基于特征学习的广告点击率预估技术研究

张志强 周 永 谢晓芹 潘海为
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

摘 要 搜索广告中的点击率预估问题在信息检索和机器学习等领域一直是研究的热点.目前通过设计特征提取方案获得特征和针对用户点击行为建模等方法,并没有充分考虑广告数据具有的高维稀疏性、特征之间存在高度非线性关联的特点,致使信息利用不充分.为了降低数据稀疏性和充分挖掘广告数据中隐藏的规律,该文提出了面向广告数据的稀疏特征学习方法.该方法基于张量分解实现特征降维,并充分利用深度学习技术刻画数据中的非线性关联,以解决高维稀疏广告数据的特征学习问题,实验结果验证了文中提出的方法能够有效地提升广告点击率的预估精度,达到了预期效果.

关键词 搜索广告;点击率;张量分解;深度学习;社交网络;社会媒体;计算广告学
中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2016.00780

Research on Advertising Click-Through Rate Estimation Based on Feature Learning

ZHANG Zhi-Qiang ZHOU Yong XIE Xiao-Qin PAN Hai-Wei
(*College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)

Abstract The issue of click through rate estimation in sponsored search has been widely studied in information retrieval, machine learning and query recommendation etc. Some related studies, such as the methods in which features are obtained by setting the feature extraction scheme or aiming at user behavior modeling, did not take into account those essential characteristics including the sparseness of advertising data and highly nonlinear association between features. In order to fully mining the hidden rules in advertising data, this paper proposes a method that can learn the sparse feature of advertising data. Our method combines dimension reduction based on tensor decomposition and takes full advantage of feature learning to portraying the nonlinear associated relationship of data to solve sparse feature learning problems. Finally, the comparison experiment shows this method has the desired effect of improving the accuracy of CTR estimation.

Keywords sponsored search; click through rate; tensor decomposition; deep learning; social networks; social media; computational advertising

1 引 言

搜索广告又称赞助商搜索(Sponsored Search),是指广告主根据自己的产品或服务,确定相关的关键词,制定广告创意、自主竞价并投放的广告.当用户检索到广告主购买的关键词时,对应的广告会被触发并展示,用户点击后按照计费规则对广告主收

收稿日期:2015-01-06;在线出版日期:2015-10-09.本课题得到国家自然科学基金(61370084,61202090,61272184)、教育部新世纪人才支持计划(NCET-11-0829)、黑龙江省自然科学基金(F201130)、中央高校基本科研业务费专项资金(HEUCF100609,HEUCFT1202)资助.
张志强,男,1973 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为信息检索、数据库、智能信息处理.E-mail:zqzhang@hrbeu.edu.cn.**周 永**,男,1988 年生,硕士研究生,主要研究方向为信息检索、数据挖掘.**谢晓芹**,女,1973 年生,博士,副教授,主要研究方向为服务计算、社会网络、智能信息处理.**潘海为**,男,1974 年生,博士,副教授,主要研究方向为数据挖掘、智能信息处理.

费.点击率(Click Through Rate,CTR)是指用户点击广告的概率,点击率预估是根据给定的〈查询,广告〉信息以及上下文环境信息等,预估用户点击广告的概率.

随着在线推广技术的发展,广告由过去"粗放式"投放正在向"精准化"投放转变,以数据驱动的广告精准投放已成为在线推广的主流趋势.在广告需求方平台(Demand Side Platform,DSP)的程序化购买和搜索广告投放的过程中,都需要评估用户对广告的偏好程度,而衡量这一偏好程度的重要指标就是广告的点击率[1-2].

搜索广告展现过程与网页搜索结果展现过程十分类似,包括查询分析,广告检索,广告排序等阶段.其中在广告排序阶段,目前相对成熟的搜索广告系统采用的排序规则是按照广告预期收益进行排序,预期收益等于广告质量度与广告竞价的乘积,其中,广告点击率是广告质量度最重要的衡量指标.广告预期收益[3]可以简写为式(1)形式:

$$E_{\text{ad}}(\text{revenue})=P_{\text{ad}}(\text{click})\times CPC_{\text{ad}} \quad (1)$$

$P_{\text{ad}}(\text{click})$指广告的预估点击率,$CPC_{\text{ad}}$表示发生一次点击行为时,搜索引擎的收益.也就是说,点击率预估是广告排序阶段的核心技术,直接影响搜索引擎的收益和用户体验,并且预估点击率对广告的后续投放具有非常重要的指导意义.

本文的主要贡献在于,针对广告数据呈现的高维稀疏性和特征之间存在着高度非线性关联的特点,从特征学习的角度出发提出了面向广告数据的稀疏特征学习方法;在数据降维和特征之间非线性关联深度学习方面进行了有意义的探索;同时,通过大量的实验与相类似方法进行了详细的比较分析,证明了本文方法的有效性.

本文第 1 节介绍研究问题的背景;第 2 节简要概括搜索广告点击率预估的相关工作以及存在的问题;第 3 节是本文的核心内容,根据当前研究工作存在的问题和广告数据自身的特点,提出面向广告数据的稀疏特征学习方法,并描述算法思想和具体过程;第 4 节给出实验方案设计,通过对比实验验证本文提出方法的预估效果;第 5 节总结全文工作,并指出存在的不足之处,对将来的工作方向进行讨论.

2　广告点击率预估

点击率预估一直是信息检索和机器学习等领域的热点问题.最初的研究是用来预估查询关键词与文档之间的真实相关性[4-5],建立预估模型消除点击数据中的各种偏倚因素,如页面上下文环境偏倚、广告位置偏倚、用户信息偏倚等,以获得查询与文档的真实相关性.后来将该研究应用在排序结果的优化和点击预估等方面.

在排序结果优化方面,通常将从点击预估模型中得到的真实相关性作为新特征加入到排序算法中,之后借助 A/B test 方法来检验新特征对算法排序能力的影响效果[6].例如 Dupret 等人[7]使用这种方式训练模型,实验表明该方法能有效提升结果排序的质量.

在点击预估方面,点击模型通常用来预估一个查询会话中各个文档的点击概率或一次会话中点击序列的概率.通过模型预估返回列表中各个文档的点击概率,有助于排序算法调整返回的文档顺序,使用户点击文档的概率最大化.这方面已有许多研究成果,如点击链模型[8]和动态贝叶斯网络模型[9]等.

2.1　相关工作

本文的核心为点击率预估模型,因此我们将从特征学习、用户行为和数据特点等 3 个角度来分别介绍点击率预估方面的相关研究工作.

(1)特征学习.影响点击率的特征很多,并不是考虑的特征越多,效果就会越好.在实际中往往要在精度与效益方面进行权衡.因此,需要尽可能获取与点击率高度相关的特征,以提高点击率预估的准确率.当前影响 CTR 预估最主要的特征是位置和相关性.位置决定了广告的曝光程度,Zhang 和 Jones[10]考察了原始查询和重写后的查询之间的相关性与广告点击率之间的关系,目的是提升查询重写的质量,从而提高广告的点击率.虽然其实际工作重点没有放到广告点击率预估上面,但是该工作考察了一些特征对广告点击率的影响,如次序(Rank)、长度差异(Length Difference)、编辑距离(Edit Distance)等.文献[3,8]将位置因素和广告查询相关性作为特征,同时考虑了根据相同或相似项(Term)的已知广告来解决稀疏广告或新广告的点击率预估问题.文献[11]综合了协同过滤,贝叶斯网络和特征工程等模型来预测点击率.该工作本质上是一种对多种模型的组合式运用,而非单一的预估模型.鉴于实际中"查询-文档"相关性模型并非对所有用户都是一致的,有很多被不同用户提交的相似查询往往导致不同的信息需求,因此文献[12]考虑了用户个性化信息,提出了基于协同过滤和张量分解来提取个性化特征.Hu 等人[13]结合用户查询意图,认为影响点击

率预估的因素不仅受到位置和文档因素的影响,还受到用户查询的真正意图与实际查询语句之间偏差的影响,作者利用贝叶斯方法基于意图假设进行建模.另外一些研究人员通过构造同一页面广告之间的相关性特征、广告与自然结果的相关性特征并融入点击预估模型,来提高点击率预估的准确率.

(2)用户行为建模.通过假设检验,借助贝叶斯网络刻画用户浏览场景,进而估计出用户点击广告的概率.对用户行为建模通常是基于一个前提假设,即搜索结果返回列表中的任意一个文档,只有先被查阅到,用户才有可能发生点击行为,如果文档没有被查阅到,则一定不会被点击.Taylor 等人[14]基于这种最简单的浏览行为假设提出级联模型.进一步考虑,如果用户点击一个文档后,该文档不能满足用户的查询需求,则用户可能仍会向后检视搜索结果并有可能发生点击行为,因此 Guo 等人[15]对级联模型的假设做了扩展,扩展至多次点击.多次点击是指用户点击一个文档后,仍可能浏览后续文档并计算下一个位置文档点击发生的概率.点击链模型[7]是针对用户与搜索结果交互行为而建模的生成模型,被点击文档 d_i 的相关性影响继续浏览的可能性,被点击文档 d_i 的相关性越大,则继续浏览下一文档 d_{i+1} 的可能性就越小,这说明相关性越大的文档越能满足用户的查询意图.动态贝叶斯网络模型[8]的建模过程中,引入了两个文档相关性变量,即观察相关性(perceived relevance)和实际相关性(actual relevance).观察相关性用来衡量用户点击广告链接(URL)的概率;实际相关性用来衡量用户进入广告链接后,该搜索结果的真实满意度.用户浏览模型[4]认为用户点击下一个位置文档的概率跟上次点击位置的距离有关,同时也与当前文档的位置有关,引入距离变量是因为当用户浏览了一系列不相关文档后,则倾向于放弃此次搜索结果.

(3)数据稀疏性特点.从广告数据稀疏性特点出发,研究广告点击率预估问题.Richardson 等人[10]利用包含稀疏广告相同或相似项的已知广告来预估其点击率.CTR 预估中最大的挑战之一就是信息的缺失,尤其对于新广告而言,历史展示数据信息过少,无法给点击预估模型提供预估基准.因此,针对广告数据的特点,Rain 等人[16]基于"竞拍词-广告主"矩阵,提出了层次聚类的方法解决历史数据不充分的广告 CTR 预估问题.

Agarwal 等人[17]从建模的角度设计适应稀疏广告或新广告的点击预估模型,分别提出了基于层级结构的预估模型和基于 Time-Spatial 的预估模型[18].文献[19]提出的基于经验贝叶斯的自然数据分层和基于数据一致性的两种平滑计算方法对层级模型做了改进.

由于不同模型之间的兼容性问题,实际可计算问题和问题自身的复杂性,目前还没有出现一个大一统的模型,能够覆盖所有方面,绝大部分工作都是通过不同角度来考察单一模型的预估效果,本文工作亦属此类.文献[11]做出了一种有意义的尝试,它通过综合运用多种模型来预测点击率,试图利用不同模型和方法的互补性来实现更高精度的预估结果.虽然结果显示这种组合式方法比单一模型好,但该工作只是单纯的从预估效果角度来做调整,并没有对如何组合运用现有模型给出清晰合理的解释,相关做法难于理解.该工作只说明了这种方式有效,至于为什么有效及如何更有效没有给出具体理论指导意义的结论.对于各种不同模型之间关联问题的研究是个挑战,而该项工作正是我们后续的一个研究课题.

2.2 存在的问题

尽管点击率预估方法已经得到了广泛研究,但是仍然存在一些问题.目前,通过人工构造特征的方法,存在效率低、可扩展性和性能提升困难的问题.而贝叶斯网络刻画用户浏览行为,存在信息利用不充分,并且没有考虑到广告数据具有高度稀疏性、特征之间存在高度非线性关联的本质特点.已有的从广告数据特点出发进行点击率预估问题的研究中,仅仅考虑了稀疏广告(即展示不充分、统计量不足的广告)点击率预估问题,实际上并未从整体角度考虑广告数据的本质特点.

广告数据具有高维稀疏性特征,高维特征中有效信息(非 0 值)的维度很低,其中包含的噪声会对真实信息干扰很大.已有的研究成果很少从广告数据的特点考虑,使得大多数 CTR 预估方法无法高效地在高维、高稀疏的广告数据上准确地预估点击率.如何解决高维稀疏数据给 CTR 预估准确率带来的影响,是一个值得研究的问题.

广告数据另一个特点是特征之间存在高度非线性关联关系,复杂度高.传统方法采用人工构造组合特征(又称"人工特征工程")的方式,挖掘特征之间的关联,但是该方式存在效率低、领域知识无法迁移等诸多问题.因此,如何在减少人工干预的情况下,通过算法自动挖掘特征之间的关联是文中要研究的重点.

3　基于特征学习的 CTR 预估方法

特征是数据的抽象表示形式，是用于表达数据中隐藏的、且对具体任务有帮助的标签系统. 从原始广告数据中挖掘与预估任务高度关联的特征是点击率预估系统的关键步骤之一. 然而，广告数据呈现的高维稀疏性且特征之间存在高度非线性关联的特点，使得已有方法无法高效地进行 CTR 预估. 因此，本文从如何降低特征的高维稀疏性以及如何刻画特征之间的非线性关联的角度出发，提出了面向广告数据的稀疏特征学习方法(Advertising Data-oriented Sparse Feature Learning Method，ADoSFLM)，以解决高维稀疏广告数据的特征学习问题.

3.1　问题描述

本文研究的广告点击率预估问题可以描述为：给定一个用户查询和其他信息(如性别、年龄、地域、兴趣爱好等)，经过查询分析、广告检索步骤后，得到一个与查询相关的广告候选集. 点击率预估系统需要计算用户点击广告候选集中每一则广告的概率，即广告的预估点击率. 图 1 中的灰色模块描述了一个广告点击率预估系统的工作流程.

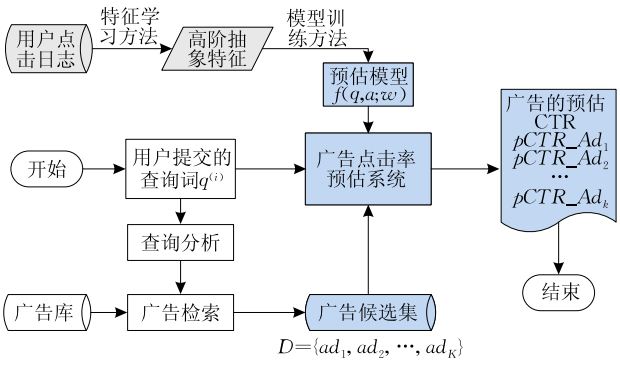

图 1　点击率预估系统工作流程

搜索广告中，广告主通过购买竞拍词的方式设定了广告被触发的场景，广告投放系统根据用户的查询关键词，匹配相应的广告，对广告排序后与自然搜索结果一起返回给用户.

3.2　算法基本思想

本文针对广告数据的特点提出 ADoSFLM 方法用于挖掘特征之间的内在关联，获取对数据有更强表达能力的特征集合，进而得到更加精确的点击率预估模型. 该方法主要包含以下 3 个环节：

(1) 数据降维. 数据降维是解决数据稀疏性的一个有效手段. 针对广告数据中相同类型对象内部之间存在相似性关系，首先对相似对象进行聚类，获得初始的聚合数据；然后，对于不同类型对象之间存在的复杂关联关系，采用张量结构对其建模，并运用张量分解法得到近似张量.

(2) 复杂特征学习. 广告数据的特征之间存在高度非线性关联的特点，而高阶多项式函数可以有效地刻画高度关联关系. 本文研究并利用深度学习模型——栈式自编码神经网络算法——利用其多层网络结构逐层学习特征之间的非线性关联，将学习到的组合特征集合用于描述广告数据中隐藏的内在规律.

(3) CTR 预估模型. 利用学到的特征集合训练预估模型，得到点击预估模型. 这样，给定一个新的样本，经过降维和特征学习的变换后，点击模型可以预估广告的点击率.

3.3　数据降维方法

降维是为高维数据获取一个能反映原始数据内在结构特性的低维表示，同时达到降噪、降低稀疏性的目的. 点击日志数据中包含了用户、查询、广告等多种类型的对象，这些对象之间的关系很复杂. 相同对象内部之间存在关系，如广告对象内部之间存在相似性关系. 同时，不同类型对象之间也存在着复杂的关系，如给定一个特定用户和该用户提交的查询，需要预估用户是否会点击广告，以及点击广告的概率大小，用户、查询和广告 3 个对象之间存在复杂的隐含关系.

本文结合点击日志数据的特点，分别从相同对象内部之间存在相似性关系和不同对象之间存在关联关系这两个角度出发进行降维.

3.3.1　相同对象聚类表示

点击日志数据中，用户之间、查询之间以及广告之间都存在相似性，例如用户输入的同一个查询所返回的广告之间具有相似性，触发同一个广告的不同查询之间同样存在相似性. 同样的，输入相似查询的不同用户由于查询意图近似也具有相似性. 因此，首先从相似性的角度对用户、查询和广告 3 个维度进行降维.

本文采用基于距离划分的 K-means 聚类算法[20]对查询、广告和用户进行聚类. 目的是通过聚类使得相似对象聚合到同一簇中，同一簇中的对象相似度尽可能的高，获得初始的聚合数据. 直接利用传统 IR 领域的文本相似性技术实现对查询甚至广告的聚类，虽然可以完成聚类任务，但是这种做法没有考虑到实际数据中蕴含的广告与查询之间的关联关系，得到的聚类结果对点击率预估没有任何参考

意义.因此为了能够更好的挖掘和利用广告与查询之间的关联关系,本文提出了新的方法,用实验数据中提供的广告展示次数作为广告 A_i 与查询 Q_j 的权重,来建立广告-查询矩阵 $\boldsymbol{W}_{N_a\times N_q}$,其中 N_a 表示广告数,N_q 表示查询数,w_{ij} 表示〈A_i,Q_j〉之间的权重.对该广告-查询矩阵采用 K-means 算法进行聚类,得到一个相对密集的数据集合.以广告聚类为例,图 2 介绍了聚类算法的流程,对查询的聚类采取同样处理方式.

输入:广告-查询矩阵 $\boldsymbol{W}_{M\times N}$,聚类簇数 K
输出:K 个广告簇集合
1. 对广告-查询矩阵 $\boldsymbol{W}_{M\times N}$ 扫描,得到所有的 M 个广告和 N 个查询,分别记作 $\boldsymbol{A}=\{a_1,a_2,\cdots,a_M\}$ 和 $\boldsymbol{Q}=\{q_1,q_2,\cdots,q_N\}$;
2. 从 M 个广告中随机抽取 K 个作为最初的聚类中心点,记作 $\boldsymbol{T}=\{t_1,t_2,\cdots,t_k\}$;
3. 初始化 K 个聚类集合$\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_K\}$为空集;
4. 计算每个广告 a_i 与各个聚类中心点 t_j 之间的距离,计算公式如下:
$$Dis(a_i,t_j)=\sqrt{\sum_{c\in G_{ij}}(W_{a_i}-W_{q_j})^2}$$
(其中 G_{ij} 表示广告 a_i 与作为聚类中心的广告 t_j 共同展现的查询集合,W_{a_i},W_{q_j} 分别是广告 a_i 与 t_j 的权重(展示次数),$Dis(a_i,t_j)$表示 a_i 与 t_j 的距离);
5. 若 $Dis(a_i,t_j)=\max\{D(a_i,t_1),D(a_i,t_2),\cdots,D(a_i,t_k)\}$,则广告 a_i 属于簇 $\boldsymbol{P}_j$;
6. 计算同一聚类集合中所有广告的平均权重值,重新生成聚类中心;
7. 如果聚类中心的偏差达到了设定的阈值,则聚类完成;否则转到步 4 重新计算.

图 2 聚类算法

基于广告-查询矩阵 $\boldsymbol{W}_{N_a\times N_q}$,通过图 2 所示的聚类算法完成对广告/查询的聚类,使得同一簇中的广告/查询相似度尽可能的高.本文具体作法是基于同一个广告-查询矩阵分别对广告和查询作聚类,两次聚类相互独立,聚类顺序不影响后续的计算.对于用户维度的聚类,考虑到具有相似查询需求的用户具有相似性,本文直接根据前面得到的查询聚类结果,将同一簇中的查询所对应的用户聚在一起组成一个用户簇.

初始数据中的用户数、查询数和广告数分别用 N_u,N_q 和 N_a 表示,相同类型对象内部聚类后,属于同一个簇中的对象用同一个 ID 表示,将聚类后的用户、查询和广告的簇数分别用 K_u,K_q,K_a 表示.这样,初始数据集中的用户数、查询数和广告数由原来的 N_u,N_q 和 N_a 分别降维到 K_u,K_q 和 K_a.

假设 T_q 和 T_a 分别表示对查询和广告调用聚类算法完成聚类任务而需要执行的迭代次数,K 表示聚类的个数,则完成对查询聚类的时间复杂性为 $O(KT_qN_q)$,而对广告进行聚类的时间复杂性为 $O(KT_aN_a)$.由于用户的聚类没有调用图 2 的算法,而是根据查询的聚类结果直接得到的,其复杂性为 $O(N_q+N_u)$,令 $T=\max\{T_q,T_a\}$,$N=\max\{N_u,N_q,N_a\}$,则聚类环节的复杂性表示为 $O(KT_qN_q)+O(KT_aN_a)+O(N_q+N_u)=O(KTN)$.

3.3.2 不同对象复杂关联关系求精

点击日志数据中的用户-查询-广告之间存在三元关系.传统的降维方法(如 PCA 等)不仅破坏了三者之间的内在关系,当数据维度数很大时,容易导致维数灾难.为此,本文用三维张量结构模型表示用户、查询和广告三维数据,然后利用张量分解法进行降维.张量模式降维充分保留了用户、查询和广告之间的结构信息和内在关联,由于参数更少,对于高维数据来说,张量模式的降维要比向量模式有更好的约简效果.

定义 1. 张量(Tensor)[21].

张量是一个定义在向量空间和对偶空间笛卡尔积上的多线性函数.在 n 维空间内,有 n^r 个分量(r 是张量的秩或阶),其中每个分量都是坐标的函数,在进行坐标变换时,这些分量同样根据相应规则做线性变换.其中一阶张量($r=1$)称为向量,二阶张量($r=2$)称为矩阵,矩阵分解是张量分解的特殊形式.

基于聚类后的数据,建立"用户-查询-广告-权重"四元关系〈u_i,q_j,a_k,$w_{i,j,k}$〉,关于权重的计算有很多种方式,本文结合实验用的数据,利用聚类后广告簇中广告的展示数之和作为三维空间中元素的权重来构建三维张量模型,用 $\boldsymbol{H}\in R^{K_u\times K_q\times K_a}$ 表示,模型如图 3 所示,三个维度的维度数分别是 K_u,K_q,K_a.构建三维张量 $\boldsymbol{H}\in R^{K_u\times K_q\times K_a}$ 后,本文利用塔克分解法(Tucker Factorization)[22],分解张量 $\boldsymbol{H}$,公式表示如下:

$$\boldsymbol{H}=\boldsymbol{C}\times_u\boldsymbol{U}\times_q\boldsymbol{Q}\times_a\boldsymbol{A}=\sum_{p=1}^{P}\sum_{t=1}^{T}\sum_{r=1}^{R}c_{pqr}\boldsymbol{u}_p\circ q_t\circ a_r$$
$$=[\boldsymbol{C};\boldsymbol{U},\boldsymbol{Q},\boldsymbol{A}] \quad (2)$$

式(2)中的 $\boldsymbol{C}$ 表示张量 $\boldsymbol{H}$ 的核心张量(Core Tensor),类似于奇异值分解的对角矩阵,本文用 $\boldsymbol{U}$,$\boldsymbol{Q}$,$\boldsymbol{A}$ 表示张量 $\boldsymbol{H}$ 在维度 K_u,K_q,K_a 上对应的特征矩阵,是张量 $\boldsymbol{H}$ 在对应 3 个维度上的主成分.

Tucker 分解原理示意图,如图 3 所示.

Tucker 分解的目的是找到一个与原始张量 $\boldsymbol{H}$ 的近似张量,并且最大程度保留原始的张量信息和结构信息[23].Tucker 分解计算可以得到一个与原始张量相近的张量表示 $\hat{\boldsymbol{H}}$,如最小化公式(3)所示:

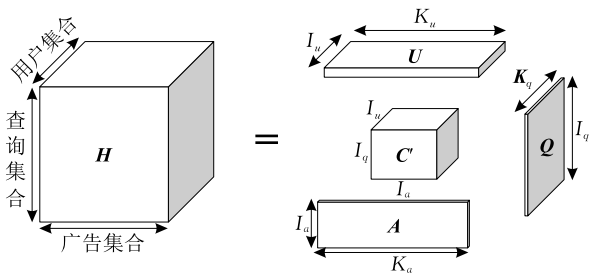

图 3　Tucker 分解原理示意图

$$\min_{\hat{\boldsymbol{H}}}\|\boldsymbol{H}-\hat{\boldsymbol{H}}\|,$$
$$\hat{\boldsymbol{H}}=\boldsymbol{C}\times_u\boldsymbol{U}\times_q\boldsymbol{Q}\times_a\boldsymbol{A}=[\boldsymbol{C};\boldsymbol{U},\boldsymbol{Q},\boldsymbol{A}] \tag{3}$$

式(3)表示原始张量与近似张量的近似程度，是优化的目标函数. 根据式(3)可以得到核心张量的表达式：

$$\boldsymbol{C}=\boldsymbol{H}\times_u\boldsymbol{U}^{\mathrm{T}}\times_q\boldsymbol{Q}^{\mathrm{T}}\times_a\boldsymbol{A}^{\mathrm{T}} \tag{4}$$

目标函数可以写成平方形式，即

$$\begin{aligned}&\|\boldsymbol{H}-[\boldsymbol{C};\boldsymbol{U},\boldsymbol{Q},\boldsymbol{A}]\|^2\\&=\|\boldsymbol{H}\|^2-2\langle\boldsymbol{H}\times_u\boldsymbol{U}^{\mathrm{T}}\times_q\boldsymbol{Q}^{\mathrm{T}}\times_a\boldsymbol{A}^{\mathrm{T}},\boldsymbol{C}\rangle+\|\boldsymbol{C}\|^2\\&=\|\boldsymbol{H}\|^2-2\langle\boldsymbol{C},\boldsymbol{C}\rangle+\|\boldsymbol{C}\|^2\\&=\|\boldsymbol{H}\|^2-\|\boldsymbol{C}\|^2\\&=\|\boldsymbol{A}\|^2-\|\boldsymbol{H}\times_u\boldsymbol{U}^{\mathrm{T}}\times_q\boldsymbol{Q}^{\mathrm{T}}\times_a\boldsymbol{A}^{\mathrm{T}}\|^2\end{aligned} \tag{5}$$

$\|\boldsymbol{H}\|^2$是一个常数，由原来的张量 $\boldsymbol{H}\in R^{K_u\times K_q\times K_a}$ 确定. 因此，目标函数转化为式(3)中右边的最大化问题的最优解，即

$$\max\|\boldsymbol{H}\times_u\boldsymbol{U}^{\mathrm{T}}\times_q\boldsymbol{Q}^{\mathrm{T}}\times_a\boldsymbol{A}^{\mathrm{T}}\|^2 \tag{6}$$

式(6)中的目标可以写成如下形式：

$$\begin{aligned}&\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}\|,\ \boldsymbol{W}=\boldsymbol{H}\times_q\boldsymbol{Q}^{\mathrm{T}}\times_a\boldsymbol{A}^{\mathrm{T}};\\&\|\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{W}\|,\ \boldsymbol{W}=\boldsymbol{H}\times_q\boldsymbol{U}^{\mathrm{T}}\times_a\boldsymbol{A}^{\mathrm{T}};\\&\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\|,\ \boldsymbol{W}=\boldsymbol{H}\times_q\boldsymbol{U}^{\mathrm{T}}\times_a\boldsymbol{Q}^{\mathrm{T}}\end{aligned} \tag{7}$$

在求最优解的过程中，需要固定其他维度的特征矩阵，即变量 $\boldsymbol{W}$，依次求解 $\boldsymbol{U}^{\mathrm{T}},\boldsymbol{Q}^{\mathrm{T}},\boldsymbol{A}^{\mathrm{T}}$，然后对 $\boldsymbol{U}^{\mathrm{T}},\boldsymbol{Q}^{\mathrm{T}},\boldsymbol{A}^{\mathrm{T}}$进行 SVD 分解. 对 $\boldsymbol{U}^{\mathrm{T}},\boldsymbol{Q}^{\mathrm{T}},\boldsymbol{A}^{\mathrm{T}}$进行 SVD 分解时，首先展开张量 $\boldsymbol{H}$，在用户、查询、广告维度展开张量 $\boldsymbol{H}$ 成为矩阵，分别记作 $\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3$，然后在这 3 个矩阵 $\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3$上应用奇异值分解，可得到

$$\begin{aligned}\boldsymbol{H}_1&=\boldsymbol{U}\cdot\boldsymbol{C}_1\cdot\boldsymbol{V}_1^{\mathrm{T}};\\\boldsymbol{H}_2&=\boldsymbol{Q}\cdot\boldsymbol{C}_2\cdot\boldsymbol{V}_2^{\mathrm{T}};\\\boldsymbol{H}_3&=\boldsymbol{A}\cdot\boldsymbol{C}_3\cdot\boldsymbol{V}_3^{\mathrm{T}}\end{aligned} \tag{8}$$

对于矩阵 $\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3$，需要确定 3 个维数的参数，分别是左奇异值矩阵 $\boldsymbol{U},\boldsymbol{Q},\boldsymbol{A}$ 中的维数 c_1,c_2,c_3. 这 3 个参数决定张量 $\boldsymbol{H}$ 的核心张量 $\boldsymbol{C}$ 的维数. 3 个对角的奇异值矩阵 $\boldsymbol{C}_1,\boldsymbol{C}_2$ 和 $\boldsymbol{C}_3$ 是通过对张量 $\boldsymbol{H}$ 的展开矩阵 $\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3$ 进行奇异值分解得到的，而核心张量 $\boldsymbol{C}$ 的计算则是通过 3 个对角奇异值矩阵 $\boldsymbol{C}_1,\boldsymbol{C}_2$ 和 $\boldsymbol{C}_3$ 求得. 维数 c_1,c_2,c_3 的计算则通过对 $\boldsymbol{C}_1,\boldsymbol{C}_2$ 和 $\boldsymbol{C}_3$ 的对角奇异值从大到小按照比例计算而得. 保留大的奇异值，按照比例删减小的奇异值，从而达到维数的归约，对原始张量降维的目的. 本文在降维的过程中将删减奇异值的比例分别设置为 50%. 这样，降维后新的核心张量 $\boldsymbol{C}'$计算公式如下：

$$\boldsymbol{C}'=\boldsymbol{H}\times_u\boldsymbol{U}_{r_1}^{\mathrm{T}}\times_q\boldsymbol{Q}_{r_2}^{\mathrm{T}}\times_a\boldsymbol{A}_{r_3}^{\mathrm{T}} \tag{9}$$

确定新核心张量 $\boldsymbol{C}'$以及新的特征矩阵 $\boldsymbol{U}_{r_1}$，$\boldsymbol{Q}_{r_2}$，$\boldsymbol{A}_{r_3}$后，构建新的近似张量 $\boldsymbol{H}'$：

$$\boldsymbol{H}'=\boldsymbol{C}'\times_u\boldsymbol{U}_{r_1}\times_u\boldsymbol{Q}_{r_2}\times_a\boldsymbol{A}_{r_3} \tag{10}$$

初始张量 $\boldsymbol{H}$ 的 3 个维度数分别是 K_u,K_q,K_a，经过降维后的近似张量 $\boldsymbol{H}'$的 3 个维度数分别用 I_u，I_q,I_a表示. Tucker 分解算法的时间复杂性与张量的各个维度成正比，可以表示为 $O(K_uK_qK_a)$. 由于我们之前利用聚类方法已经实现了对原始矩阵的降维，使得此处的 Tucker 分解开销大大降低，效率和精度都有显著提高.

3.4　复杂特征学习方法

机器学习领域的研究表明，深度或层次结构的模型对于刻画数据中的非线性关系和复杂模式更有效[24]. 受到这项研究的启发，本文利用一种能够刻画特征之间高度非线性关联的方法——栈式自编码网络算法(Stacked Auto Encoder Network，SAEN)[25]用于学习广告数据中的结构信息，使用该算法自动学习数据中的模式特征，并将学到的特征融入到建模(如分类、预测)的过程中，从而克服人工特征工程的不完备性缺陷.

3.4.1　输入层特征构成分析

广告数据中的特征之间存在高度非线性关联，虽然通过 Tucker 分解获得原始张量降维后的近似张量，但仅仅反映 User，Query，Ad 三个特征维度之间的信息，数据中其他对点击率预估有用的信息没有充分利用，如广告在返回页面的位置、广告数量以及用户年龄、性别等信息. 本文结合张量降维后的〈User，Query，Ad〉特征以及日志数据中其他有效信息作为特征学习的对象，输入层特征的构成总结如下：

(1) ID 类特征. ID 类特征可以唯一标识某类实体，在实际的点击日志中，通常使用一组数字字符串来表示此类变量，如‘10010’可标识唯一一个用户群体. 文中用到的 ID 类特征有用户 ID(UserID)，查询 ID(QueryID)，广告 ID(AdID)，广告位置 Position 以及返回页上的广告数 Depth.

值得注意的是，这里的 UserID，QueryID 和 AdID 是经过 K-means 聚类和张量降维后得到的“虚拟”

ID 类集合. 最初的用户数、查询数和广告数分别是 N_u, N_q, N_a，经过 K-means 聚类后的数量分别是 K_u, K_q, K_a，再经过张量降维后的用户数、查询数和广告数分别是 I_u, I_q, I_a.

(2) 属性类特征. ID 类特征仅仅是一个标识，此类特征无法从新出现的实体数据中获得，泛化能力比较弱. 而属性类特征可以用于描述某类用户集合、广告集合等，有较好的泛化能力，该类特征可命中多个实例. 例如，通过 IP 属性类特征可以知道用户的地理位置信息，那么对地域有要求的广告，如肯德基等餐饮类广告，IP 能够提供较大的信息量. 因此，有必要将属性类特征作为深度学习的输入层特征. 常用的属性类特征有：用户所在网址 IP，广告被触发的时间 Time，用户性别 Gender，用户年龄 Age，查询关键词 Keywords 等.

Time 类特征是一种对用户行为影响比较强的特征，用于记录用户行为发生的时段. Keywords 是由查询字符串去除停用词后切词得到的.

(3) 统计类特征. 统计类特征尝试使用历史数据的统计信息给点击预估模型提供预估基准，例如，在某一广告位上所有展现的广告平均 CTR 值. 文中的统计类特征有广告历史展现次数 Shows，广告历史点击次数 Clicks 以及广告位置归一化后的点击率 COEC.

关于 COEC 的计算，文献[9]根据式(11)计算得到去除位置偏倚后的历史点击率，COEC 计算公式为

$$\mathrm{COEC}(a)=\sum_{p}c(p,a)\Big/\sum_{p}ec(p,a) \tag{11}$$

其中，式(11)的分子表示广告 a 在所有位置上的点击次数之和，分母表示广告 a 在各个位置上的期望点击次数之和.

综上所述，在本文的实验中，SAEN 算法的输入层特征构成情况如图 4 所示.

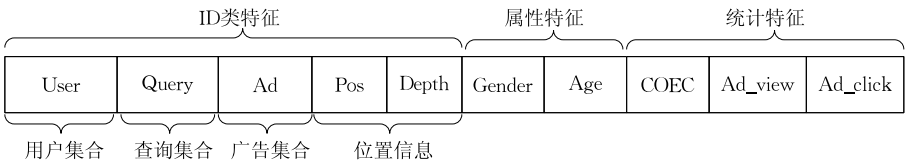

图 4 输入层特征构成情况

3.4.2 自编码器

自编码器(Auto Encoder)[26]是一个尽可能复现初始特征的深度学习算法，通常被用来学习原始数据更好的特征表示，由 3 层网络结构组成：底层是输入层 I、中间为隐藏层 H(新的数据表示层)以及输出层 O. 自编码网络如图 5 所示.

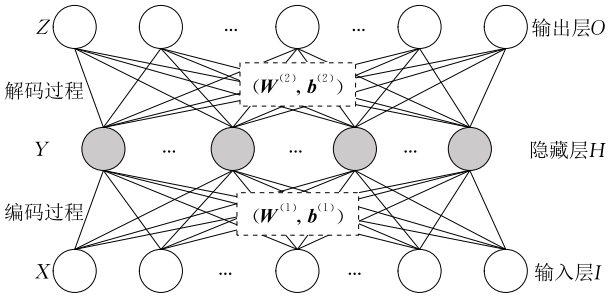

图 5 自编码网络示意图

输入数据经过隐藏层，在输出层重构，通过最小化输入和输出的重构误差来校准网络权重，学习输入数据的潜在特征或数据的压缩表示. 本文用 ins 和 $hids$ 分别表示输入层单元的个数和隐藏层单元个数，给定一个训练数据集合 $\boldsymbol{X}=\{\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\cdots,\boldsymbol{x}^{(M)}\}$，其中 $\boldsymbol{x}^{(i)}\in\boldsymbol{R}^{ins}$，自编码器将样本 $\boldsymbol{x}^{(i)}$ 映射到 $\boldsymbol{y}^{(i)}$(其中 $\boldsymbol{y}^{(i)}\in\boldsymbol{R}^{hids}$)，映射的过程就是一个编码的过程，用 Sigmoid 函数作为连接函数完成编码过程，表示如下：

$$\boldsymbol{y}^{(i)}=f(\boldsymbol{W}^{(1)}\cdot\boldsymbol{x}^{(i)}+\boldsymbol{b}^{(1)}) \tag{12}$$

参数 $\boldsymbol{W}^{(1)}\in\boldsymbol{R}^{hids\times ins}$ 是一个编码权重矩阵，偏置向量 $\boldsymbol{b}^{(1)}\in\boldsymbol{R}^{hids}$. 之后，由隐藏层 $\boldsymbol{y}^{(i)}$ 到输出层 $\boldsymbol{z}^{(i)}\in\boldsymbol{R}^{ins}$，是重构输入向量的过程，也是一个解码过程，目的是尽可能重构输入向量 $\boldsymbol{x}^{(i)}$，由 $\boldsymbol{y}^{(i)}$ 映射到 $\boldsymbol{z}^{(i)}$ 用一个线性映射表示如下：

$$\boldsymbol{z}^{(i)}=\boldsymbol{W}^{(2)}\boldsymbol{y}^{(i)}+\boldsymbol{b}^{(2)}\approx\boldsymbol{x}^{(i)} \tag{13}$$

这里 $\boldsymbol{W}^{(2)}\in\boldsymbol{R}^{ins\times hids}$ 和 $\boldsymbol{b}^{(2)}\in\boldsymbol{R}^{ins}$ 分别表示解码过程的权重矩阵和偏置向量. 从学习的角度，自编码算法旨在最小化输入 $\boldsymbol{x}^{(i)}$ 与输出 $\boldsymbol{z}^{(i)}$ 之间的重构误差，得到编码和解码过程的参数集合. 这里 $\boldsymbol{Z}=\{\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\cdots,\boldsymbol{z}^{(N)}\}$，用 $J(\boldsymbol{X},\boldsymbol{Z})$ 表示重构误差：

$$J(\boldsymbol{X},\boldsymbol{Z})=\frac{1}{2}\sum_{i=1}^{M}\|\boldsymbol{x}^{(i)}-\boldsymbol{z}^{(i)}\|^2 \tag{14}$$

图 6 描述了自编码器算法的学习流程.

虽然自编码器可以通过最小化输入输出的重构误差获得输入数据的隐藏表示，但是它只有一个隐层，属于浅层结构模型，限制了对数据的表示能力，无法刻画广告数据中特征之间的高度非线性关联. 因此，本文利用含有多个隐藏层的栈式自编码网络算法学习特征之间非线性的关联，本质上是学习单

输入：$\boldsymbol{X}$={数据集合}，ε=随机梯度下降算法的学习率，$iters$=参数迭代次数，num_i=输入层节点数，num_h=隐藏层节点数

输出：隐层结果 $Result$，网络连接权重矩阵 $\boldsymbol{W}^{(1)}$，输入层偏置向量 $\boldsymbol{b}$

1. 初始化网络连接权重矩阵 $\boldsymbol{W}$，矩阵大小为 $num\_h \times num\_i$，输入层偏置向量 $\boldsymbol{b}$，隐藏层偏置向量 $\boldsymbol{c}$；
2. 利用式(12)连接函数求出隐层 $\boldsymbol{Y}(\boldsymbol{X},\boldsymbol{W}^{(1)},\boldsymbol{b})$，利用式(14)求出重构层 $Z(\boldsymbol{Y},\boldsymbol{W}^{(2)},\boldsymbol{c})$. 其中 $\boldsymbol{W}^{(2)}$ 是 $\boldsymbol{W}^{(1)}$ 的转置矩阵；
3. 根据重构层 $\boldsymbol{Z}$ 和输入层 $\boldsymbol{X}$，构造重构误差 $J(\boldsymbol{X},\boldsymbol{Z})$. 即损失函数；
4. 分别对损失函数求参数 $\boldsymbol{W}^{(1)},\boldsymbol{b},\boldsymbol{c}$ 的偏导，分别用表示；
5. For i from 1 to iters do
6. $\quad \boldsymbol{W}^{(1)} \leftarrow \boldsymbol{W}^{(1)} + \varepsilon \times \partial J(\boldsymbol{X},\boldsymbol{Z})/\partial \boldsymbol{W}^{(1)}$
7. $\quad \boldsymbol{b} \leftarrow \boldsymbol{b} + \varepsilon \times \partial J(\boldsymbol{X},\boldsymbol{Z})/\partial \boldsymbol{b}$
8. $\quad \boldsymbol{c} \leftarrow \boldsymbol{c} + \varepsilon \times \partial J(\boldsymbol{X},\boldsymbol{Z})/\partial \boldsymbol{c}$
9. End-For
10. 计算隐藏层潜在表示 $Result = f(\boldsymbol{W} \times \boldsymbol{X} + \boldsymbol{b})$
11. 返回 $Result,\boldsymbol{W},\boldsymbol{b}$.

图 6　复杂特征学习的自编码器算法

特征之间的组合特征形式，构造最佳的(组合)特征集合用以提高模型预估 CTR 的准确率.

3.4.3　基于 SAEN 的特征学习方法

SAEN 算法是由多个自编码器组成的多层深度网络结构，其特点是每一个隐藏层都是对上一层的输出进行非线性变换得到的. SAEN 特征学习示意图，如图 7 粗矩形框部分所示.

SAEN 算法的一个很重要的特点是从输入层特征(原始特征)中学习或发现高度非线性或复杂的模式，直接从数据中自动学习潜在特征表示. 本文利用 SAEN 算法学习广告数据中的高阶组合特征过程，描述如下：

(1)将 3.4.1 节中提取的初始特征作为模型的输入，对初始特征做特征非线性变换得到第 1 隐藏层，即低阶组合特征，如图 7 所示.

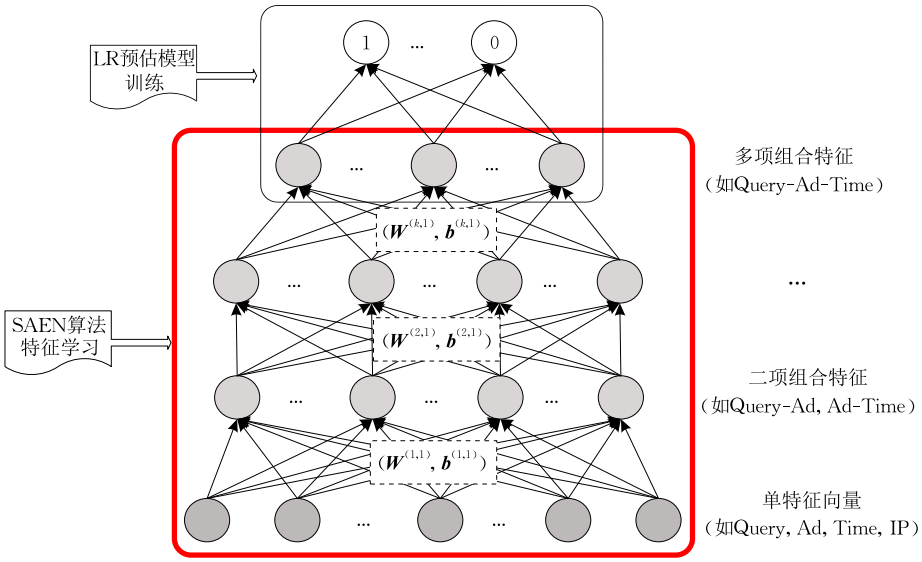

图 7　特征学习过程与 LR 预估模型训练示意图

(2)将低阶组合特征作为新的学习的对象，再次经过非线性变换得到相对高阶的组合特征，此过程重复下去，直到达到设定的隐藏层数为止.

其中，多层网络中不同的隐藏层是对输入层的不同潜在表示，高层的特征是低层特征的组合，从低层到高层的特征表达越抽象和概念化，也就越能挖掘数据中蕴藏的有价值信息.

为了更好的学习网络权重参数，本文采用文献[27]中提出的基于逐层贪婪训练的无监督学习算法. 逐层贪婪学习的关键是逐层训练网络权重参数，每次只学习相邻两层节点的连接权重，通过逐层学习以获得全局的 SAEN 模型参数. 逐层贪婪方法学习 SAEN 权重参数的过程如下：

(1)由输入层到第 1 个隐藏层，通过最小化输入输出的重构误差，利用反向传播算法训练参数，得到输入数据的第 1 个潜在表示(即第 1 隐藏层).

(2)将上一层特征向量作为训练下一层的输入，采用同样的方法训练权重参数，得到数据的另一个潜在表示(即第 2 隐藏层)，依次类推.

也就是说，第 i 隐藏层的特征作为训练第 $i+1$ 层的输入，逐层贪婪的学习过程是把 SAEN 网络进行分层，对每一层进行无监督学习，又称为预训练过程. 图 8 描述了基于逐层贪婪学习的训练过程.

输入：稀疏训练样本集合 $\boldsymbol{X}=\{\boldsymbol{x}^{(i)},1\leq i\leq M,\boldsymbol{x}^{(i)}\in\boldsymbol{R}^{D(I)}\}$，隐藏层数 k

输出：网络连接权重参数矩阵集合$\{\boldsymbol{W}_h,1\leq h\leq k\}$，偏置向量集合$\{\boldsymbol{b}_h,1\leq h\leq k\}$第 k 隐藏层输出结果 $\boldsymbol{Y}^{(k)}$

1. 初始化：$\boldsymbol{Y}^{(0)}=\boldsymbol{X}$；
2. 用原始训练样本作为输入，通过图 5 训练出第 1 个隐层结构的网络参数 W_1，利用式(12)，将训练好的参数算法第 1 个隐层的输出，得到 $\boldsymbol{Y}^{(1)}$；
3. 把上一步的输出 $\boldsymbol{Y}^{(1)}$ 作为下一个自编码的输入，同样用图 5 训练下一个隐层网络的参数 W_2，并根据式(12)计算下一个隐层的输出，得到 $\boldsymbol{Y}^{(2)}$；
4. 重复步 3，直至第 k 个隐层，并根据式(12)计算出第 k 个隐层的输出 $\boldsymbol{Y}^{(k)}$；
5. 返回 SAEN 网络连接权重矩阵集合，偏置向量集合以及第 k 层的隐层 $\boldsymbol{Y}^{(k)}$.

图 8 基于逐层贪婪学习的训练算法

上述算法复杂性与自编码器的网络结构密切相关，假设 k 为网络隐藏层的数目，则本文采用的逐层贪婪训练学习算法的复杂性可以表示为 $O(ins^k\cdot hids)$，其中 ins 和 $hids$ 分别表示输入层单元的个数和隐藏层单元的个数，k 表示网络隐藏层数.

3.5 点击率预估模型

上一小节中，通过 SAEN 算法学习得到单特征之间的关联特征，新特征有更强的表达能力，对原始数据有着更本质的刻画. 因此本小节使用新特征作为点击预估模型的训练对象.

点击率预估问题实质上是一个基于概率的二分类问题，本文使用逻辑回归作为点击预估模型，逻辑回归(Logistic Regression，LR)模型[28]是机器学习中十分常用的一种广义线性分类模型，通常用来描述事件发生的概率，在互联网领域得到广泛应用. 除了广告系统的 CTR 预估外，推荐系统中预估转化率、反垃圾系统中垃圾识别等也广泛使用该模型. LR 预估模型训练过程示意图，如图 7 中黑色框部分所示.

定义 2. 正负样本.

第 i 个样本用$(\boldsymbol{X}^{(i)},y^{(i)})$表示，其中 $\boldsymbol{X}^{(i)}$ 为第 i 个样本的组合特征向量描述，$y^{(i)}$ 表示该样本是否点击，取值为 0 或 1，$y^{(i)}$ 值为 1 表示该则广告被点击，$y^{(i)}$ 的值为 0 表示该则广告未被点击. 本文把 $y^{(i)}$ 取值为 1 的样本定义为正样本，$y^{(i)}$ 取值为 0 的样本定义为负样本.

给定 M 条训练样本$\{(\boldsymbol{X}^{(i)},y^{(i)}),i=1,2,\cdots,M\}$，$N$ 表示特征数. 对于第 i 个样本中对应的广告，用户点击广告的概率表示为

$$P(y^{(i)}=1|\boldsymbol{\theta},\boldsymbol{X}^{(i)})=\frac{1}{1+\exp(-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{(i)})} \quad (15)$$

其中 $\boldsymbol{\theta}$ 是要求解的参数. 等式右侧的函数称为 Sigmoid 函数，其值域在(0,1)之间. 点击变量 $y^{(i)}$ 服从二项分布，拟合二项分布通常用极大似然估计法，假设样本之间独立，所以它们的联合分布可表示为各边缘分布的乘积，即

$$\max_{\boldsymbol{\theta}}P(\boldsymbol{\theta})=\prod_{i=1}^{M}P(y^{(i)}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}) \quad (16)$$

通过对式(16)进行极大似然估计，即可求得参数. 本文使用 L-BFGS优化算法[29]求解参数，该方法收敛更快且节省内存. 该算法的复杂性为$O(iters\cdot L^3)$，其中 $iters$ 表示迭代次数，$L=|\boldsymbol{Y}^{(k)}|$，$\boldsymbol{Y}^{(k)}$ 为 SAEN 算法最后第 k 隐藏层输出的结果向量，即学习得到的新特征.

3.6 算法复杂性分析

本文算法的整体复杂性为上述各个环节的累加，即聚类、张量分解、深度学习和点击率预估，最后算法的复杂性可以如下描述：

$O(KTN)+O(K_uK_qK_a)+O(ins^k\cdot hids)+O(iters\cdot L^3)$，其中 K 表示聚类的个数，$T=\max\{T_q,T_a\}$，T_q和 T_a 分别表示对查询和广告调用聚类算法完成聚类任务而需要执行的迭代次数，$N=\max\{N_u,N_q,N_a\}$；K_u，K_q和 K_a 分别为张量 $\boldsymbol{H}$ 的 3 个维度数；ins 和 $hids$ 分别表示输入层单元的个数和隐藏层单元个数，k 表示网络隐藏层数；$iters$ 表示迭代次数，$L=|\boldsymbol{Y}^{(k)}|$，$\boldsymbol{Y}^{(k)}$ 为 SAEN 算法最后第 k 隐藏层输出的结果向量.

由于实际中 T 的取值远小于 N，且 $K\leq N$；K_u，K_q，$K_a<N$，$L\leq ins=O(K_u+K_q+K_a)$，深度学习网络确定后，可将 $hids$，$iters$ 看作常数，令 $P=\max\{k,3\}$，则算法的复杂性可以简化为 $O(N^P)$，其中 $P\geq 3$.

4 实验与结果分析

实验的目的是想通过实验验证本文提出的稀疏特征学习方法，并进一步验证与已有方法相比本文方法在模型运行时间上是否高效，预估效果是否更有效，以及观察不同数据规模对点击率预估结果的影响趋势.

4.1 实验环境

硬件环境：中科曙光服务器 1 台，Intel(R) Xeon (R) E5-2670@2.60GHz 32 核 CPU，32GB 内存.

软件环境：CentOS 6.2 操作系统，Python 2.7.6

开发环境以及 pyTensor，scikit-tensor，scikit-learn，theano 相关工具包.

4.2 实验数据

本文的实验数据来自 SIGKDD Cup2012 track2[①]，由腾讯公司旗下的搜索引擎搜搜(2013 年 9 月并入搜狗)提供的广告点击日志数据. KDD2012 CUP track2 对应的研究问题是根据给出的真实点击数据信息，包括用户查询，返回广告信息，返回页信息等，来预测该广告的点击率.

比赛提供的训练数据集共有 149 639 105 条记录，9.87 GB 大小. 测试数据集中除了没有点击数和展示数之外，其他信息与训练集一致，共有 20 257 594 条记录，1.26 GB 大小. 数据集中一条记录表示用户的一次检索行为所展示的 k 条广告中一条广告包含的所有信息，又称为一个实例.

4.3 实验设计过程

(1)实验数据划分

本文经过对无效数据清洗，以及数据预处理后，从候选数据集中随机抽取 330 万条样本用于实验，最终实验所用数据统计，如表 1 所示.

表 1　实验所用数据统计表

	样本数/万	用户数	查询数	广告数	占比/%
总数据	330	117 264	142 943	26 428	100.0
训练数据	300				90.9
测试数据	30				9.1

实验过程中，本文通过划分训练数据分别在 7 个不同规模的数据集上训练模型，并且在同一测试集上验证不同方法的预估性能. 训练数据集划分情况，如表 1 所示. 以上训练和测试数据的划分以及训练数据的分组均采用随机划分. 7 种不同规模数据集的每组样本数分别为 15 万、20 万、30 万、50 万、60 万、75 万和 100 万. 每种规模的数据都至少选取 3 组以上完全不同的数据，最终的结果取所有组实验结果的平均值，以此来保证实验结果的可靠性.

(2)对比方法

在进行对比实验时，为了更有针对性和公平性，我们按照如下标准选择对比方法：① 最新的点击率预估模型；② 针对广告数据的点击率预估(在广告数据上有过实验的工作)；③ 单一点击率预估模型，而非综合多种模型的组合式模型.

由于本文工作的核心是预估模型，因此第 1 个原则保证我们与最新的预估模型进行对比. 由于点击率预估方面的工作目前主要分为两类，Web 搜索文档和广告，第 2 个原则是为了选择真正在广告数据上有过测试的模型；第 3 个原则是要求选择那些单一模型进行对比，可以更好的分析不同模型之间的异同，当然本文的方法可以很容易的集成到文献[11]的框架中.

本文预估模型最大特点有两个方面：① 降维方法在保证降低维数的前提下，并没有丢失关键信息；② 在特征关联的深度学习方面，通过自动学习挖掘关键特征和关联. 其属于基于特征学习的点击率预估方法. 因此根据上述原则和本文方法的特点，选取了与本文方法相近的两个代表性工作进行对比.

文献[3]与本文提出的 ADoSFLM 方法都使用 LR 作为预估模型，从数据中抽取与点击率预估相关的特征，是通过人工设定特征提取方案后得到的，如通过数据中的信息得到广告质量特征，相似广告 CTR 值等. 而本文提出的方法是利用深度学习算法，通过挖掘特征之间的关联自动获得特征.

文献[12]使用矩阵分解获得查询-广告内在关联以及通过张量分解获得用户-查询-广告之间的关联信息，然后集合广告位置和广告与查询相关性信息，利用贝叶斯网络建模，得到的概率模型用 EM 算法求解结果. 该文使用张量分解法是为了获得用户个性化信息，最后得到的是基于个性化点击率预估模型. 通过对用户行为建模，考虑到了用户信息，通过对比实验，验证该方法好于其他已有的概率图模型方法. 然而，与本文方法相比，两者最大的差异是对信息的利用方式不同. 本文最大特点是使用深度学习算法挖掘特征之间的内在关联，获得的是组合特征集合，而非简单的对用户行为进行建模.

通过与文献[3,12]中的方法进行比较，可以验证本文通过深度学习获得的特征与人工方式获得特征相比是否更高效. 同时，也可以验证经过张量分解后得到的低维核心张量是否仍然具有足够的信息量用于深度学习点击率预估的关键特征. 表 2 简要的介绍了本文进行对比的几种方法.

表 2　对比方法介绍

方法	介绍
Human_LR	人工抽取特征＋LR 预估模型
HPCM	矩阵分解和张量分解的混合点击预估模型＋EM 算法训练模型
ADoSFLM	基于 K-means 聚类的张量分解＋深度学习 SAEN 算法＋LR 模型

(3)实验评估方法

本文采用 ROC 曲线下面积 AUC[30](Area Under

① SIGKDD12 Cup. http://www.kddcup2012.org/c/kddcup2012-track2

the Receiver Operating Characteristic)作为模型预估性能评价标准. ROC 曲线基于混淆矩阵,它主要用于比较预估结果和真实结果. 矩阵中的行表示实例的预测结果,列表示实例的真实结果.

在 ROC 空间中,每个点的纵坐标是真正率,横坐标是假正率,描绘了分类器在 TP 和 FP 之间的权衡. 对于本文的二值分类问题来说,实例值往往是连续值,需要设定一个阈值,将实例分类到正类或负类. 曲线下面积(AUC)就是 ROC 曲线下方的那部分面积大小,该值通常介于[0.5,1)之间,AUC 越大,代表点击预估模型性能越好[30].

4.4 实验结果分析

主要包括 3 个部分:(1)参数对模型预估效果的影响;(2)模型运行时间比较;(3)预估效果的比较与分析.

4.4.1 参数对模型预估效果的影响

本文提出的 ADoSFLM 方法涉及到很多参数,最关键的参数如深度学习阶段中网络层数 k 和模型训练阶段中迭代次数 $iters$. 这些参数对于模型最后的预估结果产生直接的影响,因此本文首先针对参数进行了实验,以选出最佳的参数组合.

本文使用数据规模为 500 000 样本的一组抽样数据训练模型,在测试集上测试,用于选取最佳参数. 固定 SAEN 算法的网络层数($k=2,3,4,5,6$)时,分析不同 $iters$ 对模型性能的影响,所得结果如表 3 所示.

表 3 不同网络层数与 LR 模型训练迭代次数之间的关系

迭代次数	平均 AUC 值				
	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
$iters=10$	0.6634	0.6577	0.6684	0.6644	0.6639
$iters=20$	0.6757	0.6591	0.6842	0.6731	0.6693
$iters=30$	0.6793	0.6685	0.6954	0.6739	0.6802
$iters=50$	0.6989	0.6932	0.7097	0.6983	0.6937
$iters=70$	0.7169	0.7149	0.7224	0.7121	0.7139
$iters=90$	0.7246	0.7265	0.7347	0.7238	0.7354
$iters=100$	0.7283	0.7294	0.7384	0.7377	0.7335
$iters=110$	0.7305	0.7308	0.7409	0.7394	0.7390
$iters=130$	0.7311	0.7321	0.7408	0.7411	0.7368
$iters=150$	0.7324	0.7320	0.7404	0.7385	0.7382

根据表 3 生成曲线图,如图 9 所示,图 9 反映了不同的网络隐层数 k 值与 LR 模型迭代次数 $iters$ 所对应的 AUC 变化情况,从图 9 中不难看出,当迭代次数为 100~120 次时,几条曲线的 AUC 值已经变化不大,趋于稳定. 因此在下面的对比实验中,本文选择 110 作为训练预估模型的迭代数.

值得注意的是,SAEN 算法的复杂度与参数 k 直接相关,k 表示隐层数,其值越大,SAEN 算法的计算开销越大. 通过图 9 可知,达到收敛时,$k=2,3,4,5,6$ 对应的 AUC 值分别是 0.7324, 0.732, 0.7404, 0.7385 和 0.7382. 从曲线上看,随着迭代次数的变化,各条曲线波动较大,其中 $k=4$ 时相对稳定,并且当迭代数达到一定值时,AUC 值趋于收敛. 因此,就本实验所用的数据集、数据规模以及计算开销而言,综合考虑 $k=4$(即隐层数为 4)是一个合理的选择.

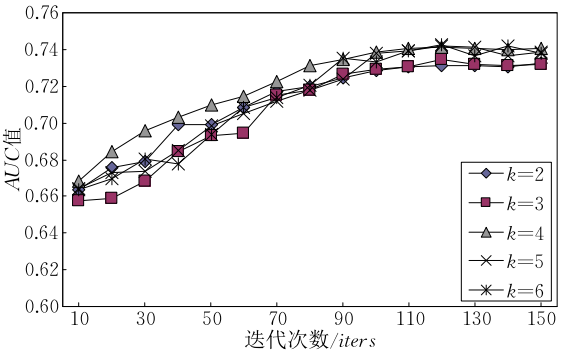

图 9 LR 模型的 AUC 值随不同迭代次数变化情况

4.4.2 模型运行时间比较

本文记录了 3 种方法在不同数据规模下的运行时间,表 4 给出了分别在 15 万、30 万、50 万、75 万和 100 万数据规模下的平均运行时间.

表 4 不同数据规模下模型平均时间

平均时间/s	数据规模(样本数/万)				
对比方法	15	30	50	75	100
Human_LR	87	182	298	461	578
HPCM	2624	4238	6140	8924	12624
ADoSFLM	1975	3750	5637	8347	11327

从表 4 可知,3 种方法在模型运行时间上相差很大. Human_LR 运行时间只有预估模型训练阶段,特征抽取过程依赖领域知识,特征数相对较少,模型运行时间相对也较小. 而 HPCM 和 ADoSFLM 模型涉及到张量分解的复杂运算以及 EM 算法和深度学习阶段算法的计算开销也很大,并且涉及到的特征数较多. 因此,HPCM 和 ADoSFLM 方法运行时间相对较长. 虽然本文提出的 ADoSFLM 方法在运行时间方面并没有特别优势. 但是,最耗时的广告点击率预估模型是基于海量数据训练得到的,其计算过程是在离线环境下进行的. 因此,运行时间对在线预估 CTR 值并没有影响.

4.4.3 预估效果的比较与分析

本文分别在 7 个不同规模的数据集上训练模型,在同一测试集上评估预估效果. 分别考察不同方法之间的预估效果,以及数据规模对预估质量的影

响. 表 5 描述了不同方法在不同数据规模下的预估结果.

表 5　3 种方法在不同数据规模下的平均预估效果

数据规模/万	平均 *AUC* 值		
	Human_LR	HPCM	ADoSFLM
15	0.6674	0.6651	0.6696
20	0.6740	0.6779	0.6827
30	0.6866	0.6927	0.7115
50	0.6931	0.7116	0.7303
60	0.6967	0.7195	0.7429
75	0.7003	0.7235	0.7556
100	0.7015	0.7281	0.7601

从表 5 可以看到 3 种方法在不同规模数据集下的预估结果，ADoSFLM 模型相对于 Human_LR 的预估效果随着数据规模的加大提升效果越发明显，以数据规模 100 万为例，预估性能提升了 8.35%. ADoSFLM 之所以预估性能更好，主要有 3 个原因：(1) 原始数据中存在很多长尾数据和无效数据，本文经过数据清洗后，通过对原始数据进行 *K*-means 聚类和张量分解两次降维，降低了数据中的噪声对真实信息的干扰，同时也降低了数据的稀疏性，这对于提高预估准确率是有利的；(2) ADoSFLM 方法采用了深度 SAEN 算法学习特征之间的高度非线性关联，多层结构的特征学习模型对于挖掘特征之间深层次的规律是有效的；(3) 实验过程中，利用 ADoSFLM 方法学到的特征数大于 Human_LR 方法中人工抽取的特征，ADoSFLM 方法学到的特征对数据有更强的表达能力. 通过人工方式提取出对点击率预估高度关联的特征越来困难，而 SAEN 算法通过对广告数据进行深度学习，得到更多的对点击率预估有意义的特征，这对于提高模型预估准确率是有帮助的.

从表 5 还可知，采用 ADoSFLM 方法的点击率预估效果也要好于 HPCM. 主要原因有：就模型而言，两种方法都使用张量分解，获得近似张量用于点击预估模型，ADoSFLM 方法使用 LR 模型预估点击率，后者使用概率模型预估点击率，其中最大的区别是本文提出的方法含有深度特征学习的阶段. ADoSFLM 中的深度学习方法，充分学习了张量分解后的近似张量和数据中其他属性特征之间的非线性关系，SAEN 算法刻画特征之间的高度非线性关联，最大限度的挖掘隐藏在特征之间的规律，这是 HPCM 模型中没有的. 因此，从预估质量上来看，本文使用深度结构模型学习特征之间的高度非线性关联，对于提高点击率预估的准确性是可行的，预估质量也有所提升.

为了能清晰的看出 3 种方法的预估质量在不同规模数据下的变化趋势，本文根据表 7 得到相应的平均预估 *AUC* 值折线图，如图 10 所示.

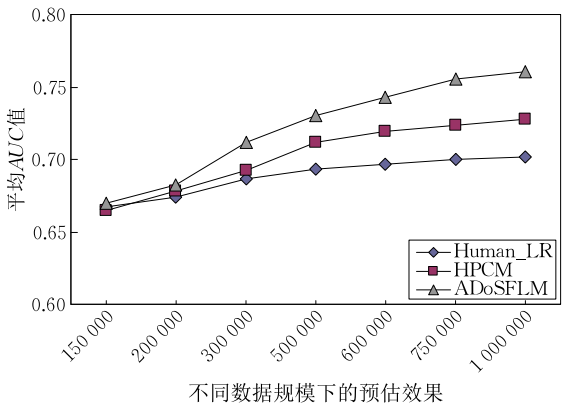

图 10　不同数据规模下 3 种方法的预估效果比较

从图 10 可以看出，3 种方法随着数据规模（即训练样本数）的增加，在同一测试集上的预估质量都有所提升. 这主要是因为，最初的训练样本数比较少，对于 3 种方法来说，模型都处于过拟合状态，随着数据规模的增加，训练得到的参数稳定性和健壮性都有所增强. 值得注意的是，随着数据规模的增长，3 条曲线的变化趋势并不相同，即模型预估效果的收敛程度不同. 开始时，3 种方法预估结果基本无差异，随着数据规模的增长，本文提出的 ADoSFLM 方法与 Human_LR 和 HPCM 方法相比，预估效果增长明显加快. 在相同数据规模下，增长幅度也大于其他两个方法.

Human_LR 方法在数据规模为 600 000 时，已基本趋于收敛，这说明人工抽取特征方法，在较小规模数据集上表现良好. 当数据规模较小时，数据中隐藏的结构和规律不明显，人工抽取的特征可以很好的表示数据. HPCM 方法中用到的张量分解，一方面是为了获得数据的降维、降噪表示，另一方面也是为了挖掘广告日志数据中隐藏的用户个性化信息，从图 10 可以看出，随着数据规模的增加，预估效果也在提升.

本文提出的 ADoSFLM 方法，随着数据规模的增长，预估结果好于 HPCM. 这主要是因为 ADoSFLM 方法中包括特征学习阶段，借助 SAEN 算法的多层网络结构，学习特征之间的高度非线性关联关系. 当数据规模很小时，数据无法体现特征之间复杂的内在关联，此时 SAEN 也不能很好的刻画这种关联. 随着数据规模的增加，SAEN 算法可以更好的学习

数据中隐藏的规律,也就是说,本文使用深度学习算法来学习广告数据单特征之间的组合特征(即特征之间的关联).从理论上来说,训练数据越多,该算法学到的特征对数据的表达能力就越强,广告 CTR 预估的准确率也就越高.虽然受到当前实验所用硬件设备的限制,实验过程中没有使用更大规模的数据来训练模型,但是上述结果也证明本文提出的方法比其他方法更适合于大数据量情况下的应用.

5 结论与展望

本文基于最基本的搜索广告点击数据,从特征学习的角度,提出了面向广告数据的稀疏特征学习方法.考虑到广告数据的高维和高稀疏性特点,本文利用降维的方法,首先基于相似度分别对相似的广告、查询和用户进行聚类,使数据具有初始聚合性;其次对降维后的三元组建立三维张量模型,利用 Tucker 分解获得低阶近似张量.针对特征之间存在的高度非线性关联关系,文中研究了基于深度学习的特征学习方法,首先分析了输入层特征构成,结合栈式自编码网络算法学习特征之间的高阶组合特征,作为点击预估模型的训练对象.实际中如果还可以获得其他类型的数据,如不同用户群体的偏好,不同类型广告的投放规律等,则可以结合使用不同类型的方法,以此获得更高的预估准确率,因为不同类型的方法可以覆盖问题的不同方面.

尽管本研究从特征学习的角度研究点击率预估问题取得了较好的效果,但仍存在一些不足之处.因此,在后续的工作中,可在本文基础上,从以下几个角度进一步研究:(1)文中基于深度学习算法的时间复杂度比较高,计算开销较大,该算法复杂度与每一层的单元数和网络层数有关,同时也与连接函数有关.在保证特征学习效果的前提下,如何改进连接函数、调节每一层单元数和网络层数,以降低时间开销的问题有待解决;(2)文中提出的特征学习方法预估广告点击率,仅考虑了展示充分的广告数据,未考虑展示不充分的广告(稀疏广告).接下来的工作中,如何从特征学习的角度,预估稀疏广告的点击率,是一个值得研究的问题,也是实际中目前亟需解决的一个问题;(3)不同预估模型之间的深入分析与对比,开展不同模型融合方面的机理性研究.

参 考 文 献

[1] Broder A Z. Computational advertising//Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA. San Francisco, USA, 2008: 992

[2] Zhou Ao-Ying, Zhou Min-Qi, Gong Xue-Qing. Computational advertising: A data-centric comprehensive Web application. Chinese Journal of Computers, 2011, 34(10): 1805-1819(in Chinese)
(周傲英, 周敏奇, 宫学庆. 计算广告: 以数据为核心的文本应用. 计算机学报, 2011, 34(10): 1805-1819)

[3] Richardson M, Dominowska E, Ragno R. Predicting clicks: Estimating the click-through rate for new ads//Proceedings of the 16th International Conference on World Wide Web. Banff, Canada, 2007: 521-530

[4] Srikant R, Basu S, Wang N, et al. User browsing models: Relevance versus examination//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Indianapolis, USA, 2010: 223-232

[5] Hillard D, Schroedl S, Manavoglu E, et al. Improving ad relevance in sponsored search//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 361-370

[6] Dupret G, Liao C. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 181-190

[7] Dupret G E, Piwowarski B. A user browsing model to predict search engine click data from past observations//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008: 331-338

[8] Guo F, Liu C, Kannan A, et al. Click chain model in web search//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 11-20

[9] Chapelle O, Zhang Y. A dynamic Bayesian network click model for web search ranking//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 1-10

[10] Zhang W V, Jones R. Comparing click logs and editorial labels for training query rewriting//Proceedings of the WWW 2007 Workshop on Query Log Analysis: Social and Technological Challenges. Banff, Canada, 2007

[11] Jahrer M, Toscher A, Lee J Y, et al. Ensemble of collaborative filtering and feature engineered models for click through rate prediction//Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDDCup Workshop. Beijing, China, 2012

[12] Shen S, Hu B, Chen W, et al. Personalized click model through collaborative filtering//Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, USA, 2012: 323-332

[13] Hu B, Zhang Y, Chen W, et al. Characterizing search intent diversity into click models//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 17-26

[14] Kempe D, Mahdian M. A cascade model for externalities in sponsored search//Proceedings of the 4th International Workshop on Internet and Network Economics. Chicago, USA, 2008: 585-596

[15] Guo F, Liu C, Wang Y M. Efficient multiple-click models in web search//Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. Barcelona, Spain, 2009: 124-131

[16] Regelson M, Fain D. Predicting click-through rate using keyword clusters//Proceedings of the Second Workshop on Sponsored Search Auctions. Ann Arbor, USA, 2006: 9623-9628

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

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

[19] Wang X, Li W, Cui Y, et al. Click-through rate estimation for rare events in online advertising//Hua Xian-Sheng, Mei Tao, Hanjalic A eds. Online Multimedia Advertising: Techniques and Technologies. Hershey Pennsylvania, USA: IGI Global, 2010: 1-12

[20] Hartigan J A, Wong M A. A *K*-means clustering algorithm. Applied Statistics, 1979, 28(1): 100-108

[21] Cichocki A, Zdunek R, Phan A H, et al. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. New Jersey, USA: John Wiley & Sons, 2009

[22] Oseledets I V, Savostyanov D V, Tyrtyshnikov E E. Linear algebra for tensor problems. Computing, 2009, 85(3): 169-188

[23] Kolda T G, Sun J. Scalable tensor decompositions for multi-aspect data mining//Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08). Pisa, Italy, 2008: 363-372

[24] Ng A Y. On feature selection: Learning with exponentially many irrelevant features as training examples//Proceedings of the 15th International Conference on Machine Learning. Madison Wisconsin, USA, 1998: 404-412

[25] Rifai S, Mesnil G, Vincent P, et al. Higher order contractive auto-encoder//Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part II. Athens, Greece, 2011: 645-660

[26] Hinton G E, Zemel R S. Autoencoders, minimum description length, and Helmholtz free energy//Cowan J D, Tesauro G, Alspector J eds. Advances in Neural Information Processing Systems 6. San Francisco, USA: Morgan Kaufmann, 1994: 3-10

[27] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527-1554

[28] Hosmer Jr D W, Lemeshow S. Applied Logistic Regression. 3rd Edition. New Jersey, USA: John Wiley & Sons, 2013

[29] Liu D C, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 1989, 45(3): 503-528

[30] Lobo J M, Jiménez-Valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 2008, 17(2): 145-151

ZHANG Zhi-Qiang, born in 1973, Ph. D., professor. His main research interests include information retrieval, database, and intelligent information processing.

ZHOU Yong, born in 1988, M. S. His current research interests include information retrieval and data mining.

XIE Xiao-Qin, born in 1973, Ph. D., associate professor. Her current research interests include service computing, social network, intelligent information processing.

PAN Hai-Wei, born in 1974, Ph. D., associate professor. His current research interests include data mining, intelligent information processing.

Background

Click through rate estimation in sponsored search is the task of predicting probabilities that user click an ad given ⟨query, ad⟩ and context information. The existing works on this issue could be classified into two categories: first, statistical learning model which is characterized by obtaining highly correlated with CTR model to improve the precision of estimating. Designing the feature extraction scheme is a key part of program, such as extracting relevant features between advertisements in the same result page or constructing combined features and so on. Second, user behavior model which is based on Probabilistic Graphical Models, through hypothesis testing, takes advantage of Bayesian network to portraying user's browsing scenes, and then estimates the probability that the user clicks an ad. Human feature engineering

constructs combined features, but with low efficiency, low scalability and other defects. Bayesian model characterizes user browsing behavior, but it is not able to use information sufficiently, and don't take into account the sparseness of advertising data and highly nonlinear association between features.

Against these issues, this paper considers the characteristics of advertising data and proposes advertising data-oriented sparse feature learning method from the perspective of learning characteristics to estimate CTR. This method combines the advantages of tensor dimensionality reduction and feature learning to solve high-dimensional sparse feature problem of advertising data. Firstly, based-on the relationship between internal objects of the same type, dimension reduction using clustering makes an initial aggregation of data. As for correlations between different types of objects, we use tensor decomposition to reduce the dimensionality, while protecting the data associated with the original structure of ad clicks. Secondly, a stacked auto encoder algorithms in deep learning fields has been studied to mine high-order relationship between advertising data features, and obtain new abstract features, which has more representation capability to ad data and help to improve the prediction accuracy of the CTR. Thirdly, the new features will be acquired as the input for click prediction model and use the L-BFGS algorithm to learn parameters.

This paper is supported by the National Natural Science Foundation of China under Grant Nos. 61370084, 61202090, 61272184, the Program for New Century Excellent Talents in University No. NCET-11-0829, the Natural Science Foundation of Heilongjiang Province under Grant No. F201130 and the Fundamental Research Funds for the Central Universities under Grant Nos. HEUCF100609, HEUCFT1202.