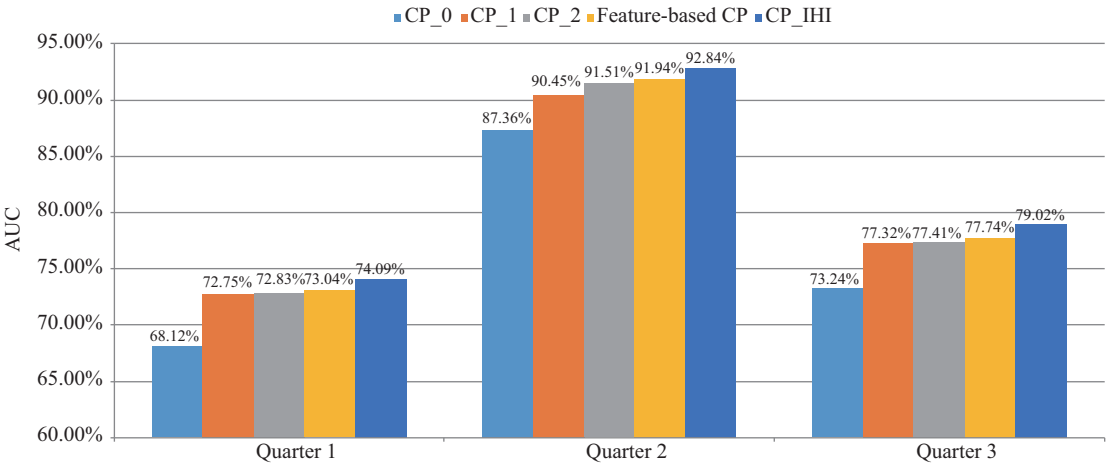


**图 8** (网络版彩图) CP 分解模型融合不同级别特征的性能对比

**Figure 8** (Color online) Performance of CP integrated different level features

其次, 将 3 类模型中性能表现最好的 3 个基于异构信息融合的协同过滤模型与 LR 模型的性能进行了对比, 在三季数据集上的预测性能对比如图 10~12 所示.

由于协同过滤因子模型能够有效地学习不同对象之间的非线性交互关系, 因此, 他们的预测性能都优于传统的逻辑回归模型, 在 3 个数据集性能上都有提升. 在第 3 季数据集上提升最多, 第 3 季数据集规模比前两季规模小很多, 说明因子分解模型在避免过拟合方面的能力优于逻辑回归. 在各种因子分解模型中, 由于张量分解模型能够自动学习用户、广告和上下文之间的三维交互关系, 因此 Tucker 和 CP 分解模型的预测性能总体优于 MF 模型. Tucker 与 CP 分解预测能力不分伯仲, 在不同数据集上表现不同, 总体来看 Tucker 分解模型由于具有较强的三维交互关系表示能力, 而 CP 分解为了获得较优的时间效率而牺牲了交互关系表示能力, 因此在不同性质的数据集上 Tucker 分解模型表现出比 CP 分解模型较好的性能稳定性.

最后, 给出一种基于回归性能的评价指标值 RMSE, 如表 7 给出了基于最小二乘法学习策略的融合前 3 类特征的各个模型的 RMSE 性能表现. 由于训练集中非常严重的类别不平衡, 点击正例平均占总体的 0.0801%, 基于回归优化的模型总是更倾向于将 CTR 预估为接近 0 的值, 因此, 预测结果的

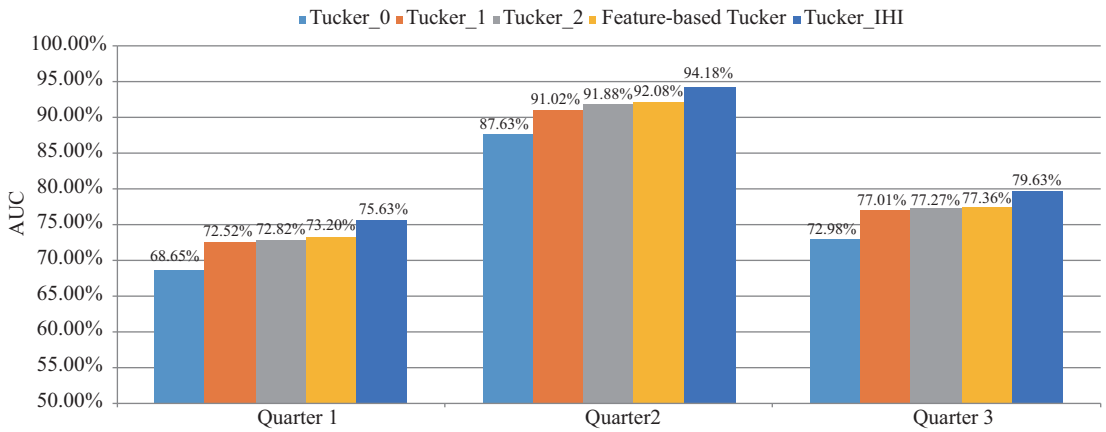


**图 9** (网络版彩图) **Tucker** 分解模型融合不同级别特征的性能对比

**Figure 9** (Color online) The performance of Tucker factorization integrated different level features

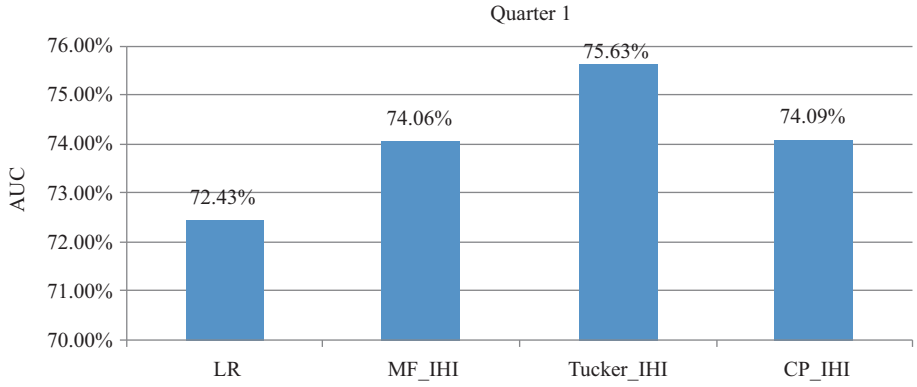


**图 10** (网络版彩图) 基于异构信息融合的不同模型在第 **1** 季数据集上性能对比

**Figure 10** (Color online) Performance of all models integrating heterogeneous information in quarter 1

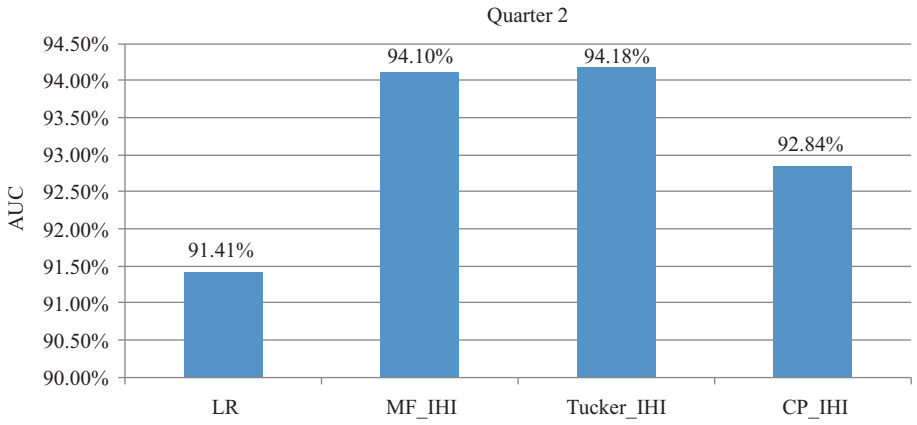


**图 11** (网络版彩图) 基于异构信息融合的不同模型在第 **2** 季数据集上性能对比

**Figure 11** (Color online) Performance of all models integrating heterogeneous information in quarter 2

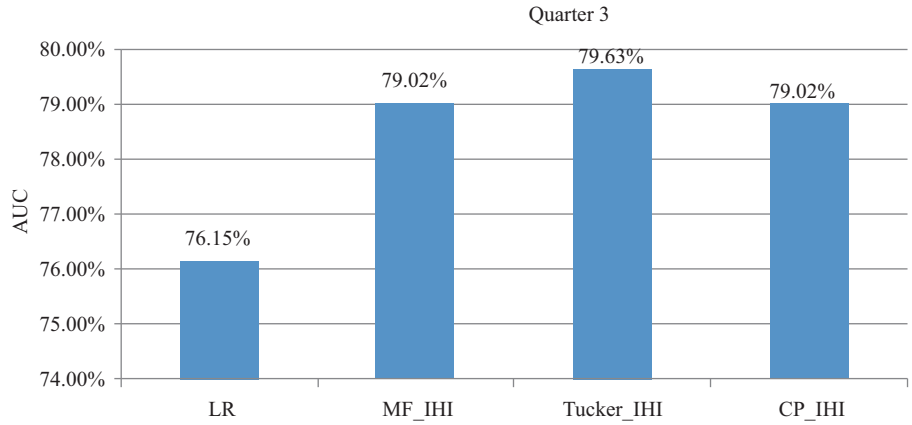


**图 12 (网络版彩图) 基于异构信息融合的不同模型在第 3 季数据集上性能对比**

**Figure 12** (Color online) Performance of all models integrating heterogeneous information in quarter 3

**表 7 融合 3 类特征模型的 RMSE 值对比表**

**Table 7** RMSE values for all models merging the first three types of features

| RMSE | LR | Feature-based MF | Feature-based tucker | Feature-based CP |
|---|---|---|---|---|
| Quarter 1 | 0.0274 | 0.0261 | 0.0235 | 0.0275 |
| Quarter 2 | 0.0262 | 0.0261 | 0.0260 | 0.0262 |
| Quarter 3 | 0.0268 | 0.0267 | 0.0267 | 0.0266 |

**表 8 融合 4 类特征模型的 RMSE 值对比表**

**Table 8** RMSE values for all models merging all features

| RMSE | LR | MF_IHI | Tucker_IHI | CP_IHI |
|---|---|---|---|---|
| Quarter 1 | 0.0274 | 0.0371 | 0.0371 | 0.0371 |
| Quarter 2 | 0.0262 | 0.0362 | 0.0362 | 0.0363 |
| Quarter 3 | 0.0268 | 0.0372 | 0.0361 | 0.0362 |

RMSE 与 AUC 值相比, 差距很小. 然而, 从这些值中依然能够发现, 因子分解模型仍然保持着不低于 LR 的预测性能. 张量分解也同样显示出了不低于 MF 的预测性能. 表 8 给出了基于三元组排序优化学习策略即融合所有 4 类特征的模型的性能比较. 需要强调的是, 由于三元组排序优化学习是以排序为优化目标, 所以模型的输出值理论上是整个实数域, 所以在计算 RMSE 之前, 首先把模型的输出结果使用与文献 [40] 类似的方法转换到了 [0,1] 区间实数, 然后再计算 RMSE. 由于基于三元组排序优化的学习重点是优化样本之间的顺序而不是估值准确性, 正相反, 最小二乘法优化目标恰好是估值的准确性, 所以正如我们所料, 基于三元组排序化的 RMSE 值比基于回归的最小二乘法要差一些.

## 6 结论

本文针对广告点击率预测时遇到的数据稀疏问题, 提出了基于异构信息融合的张量分解模型来缓解数据稀疏. 与传统方法相比, 本文的方法主要有以下 5 个特点: 第一, 能有效地集成类别特征和数值特征. 我们允许特征值为类别特征, 例如: 性别特征分为男和女, 为男特征和女特征分别学习一个隐因

子向量; 也可以是数值特征, 例如: 广告位的长度特征, 为这类特征如广告位长度学习一个隐因子向量, 对数值特征做了归一化处理. 第二, 省略对象标识符, 而直接使用对象特征组合来描述对象, 提升特征粒度, 既能避免过拟合, 又能有效缓解数据稀疏. 第三, 能有效集成多值特征和单值特征, 对多值特征进行了归一化处理. 第四, 既能集成附加信息, 也能有效利用层次结构信息. 例如, 广告由广告商、广告活动和广告创意组合描述, 即使在历史中没有出现的新广告创意, 只要对应的广告活动或广告商曾经有历史投放数据, 新的广告创意就可以借鉴他们的表示. 第五, 既融合了历史点击信息也融合了历史转化信息, 有效利用两种信息的互补性, 取长补短, 互相促进和提升, 使得张量分解模型具有更强的抗数据稀疏能力, 提升预测性能.

## 参考文献


1 McMahan H B, Holt G, Sculley D, et al. Ad click prediction: a view from the trenches. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, 2013. 1222–1230

2 Graepel T, Candela J Q, Borchert T, et al. Web-scale bayesian clickthrough rate prediction for sponsored search advertising in microsoft's bing search engine. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), New York, 2010. 13–20

3 Alekh A, Olivier C, Miroslav D, et al. A reliable effective terascale linear learning system. J Mach Learn Res, 2014, 15: 1111–1133

4 Chapelle O, Manavoglu E, Rosales R. Simple and scalable response prediction for display advertising. ACM Trans Intel Syst Technol, 2015, 5: 1–34

5 Wu W C H, Yeh M Y, Chen M S. Predicting winning price in real time bidding with censored data. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015. 1305–1314

6 Li C, Lu Y, Mei Q Z, et al. Click-through prediction for advertising in twitter timeline. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015. 1959–1968

7 Menon A K, Chitrapura K P, Garg S, et al. Response prediction using collaborative filtering with hierarchies and side-information. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, 2011. 141–149

8 Wu K W, Ferng C S, Ho C H, et al. A two-stage ensemble of diverse models for advertisement ranking in KDD Cup 2012. https://www.csie.ntu.edu.tw/~htlin/paper/doc/wskdd12cup.pdf

9 Li S, Kawale J, Fu Y. Predicting user behavior in display advertising via dynamic collective matrix factorization. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, 2015. 875–878

10 Trofimov I, Kornetova A, Topinskiy V. Using boosted trees for click-through rate prediction for sponsored search. In: Proceedings of the 6th International Workshop on Data Mining for Online Advertising and Internet Economy, Beijing, 2012

11 Agarwal D, Agrawal R, Khanna R, et al. Estimating rates of rare events with multiple hierarchies through scalable log-linear models. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, 2010. 213–222

12 Zhang W N, Yuan S, Wang J. Real-time bidding benchmarking with iPinYou dataset. 2014. ArXiv:1407.7073

13 Zou Y Q, Jin X, Li Y, et al. Mariana: tencent deep learning platform and its applications. Proc VLDB Endow, 2014, 7: 1772–1777

14 Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer, 2009, 42: 30–37

15 Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Int Comput, 2003, 7: 76–80

16 Chen T Q, Tang L P, Liu Q, et al. Combining factorization model and additive forest for collaborative followee recommendation. 2012. http://www.cs.princeton.edu/~linpengt/papers/kddcup2012.pdf

17 Symeonidis P, Nanopoulos A, Manolopoulos Y. Tag recommendations based on tensor dimensionality reduction. In: Proceedings of the 2008 ACM Conference on Recommender Systems, Lausanne, 2008. 43–50

18 Rendle S, Balby M L, Nanopoulos A, et al. Learning optimal ranking with tensor factorization for tag recommendation. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, 2009. 727–736

19 Shen S, Hu B, Chen W Z, et al. Personalized click model through collaborative filtering. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining, Seattle, 2012. 323–332

20 Shan L L, Lin L, Shao D, et al. CTR prediction for DSP with improved cube factorization model from historical bidding log. In: Proceedings of International Conference on Neural Information Processing, Kuching, 2014. 17–24

21 Shan L L, Lin L, Sun C J, et al. Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization. Electron Com Res Appl, 2016, 16: 30–42

22 Shan L L, Lin L, Sun C J, et al. Optimizing ranking for response prediction via triplet-wise learning from historical feedback. Int J Mach Learn Cybern, 2017, 8: 1777–1793

23 Lee K, Orten B, Dasdan A, et al. Estimating conversion rate in display advertising from past erformance data. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, 2012. 768–776

24 Oentaryo R J, Lim E P, Low J W, et al. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, New York, 2014. 123–132

25 Agarwal D, Broder A Z, Chakrabarti D, et al. Estimating rates of rare events at multiple resolutions. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, 2007. 16–25

26 Wang X R, Li W, Cui Y, et al. Click-through rate estimation for rare events in online advertising. In: Online Multimedia Advertising: Techniques and Technologies. Hershey: IGI Global, 2010

27 Kota N, Agarwal D. Temporal multi-hierarchy smoothing for estimating rates of rare events. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, 2011. 1361–1369

28 Vargiu E, Giuliani A, Armano G. Improving contextual advertising by adopting collaborative filtering. ACM Trans Web, 2013, 7: 1–22

29 Dave K S, Varma V. Learning the click-through rate for rare/new Ads from similar Ads. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, 2010. 897–898

30 Agarwal D, Chen B C, Elango P. Spatio-temporal models for estimating click-through rate. In: Proceedings of the 18th International Conference on World Wide Web, Madrid, 2009. 21–30

31 Regelson M, Fain D. Predicting click-through rate using keyword clusters. In: Proceedings of the 2nd Workshop on Sponsored Search Auctions. New York: ACM, 2006

32 Richardson M, Dominowska E, Ragno R. Predicting clicks: estimating the click-through rate for new ADs. In: Proceedings of the 16th International Conference on World Wide Web, Banff, 2007. 521–530

33 Kolesnikov A, Logachev Y, Topinskiy V. Predicting CTR of new Ads via click prediction. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, 2012. 2547–2550

34 Cheng H, Zwol R V, Azimi J, et al. Multimedia features for click prediction of new Ads in display advertising. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, 2012. 777–785

35 Koren Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, 2008. 426–434

36 Menon A K, Elkan Charles. A log-linear model with latent features for dyadic prediction. In: Proceedings of the 10th International Conference on Data Mining, Piscataway, 2010. 364–373

37 Yang S H, Long B, Smola A, et al. Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of the 20th International Conference on World Wide Web, Hyderabad, 2011. 537–546

38 Chen T Q, Zheng Z, Lu Q X, et al. Feature-based matrix factorization. 2011. ArXiv:1109.2271

39 Yan L, Li W J, Xue G R, et al. Coupled group lasso for web-scale CTR prediction in display advertising. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), Beijing, 2014. 802–810

40 Tagami Y, Ono S, Yamamoto K, et al. CTR prediction for contextual advertising: learning-to-rank approach. In: Proceedings of the 7th International Workshop on Data Mining for Online Advertising, Chicago, 2013

41 Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback. In: Pro-

ceedings of the 25th Conference on Uncertainty in Artificial Intelligence, Montreal, 2009. 452–461
42 Liao H, Peng L X, Liu Z C, et al. iPinYou global RTB bidding algorithm competition dataset. In: Proceedings of the 8th International Workshop on Data Mining for Online Advertising, New York, 2014
43 Zhang W N, Yuan S, Wang J. Optimal real-time bidding for display advertising. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2014. 1077–1086
44 Zhang W N, Wang J. Statistical arbitrage mining for display advertising. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015. 1465–1474
45 Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982, 143: 29–36
46 Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn, 1997, 30: 1145–1159
47 Fawcett T. ROC graphs: notes and practical considerations for researchers. Mach Learn, 2004, 31: 1–38

# Response prediction via integration of heterogeneous information

Lili SHAN*, Lei LIN* & Chengjie SUN

*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001, *China*
* Corresponding author. E-mail: shanll@insun.hit.edu.cn, linl@insun.hit.edu.cn

**Abstract** In recent years, the tensor factorization model has been used to model complicated feature interactions involving multiple aspects, such as the user, publisher, and advertiser, for response prediction in real-time bidding. Among numerous challenges, data sparsity and cold start problems always bother researchers, particularly for ad conversion rate prediction. Such problems in prediction become difficult if only one or several types of information are considered. All types of heterogeneous information must be simultaneously integrated to address these problems. This paper proposes an availability solution for integrating heterogeneous information in the tensor factorization model to efficiently alleviate data sparsity and cold start problems. It proposes different integration strategies and implementation methods for various types of information depending on their property, category, structure, form, and function. This solution efficiently alleviates data sparsity and cold start problems, and enhances the prediction reliability and precision for the tensor factorization model in real-time bidding systems. Finally, this solution achieves a significant improvement in response prediction compared to baselines methods on the selection datasets.

**Keywords** real-time bidding, response prediction, tensor decomposition, integration of heterogeneous information, data sparsity, cold start problem, prediction method

**Lili SHAN** was born in 1976. She received her Ph.D. degree in computer science from the Harbin Institute of Technology (HIT), Harbin, in 2016. She is currently a lecturer at HIT. Her research interests include computational advertising and natural language processing.

**Chengjie SUN** was born in 1980. He received his Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, in 2008. Currently, he is an associate professor at HIT. His research interests include natural language processing, information extraction, text mining, and recommender system.

**Lei LIN** was born in 1970. He received his Ph.D. degree in computer science from HIT, Harbin, in 2004. He is currently an associate professor at HIT. His research interests include network information processing, natural language processing, and computational molecular biology.