**文章编号**:1007-5321(2021)05-0081-07 **DOI**:10.13190/j.jbupt.2021-010

# 基于图嵌入和 CaGBDT 的多模态出行推荐

孙全明, 常 磊, 马 铖, 曲志坚
(山东理工大学 计算机科学与技术学院, 淄博 255049)

**摘要**: 针对交通出行服务中推荐方式单一、容易忽略用户出行偏好等问题,借鉴多粒度级联森林结构,提出了一种级联梯度提升树模型(CaGBDT). 该模型利用级联结构增加模型的深度,进而实现了特征的深层次表示学习. 同时,为了解决样本类别不平衡问题,提出了一种基于鲍威尔算法的指标优化层,其通过为每个类别搜索一个阈值,对模型的预测结果进行权重修正,以实现最大化评价指标的目的. 此外,CaGBDT 模型可以根据用户的出行记录,构建用户出行全局关系图,利用图嵌入表示学习方法,自动提取用户出行的空间上下文关系,从而提高特征提取的效率.

**关 键 词**: 交通出行推荐; 图嵌入; 特征工程; 级联森林; 梯度提升决策树
**中图分类号**: TP391 **文献标志码**: A

# Multi-Modal Transportation Recommendation Based on Graph Embedding and CaGBDT

SUN Quan-ming, CHANG Lei, MA Cheng, QU Zhi-jian
(School of Computer Science and Technology, Shandong University of Technology, Zibo 255049, China)

**Abstract**: To solve the problems in transportation recommendation service, such as single recommended methods and ignoring the user travel preferences, a cascade gradient boosting decision tree (CaGBDT) model is proposed based on the multi-grained cascade forest structure. CaGBDT uses the cascade structure to increase the depth, and then realizes the deep-level representation learning of features. Meanwhile to solve the imbalance of sample class, an index optimization layer based on Powell algorithm is proposed. By searching a threshold for each class, the weight of the prediction results of the model is modified to maximize the evaluation indexes. In addition, a user travel global relationship graph can be constructed by the CaGBDT model via referring the user's travel record, and the spatial context relationship of the user's travel is extracted automatically by the graph embedding method to improve the efficiency of feature extraction.

**Key words**: traffic recommendation; graph embedding; feature engineering; cascade forest; gradient boosting decision tree

随着交通事业的快速发展,可供选择的出行方式越来越多样化,人们对出行方案的要求也越来越高. 面对复杂的交通系统,如何在合适的时间、合适的地点,推荐符合用户偏好的出行方案,成为业界面

**收稿日期**: 2021-01-19
**基金项目**: 山东省自然科学基金项目(ZR2017LF004); 山东省高等学校优秀青年创新团队支持计划项目(2019KJN048)
**作者简介**: 孙全明(1995—), 男, 硕士生.
**通信作者**: 曲志坚(1980—), 男, 副教授, 硕士生导师, E-mail: zhijianqu@sdut.edu.cn.

临的难题之一. Herzog 等[1]将协同过滤与基于知识的推荐方法相结合,规划最适合用户的单一模式出行方案. Du 等[2]以出行的最低成本作为优化目标,利用广度优先搜索算法规划成本最低的公共交通出行方案. Socharoentum 等[3]综合考虑用户出行目的、用户所在位置以及周围环境等多个因素,根据用户的历史行为挖掘用户出行偏好,实现在不同的情境中规划最合适的步行方案. 上述方法都取得了较好的推荐效果,但是推荐的出行方式都是单一模式. 由于现代城市交通系统的复杂性和不确定性,单一的交通规划和推荐策略很难满足用户的出行需求,并且在很大程度上忽略了用户的出行偏好,无法提供令人满意的用户体验.

多模态出行推荐旨在根据用户的出发地和目的地,向用户推荐一种单式或多式联合运输的出行方案. Liu 等[4]提出了一种基于多式联运图(MMTG, multi-modal transportation graph)的交通推荐表示学习框架,该框架从大规模地图查询数据中提取多式联运图,通过学习 5 种出行方式在联运图中的嵌入向量,实现多式联运推荐. 上述方法综合考虑了用户的出行偏好和当前的路况信息,取得了较好的多式联运推荐效果,但是推荐的出行方式较少,且选择的用户只在单一城市范围内出行,对于在不同城市出行的用户推荐效果不佳.

因此,分析用户在不同城市的多模态出行模式并为用户提供个性化的多模态出行推荐方案是一个值得研究的热点问题. 所提方法通过人工设计特征和图嵌入[5]的方法,提取用户出行潜在的时空信息,借鉴多粒度级联森林(gcForest, multi-grained cascade forest)[6]的结构,构建了级联梯度提升决策树(CaGBDT, cascade gradient boosting decision tree)预测模型. 通过给定用户信息、源-目的地(OD, origin-destination)对信息,以及情景上下文信息,从 11 种模态出行方式中,为用户推荐一种最合适的出行方案.

# 1 数据描述与分析

## 1.1 数据描述

分析数据采用百度地图提供的 KDD Cup 2019 数据集,该数据集包括北京、上海、广州、深圳 4 个城市从 2018 年 10 月 1 日至 2018 年 11 月 30 日的查询记录、展示记录、点击记录以及用户属性数据. 数据集统计信息如表 1 所示.

**表 1 数据集的统计信息**

| 数据描述 | 北京 | 上海 | 广州 | 深圳 |
|---|---|---|---|---|
| 查询记录 | 500 000 | 500 000 | 500 000 | 500 000 |
| 展示记录 | 491 054 | 490 917 | 486 200 | 470 401 |
| 点击记录 | 453 336 | 449 604 | 434 232 | 338 011 |
| 用户属性 | 46 191 | 46 119 | 38 468 | 41 579 |

1)一条查询记录代表用户在地图软件中的一次路线搜索. 每条记录包括用户编号、查询时间、出发地的经纬度和目的地的经纬度.

2)一条展示记录代表用户查询之后,地图软件展示给用户的所有可行路线. 每条记录包括用户编号、展示时间和路线列表. 路线列表中的每条路线由出行方式、路线距离、预计耗时和预计花费组成.

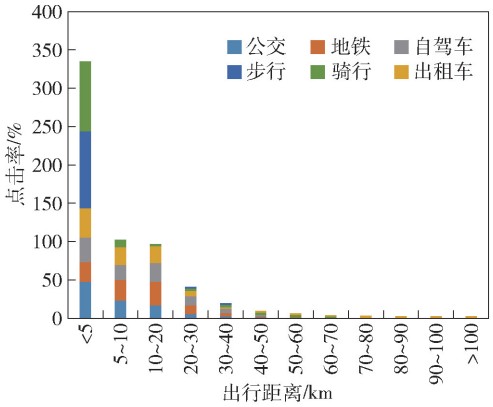
3)一条点击记录表示用户对展示记录的点击反馈,代表用户选择的具体出行方案. 每条记录包括点击时间和点击的出行方式. 由于数据集对出行方式进行了限定,只提供了 11 种(6 种单模态和 5 种多模态)出行方式的查询和点击记录,所以将路线推荐任务转化为点击预测任务,将预测的点击概率最高的出行方式作为首选推荐给用户. 另外,由表 1 可知,4 个城市都存在查询未点击的情况,且占比各不相同. 为了充分利用原始数据的信息,将查询未点击的数据标签置为 0,表示未发生点击,则所提方法不仅需要预测未来 7 天用户是否会有效点击,而且需要预测若发生点击,用户会点击哪一种出行方案.

4)出于隐私保护,不直接提供每个用户的人口统计学信息,而是通过聚类,将具有相同属性的用户合并为一个用户群体,并用一组多热编码的向量表示用户属性. 该方法既可以保护用户隐私,同时也可以有效缓解稀疏用户带来的冷启动问题. 但是这也导致约 43% 的用户群体在 4 个城市都有出行记录. 由于不同城市的交通设施以及出行规律各不相同,所以通过分析不同城市的多模态出行特点,挖掘用户在不同城市的出行规律,是实现用户在多城市多模态出行推荐服务的关键.

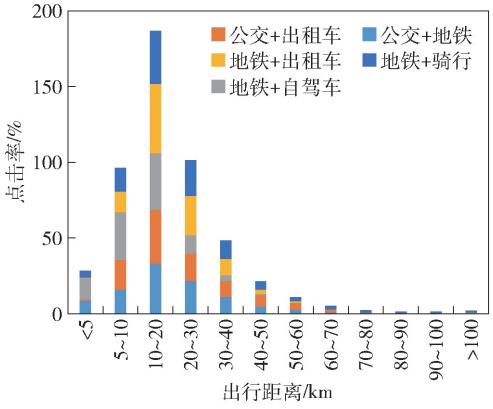
## 1.2 数据分析

首先,对 11 种出行方式的点击率与出行距离的关系进行分析,结果如图 1 所示. 由图 1 可知,步行和骑行的速度最慢,出行距离多集中在 5 km 以内;公交和地铁凭借价格和速度优势,成为 5 ~ 10 km 出行的主要选择;而自驾车和出租车虽然在速度上占

优，但是当面临高峰期时，最容易形成堵车、耽误行程，出行距离多集中在 10 km 以内. 当行程距离在 10 ~ 20 km时，多模态出行方案成为主要的出行选择，其中地铁 + 出租车的出行方案凭借速度优势，获得接近 50% 的点击率. 这些数据表明，单模态出行方案受到速度和价格的限制，更适合短距离的出行，而多模态出行方案的速度优势和价格优势在长距离出行中更加明显.




(a) 单模态出行方式的点击率与出行距离的关系




(b) 多模态出行方式的点击率与出行距离的关系

图 1　不同出行方式的点击率与出行距离的关系

其次，利用查询记录对用户的出行时间进行分析发现，用户在白天的活跃度明显高于晚上，节假日和周末明显高于工作日，而且每个城市的高峰期各不相同. 例如北京市的高峰多集中在 9 点、12 点以及 17 点，而广州市的高峰多集中在 11 点至 17 点. 根据用户的出行时间分布情况，可以通过提取周期特征或时间节点特征，如是否是节假日、是否是工作日等特征，表示用户出行的时间依赖关系.

最后，通过分析不同出行方式的点击率发现，超过 15% 的用户未发生有效点击，超过 50% 的用户会选择公交和地铁，而只有 6% 的用户选择自驾车和出租车，这些数据表明大多数用户更倾向于选择公共交通工具出行，而且数据集存在样本类别不平衡问题，因此，在设计推荐模型时，还需要充分考虑数据不均衡带来的负面影响.

## 2　特征提取与模型构建

### 2.1　特征提取方法

1) 路线特征.

用户出行会综合考虑路线的距离、耗时、花费等因素. 对于路线特征，首先从路线列表中提取路线的个数、最少花费路线、最少耗时路线、最短距离路线等特征. 然后提取每条路线所对应的出行方式、行程距离、行程时间、行程花费等属性特征. 最后，根据各路线的展示顺序，提取路线之间的排名特征，以及排名之间各属性特征的数学统计信息，例如均值、方差、极值等.

2) 时间特征.

由于用户出行规律与查询时间具有强相关性，所以时间特征主要依据用户查询时间进行特征构造. 这里采用对时间戳进行转换的方法，主要提取当前记录时刻的小时、周几、是否为节假日、是否为工作日等信息.

3) 空间特征.

空间特征主要围绕出发地和目的地的经纬度进行构造. 首先提取经纬度的查询次数，用来表示用户出行的热点地区. 其次，采用层次聚类算法，提取经纬度聚类特征，用来表示数据中的兴趣点信息. 由于各出行方式的点击率与出行距离具有强相关性，所以根据经纬度提取了两地之间的球面距离 $d_{od}$ 和方位信息 $b_{od}$，具体计算为

$$d_{od} = 2R \times \arcsin\left(\sqrt{\sin^2\left(\frac{\alpha_o - \alpha_d}{2}\right) + \cos\alpha_o \cos\alpha_d \sin^2\left(\frac{\beta_o - \beta_d}{2}\right)}\right) \tag{1}$$

$$b_{od} = \arctan\left(\frac{\sin(\beta_d - \beta_o)\cos\alpha_d}{\cos\alpha_o \sin\alpha_d - \sin\alpha_o \cos\alpha_d \cos(\beta_d - \beta_o)}\right) \tag{2}$$

其中：$R$ 为地球半径，$\alpha_o$ 和 $\alpha_d$ 分别为出发地和目的地的纬度，$\beta_o$ 和 $\beta_d$ 分别为出发地和目的地的经度.

此外，在推荐场景中数据对象之间呈现更多的是图结构，例如在出行数据中用户在不同 OD 对之间的出行，形成了用户与地点之间的全局关系图. 常规的经纬度统计特征对用户出行信息的表征能力

是有限的,为了增强特征的表达能力,设计了一种基于用户历史出行序列的编码方式,通过 node2vec[7] 图嵌入表示学习方法,学习用户与 OD 对之间的高阶协作关系. 研究发现,相似用户在相似的 OD 对之间出行,具有相似的兴趣偏好[8],基于此首先构建了用户 $u$ 与 OD 对的异构图,如图 2(a)所示. 异构图反映了不同用户在不同地点之间出行的空间关系,通过学习不同节点的嵌入向量,解决相似用户在不同城市之间出行的推荐难题. 其次,还构建了出发地与目的地的同构图,如图 2(b)所示. 同构图反映了出发地与目的地之间的潜在空间关系,用户出行的方向表示边的方向,节点的出现次数表示边的权重,通过学习节点之间的嵌入向量,解决经纬度特征表征能力弱的问题.

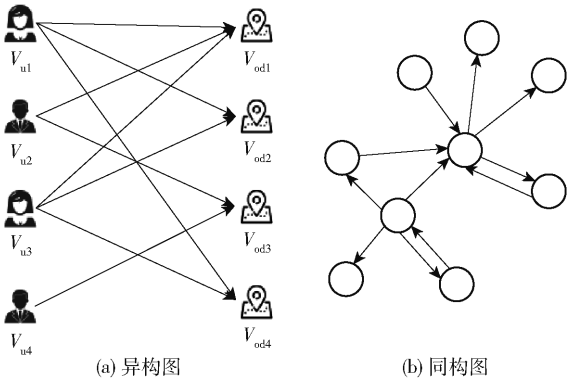



图 2 用户历史出行记录有向图

4)用户偏好特征.

根据用户的出行记录,构建用户与时空信息之间的交叉统计特征,例如用户编号与城市编号的共现次数,用户编号与时间的共现次数等,表示用户在时间和空间上的出行偏好. 为了减少人工设计特征的工作量,利用因子分解机(FM, factorization machine)[9]进行特征提取.

## 2.2 图嵌入表示学习

图嵌入表示学习旨在将大规模、高维度网络映射到低维空间,用低维稠密向量表示节点,尽可能保存原始网络结构及属性特征,常用于链接预测、节点分类、推荐系统等任务[10].

node2vec 是图嵌入代表算法之一,采用文本表示的思想,利用随机游走策略进行顶点采样,生成顶点的近邻序列,然后通过跳字模型[11]学习顶点表示. 与均匀随机游走策略[12]不同的是,node2vec 随机游走是有偏向性的,克服了均匀随机游走在检测 2 个不同网络社区时,既无法划分网络结构,也无法指导搜索方向的缺点. 为了同时满足同一个社区内的节点表示相互接近、不同社区内扮演相同角色的节点表示也要相互接近的优化目标,node2vec 引入跳转超参数 $p$ 和 $q$ 来控制随机游走策略. 假设当前随机游走经过边$(t,v)$到达顶点 $v$,设顶点 $v$ 到顶点 $x$ 的转移概率为 $\pi_{v,x}=\alpha_{p,q}(t,x)w_{v,x}$,其中 $w_{v,x}$为边的权重,

$$\alpha_{p,q}(t,x)=\begin{cases}\dfrac{1}{p}, & d_{t,x}=0\\ 1, & d_{t,x}=1\\ \dfrac{1}{q}, & d_{t,x}=2\end{cases} \tag{3}$$

其中 $d_{t,x}$为顶点 $t$ 到顶点 $x$ 之间的最短路径距离,则其顶点 $v$ 和顶点 $x$ 之间归一化的转移概率为

$$P(c_i=x\mid c_{i-1}=v)=\begin{cases}\dfrac{\pi_{v,x}}{z}, & (v,x)\in E\\ 0, & (v,x)\notin E\end{cases} \tag{4}$$

其中 $z$ 为归一化常数,$E$ 为图中边的集合. node2vec 通过半监督网络形式学习最优超参数 $p$ 和 $q$,使广度优先和深度优先达到最佳平衡,均衡网络的局部信息和全局信息.

## 2.3 CaGBDT 模型

gcForest 汲取集成学习的策略,利用级联结构,将多个随机森林模型集成,搭建强分类器. 同时,为了充分利用特征间的互补信息,感知上下文或结构关系,引入多粒度扫描算法,通过设置滑动窗口,按照步长进行特征扫描,提高特征的表示能力. 梯度提升决策树(GBDT, gradient boosting decision tree)是 Boosting 类算法之一,该算法通过拟合上一轮模型的残差,来提高模型整体的预测能力,在诸多领域和竞赛中表现都优于随机森林. 由于多粒度扫描算法会极大地增加特征维度,造成内存压力,故仅借鉴级联结构,将随机森林替换为轻量级梯度提升机(LightGBM, light gradient boosting machine)[13]和极限梯度提升(XGBoost, extreme gradient boosting)[14]模型,构建了 CaGBDT 模型,模型结构如图 3 所示.

为了增加层级模型之间的多样性,相同模型之间采用不同的超参数. 每个层级模型的输出为样本的类别概率,即长度为类别数的概率向量. 此外,由于用户出行数据存在样本类别不平衡问题,容易导致分类模型的训练出现偏差,造成对尾部类别分类准确率劣化的问题. 所以,在模型每一层的输出后面均添加一个指标优化层,通过搜索每个类别的最

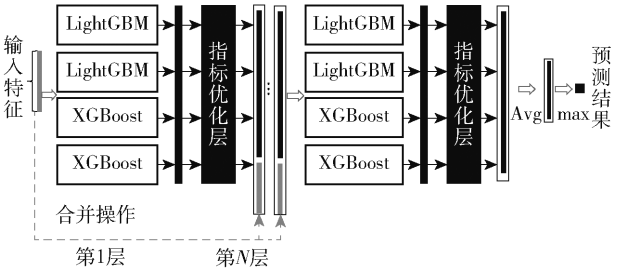



图 3　CaGBDT 模型结构图

优权重，以实现最大化评价指标的目的. 指标优化层的优化目标为

$$H=\sum_{i=1}^{k} w_i F_1(i) \tag{5}$$

其中：$F_1(i)$为模型预测的类别 $i$ 的 $F_1$ 值，$w_i$为类别权重.

对于式(5)所示的无约束优化问题，采用改进的鲍威尔优化方法搜索最优权重. 该方法无需对目标函数进行求导，大大降低了计算量，且避免了原始鲍威尔算法[15]在新一轮搜索方向和原始方向线性相关时不能收敛的缺陷.

改进的鲍威尔算法具体步骤如下：

1) 定义 12 个初始点 $\boldsymbol{X}^{(0)}=\{x^{(0,0)},x^{(0,1)},\cdots,x^{(0,11)}\}$以及 12 个线性无关的初始搜索方向 $d^{(1,0)},d^{(1,1)},\cdots,d^{(1,11)}$；定义允许误差 $\varepsilon>0$，令 $k>0$；

2) 进行基本搜索，令 $x^{(k,0)}=x^{(k-1)}$，以 $x^{(k,0)}$为起点，依次沿 $d^{(k,0)},d^{(k,1)},\cdots,d^{(k,11)}$进行一维搜索，得到点 $X^{(0)}$，求 $m$，使得 $f(x^{(k,m-1)})-f(x^{(k,m)})=\max\limits_{j=1,2,\cdots,12} f(x^{(k,j-1)}-f(x^{(k,j)}))$

3) 令加速度方向 $d^{(k,n+1)}=x^{(k,n)}-x^{(k,0)}$，若 $\|d^{(k,n+1)}\|\leqslant\varepsilon$，则停止迭代，得到 $X^{(k)}$为最优解；否则执行步骤 2；

4) 求解 $\lambda_{n+1}$，使得 $f(x^{(k,0)}+\lambda_{n+1}d^{(k,n+1)})=\min\limits_{\lambda} f(x^{(k,0)}+\lambda d^{(k,n+1)})$，令 $x^{(k+1,0)}=x^{(k)}=x^{(k,0)}+\lambda_{n+1}d^{(k,n+1)}$. 若 $\|x^{(k,n)}-x^{(k,0)}\|\leqslant\varepsilon$，则停止迭代，得到 $X^{(k)}$为最优解；否则执行步骤 3；

5) 调整搜索方向，若 $|\lambda_{n+1}|>\left[\dfrac{f(x^{(k,0)})-f(x^{(k+1,0)})}{f(x^{(k,m-1)})-f(x^{(k,m)})}\right]^{0.5}$，则令 $d^{(k+1,j)}=d^{(k,j)}$，$j=1,2,\cdots,m-1$，$d^{(k+1,j)}=d^{(k,j+1)}$，$j=m,m+1,\cdots,n$，$k=k+1$；否则令 $d^{(k+1,j)}=d^{(k,j)}$，$j=1,2,\cdots,n$，$k=k+1$；

6) 执行步骤 2.

指标优化层根据类别权重，适当放缩样本的类别概率，若一层中包含 $M$ 个 GBDT 模型，则该层的输出向量长度为 $M\times C_{\text{num}}$，其中，$C_{\text{num}}$为类别数量. 将该层的输出向量与原始的特征向量进行拼接，组成长度为 $K+M\times C_{\text{num}}$新特征向量，作为下一层的输入特征，其中 $K$ 为原始特征向量的长度. 若模型达到预设最大层数或模型第 $m$ 层的分类准确率不再升高，则该层的输出向量不再与原始特征向量进行拼接，直接对类别概率求均值，选择最大概率的类别作为预测结果.

## 3　实验与结果分析

### 3.1　实验数据与评价指标

实验过程中将 4 个城市的数据合并，选择 10 月 1 日至 11 月 16 日的数据作为训练集，11 月 17 日至 11 月 23 日的数据作为验证集，11 月 24 日至 11 月 30 日的数据作为测试集，实现用户在多城市出行的推荐任务. 实验数据的统计信息如表 2 所示.

**表 2　实验数据的统计信息**

| 数据集 | 样本数量 | 特征维度 |
|---|---|---|
| 训练集 | 1 543 004 | 323 |
| 验证集 | 237 123 | 323 |
| 测试集 | 219 873 | 323 |

另外，为了避免极端结果的影响，所有预测结果均取 5 次实验的平均值. 由于数据存在样本类别不平衡的问题，为了能更好地衡量模型表现，使模型同时兼顾较高的准确率和召回率，采用加权 $F_1$作为不平衡学习的评价指标. 加权 $F_1$评价指标的定义为

$$F_{1,\text{wt}}=\sum_{i=1}^{k} w_i F_1(i) \tag{6}$$

$$F_1(i)=2\times\frac{p_i\times r_i}{p_i+r_i} \tag{7}$$

其中：$p_i$为类别 $i$ 的精度，$r_i$为类别 $i$ 的召回率，$w_i$为类别 $i$ 的权重. $F_{1,\text{wt}}$的值越接近 1，表示模型的预测效果越好.

### 3.2　不同模型的效果比较

首先，为了验证所提模型的推荐准确度，实验选取逻辑回归模型[16]、决策树模型[17]、随机森林模型[18]、gcForest 模型、XGBoost 模型和 LightGBM 模型 6 种模型与所提模型进行对比. 这 6 种模型是常用的分类模型，凭借其运行效率高、预测效果好等优点被广泛应用于目标排序、点击率预估等机器学习任务中. 实验对以上 6 种模型进行复现，与 CaGB-

$DT_1$模型和 $CaGBDT_2$模型进行比较. 其中,$CaGBDT_1$模型表示不添加指标优化层的级联梯度提升决策树,$CaGBDT_2$模型表示添加指标优化层的级联梯度提升决策树. 不同模型的运行效率和推荐效果如表3所示.

**表3 不同模型的推荐效果和运行时间对比**

| 模型 | $F_{1,wt}$ | 时间/s |
|---|---|---|
| 逻辑回归 | 0.433 | 68 |
| 决策树 | 0.567 | 420 |
| 随机森林 | 0.669 | 1 122 |
| gcForest | 0.682 | 5 963 |
| XGBoost | 0.683 | 2 902 |
| LightGBM | 0.684 | 2 837 |
| $CaGBDT_1$ | 0.687 | 9 831 |
| $CaGBDT_2$ | 0.695 | 11 682 |

由表3可知,与集成学习模型相比,虽然逻辑回归模型和决策树模型的训练时间较短,但是模型的推荐效果较差,这一结果符合实验预期. 这是因为单模型的训练过程比较简单,面对高维非线性数据时学习能力有限,而集成学习方法通过将多个单模型集成,使得模型结构更加复杂,在高维数据中更具优势. 随机森林模型通过级联结构加深模型深度之后,推荐效果的提升也比较明显. gcForest 模型较随机森林模型的训练时间增加了5倍,但是 $F_{1,wt}$指标提升了1.3%. $CaGBDT_2$模型较 XGBoost 模型和 LightGBM 模型的训练时间也增加了5倍,虽然训练时间更长,但是由于层级模型之间的训练相对独立,所以易于并行,运行效率可以进一步提升,而且在 $F_{1,wt}$指标上也提升了1.2%,达到了0.695,这表明提出的深度模型在多模态出行推荐效果上是有效的. 另外,由于 $CaGBDT_2$模型添加了指标优化层,导致其训练时间比 $CaGBDT_1$模型更长,但是前者的 $F_{1,wt}$明显高于后者,这表明提出的指标优化层对缓解类别不平衡问题是有效的.

其次,为了验证模型的鲁棒性,在不同规模的测试集上对不同模型的推荐效果进行了比较,实验分别对未来1d,2d,3d,5d,7d 和10d 的用户点击情况进行预测,实验结果如表4所示.

由表4可知,测试集的时间越接近训练集,模型的预测效果越好,且随着测试集的时间远离训练集,集成学习的稳定性优于单个学习器. $CaGBDT_2$模型在不同规模的测试集上都可以取得较好的效果,表明该模型在中短期时间序列预测问题中的鲁棒性较强.

**表4 各模型在不同规模测试集上的预测 $F_{1,wt}$对比**

| 模型 | 1d | 2d | 3d | 5d | 7d | 10d |
|---|---|---|---|---|---|---|
| 逻辑回归 | 0.448 | 0.445 | 0.442 | 0.437 | 0.433 | 0.421 |
| 决策树 | 0.576 | 0.575 | 0.572 | 0.570 | 0.567 | 0.562 |
| 随机森林 | 0.680 | 0.677 | 0.675 | 0.674 | 0.669 | 0.660 |
| gcForest | 0.691 | 0.689 | 0.687 | 0.685 | 0.682 | 0.679 |
| XGBoost | 0.692 | 0.689 | 0.688 | 0.687 | 0.683 | 0.680 |
| LightGBM | 0.692 | 0.690 | 0.689 | 0.688 | 0.684 | 0.680 |
| $CaGBDT_1$ | 0.696 | 0.693 | 0.692 | 0.690 | 0.687 | 0.683 |
| $CaGBDT_2$ | 0.704 | 0.702 | 0.701 | 0.699 | 0.695 | 0.692 |

### 3.3 特征分析

首先,为了验证图嵌入特征对模型推荐的有效性,分别在不同嵌入特征的组合上对不同模型的运行时间和推荐效果进行了比较,实验结果如图4和表5所示. 由图4可知,逻辑回归模型,决策树模型和随机森林模型在添加图嵌入特征后,所需要的训练时间更长,而 gcForest 模型、XGBoost 模型、LightGBM 模型、$CaGBDT_1$模型以及 $CaGBDT_2$模型在添加图嵌入特征后,所需要的训练时间反而更短,这表明图嵌入特征有利于集成模型的收敛. 由表5可知,在同时添加同构图嵌入特征和异构图嵌入特征后,所有模型的推荐结果在 $F_{1,wt}$指标上都有提升,这表明构造的空间上下文特征对于模型的收敛以及多模态出行的推荐具有积极作用.

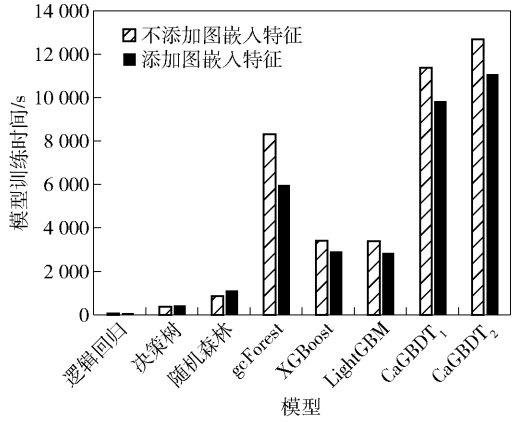



图4 添加图嵌入特征前后不同模型的运行效率

其次,为了评估特征构造的有效性,根据信息增益对特征进行排序,信息增益越高,表明该特征在构建决策树过程中的使用频率越高. 其中信息增益排

**表 5　各模型在不同嵌入特征上的 $F_{1,wt}$ 对比**

| 模型 | 不添加图特征 | 同构图特征 | 异构图特征 | 同构图 + 异构图特征 |
|---|---|---|---|---|
| 逻辑回归 | 0.432 | 0.431 | 0.432 | 0.433 |
| 决策树 | 0.563 | 0.563 | 0.565 | 0.567 |
| 随机森林 | 0.665 | 0.667 | 0.666 | 0.669 |
| gcForest | 0.676 | 0.679 | 0.679 | 0.682 |
| XGBoost | 0.681 | 0.682 | 0.681 | 0.683 |
| LightGBM | 0.682 | 0.682 | 0.683 | 0.684 |
| $CaGBDT_1$ | 0.684 | 0.684 | 0.686 | 0.687 |
| $CaGBDT_2$ | 0.693 | 0.694 | 0.693 | 0.695 |

名前 10 的特征如表 6 所示. 由表 6 可知,路线特征中,尤其是路线排名特征,如排名第 1 的出行方式、排名第 3 的出行方式、排名第 2 与第 3 的出行方式之间的行程时间差距等,对模型的预测做出了重要贡献. 这表明路线的展示顺序会对用户点击行为产生影响,排名越靠前,点击率越高,反之排名越靠后,点击率越低. 此外,如地点查询热度特征和路线属性统计特征对模型的预测也做出了重要贡献. 这是因为用户选择出行方式时一般会考虑多种因素,如行程时间、行程距离、基础设施等,所以利用统计学方法从多个角度刻画不同用户的出行偏好是有效的. 在时间特征中,如查询时间是一天中的几点几分、查询时间是否是节假日等特征的信息增益也比较高,这也符合预期.

**表 6　信息增益排名前 10 的特征**

| 排名 | 特征名称 | 信息增益 |
|---|---|---|
| 1 | 排名第 1 的出行方式 | 6 804 |
| 2 | 排名第 3 的出行方式 | 6 247 |
| 3 | 用户查询时间 | 5 567 |
| 4 | 行程时间之差 | 4 779 |
| 5 | 异构图嵌入特征 1 | 4 647 |
| 6 | 行程距离的方差 | 4 607 |
| 7 | 异构图嵌入特征 6 | 3 896 |
| 8 | 查询时间是否是假期 | 3 023 |
| 9 | 同构图嵌入特征 2 | 2 942 |
| 10 | 地点查询热度 | 2 847 |

## 4　结束语

从为用户提供个性化的多模态出行推荐的需求出发,以百度地图提供的 KDD Cup 2019 数据集为分析对象,提出了一种基于图嵌入和级联梯度提升决策树模型的多模态出行推荐方法. 该方法从用户、时间、空间和路线 4 个维度设计特征,利用统计学方法和图嵌入表示学习的方法捕获用户出行在时间和空间上的内在关系,提高了特征的表示能力,通过级联结构构建 CaGBDT 深度集成学习模型进行高性能分类. 为了解决样本类别不平衡带来的负面影响,设计了一种基于改进鲍威尔算法的指标优化方法,通过为每个类别搜索最优权重,实现最大化评价指标的目的. 利用 4 个不同城市的数据对模型进行评估与比较,证实了所提出的模型拥有较好的准确性和稳定性,能够根据用户的历史出行记录和情景上下文信息,向用户推荐一种最合适的出行方案.

**参考文献:**


[1] Herzog D, Massoud H, Wolfgang Woerndl. A mobile recommender system for personalized, multi-modal route planning[C]//25th Conference on User Modeling, Adaptation and Personalization. Bratislava: ACM, 2017: 67-75.

[2] Du Renjie, Zhang Nian, Gao Xunfei, et al. Optimal path choice based on multi-modal public transport: a case study of the Chengdu qinghua road area[C]//Fifth International Conference on Transportation Engineering. Dalian: ASCE, 2015: 1682-1688.

[3] Socharoentum M, Karimi H A. Multi-modal transportation with multi-criteria walking (MMT -MCW): personalized route recommender[J]. Computers, Environment and Urban Systems, 2016, 55: 44-54.

[4] Liu Hao, Li Ting, Hu Renjun, et al. Joint representation learning for multi-modal transportation recommendation[J]. AAAI Conference on Artificial Intelligence, 2019, 33: 1036-1043.

[5] 祁志卫, 王箭辉, 岳昆, 等. 图嵌入方法与应用: 研究综述[J]. 电子学报, 2020, 48(4): 808-818.
Qi Zhiwei, Wang Jiahui, Yue Kun, et al. Methods and applications of graphembedding: a survey[J]. Acta Electronica Sinica, 2020, 48(4): 808-818.

[6] Zhou Zhihua, Feng Ji. Deep forest: towards an alternative to deep neural networks[C]//Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne: IJCAI, 2017: 3553-3559.


(下转第 106 页)

[8] Luo Keyang, Guan Tao, Ju Lili, et al. P-MVSnet: learning patch-wise matching confidence aggregation for multi-view stereo[C]//IEEE/CVF International Conference on Computer Vision. Long Beach: IEEE, 2019: 10452-10461.

[9] Aanæs H, Jensen R R, Vogiatzis G, et al. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168.

[10] Knapitsch A, Park J, Zhou Qianyi, et al. Tanks and temples: benchmarking large-scale scene reconstruction[J]. ACM Transactions on Graphics, 2017, 36(4): 1-13.

[11] Angelova A, Long P M. Benchmarking large-scale fine-grained categorization[C]//IEEE Winter Conference on Applications of Computer Vision. Steamboat Springs: IEEE, 2014: 532-539.

[12] Hartmann W, Galliani S, Havlena M, et al. Learned multi-patch similarity[C]//IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 1586-1594.

[13] Ji Mengqi, Gall J, Zheng Haitian, et al. Surfacenet: an end-to-end 3D neural network for multiview stereopsis [C]// IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2307-2315.

[14] Choi S, Kim S, Park K, et al. Learning descriptor, confidence, and depth estimation in multi-view stereo [C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018: 276-282.

[15] Zhu Guangming, Zhang Liang, Shen Peiyi, et al. Multimodal gesture recognition using 3-D convolution and convolutional LSTM[J]. IEEE Access, 2017, 5: 4517-4524.

[16] Konolige K, Agrawal M. FrameSLAM: from bundle adjustment to real-time visual mapping [J]. IEEE Transactions on Robotics, 2008, 24(5): 1066-1077.

[17] Tola E, Strecha C, Fua P. Efficient large-scale multi-view stereo for ultra high-resolution image sets[J]. Machine Vision and Applications, 2012, 23(5): 903-920.

(上接第87页)

[7] Grover A, Leskovec J. Node2vec: scalable feature learning for networks[C]//22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016: 855-864.

[8] Liu Hao, Tong Yongxin, Zhang Panpan, et al. Hydra: a personalized and context-aware multi-modal transportation recommendation system[C]//25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage: ACM, 2019: 2314-2324.

[9] Sun Hanxiao, Wang Wenjie, Shi Zhongzhi. Parallel factorization machine recommended algorithm based on MapReduce[C]// 10th International Conference on Semantics, Knowledge and Grids. Beijing: IEEE, 2014: 120-123.

[10] 曹燕, 董一鸿, 邬少清, 等. 动态网络表示学习研究进展[J]. 电子学报, 2020, 48(10): 2047-2059.
Cao Yan, Dong Yihong, Wu Shaoqing, et al. Dynamic network representation learning: a review [J]. Acta Electronica Sinica, 2020, 48(10): 2047-2059.

[11] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality[C]//26th International Conference on Neural Information Processing Systems. North Miami Beach: Curran Associates Incorporated, 2013: 3111-3119.

[12] Cui Peng, Wang Xiao, Pei Jian, et al. A survey on network embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(5): 833-852.

[13] Ke Guolin, Meng Qi, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree [C]// 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Incorporated, 2017: 3146-3157.

[14] Chen Tianqi, Guestrin C. XGBoost: A scalable tree boosting system[C]//22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016: 785-794.

[15] Yang Gun, Zhou Fangrong, Ma Yi, et al. Identifying lightning channel-base current function parameters by Powell particle swarm optimization method[J]. IEEE Transactions on Electromagnetic Compatibility, 2018, 60(1): 182-187.

[16] Meier L, Geer S V D, Bhlmann P, et al. The group lasso for logistic regression[J]. Journal of the Royal Statistical Society Series B (Statistical Methodology), 2008, 70(1): 53-71.

[17] Lee J S. AUC4. 5: auc-based C4. 5 decision tree algorithm for imbalanced data classification [J]. IEEE Access, 2019: 106034-106042.

[18] Breiman L. Random forest [J]. Machine Learning, 2001, 45: 5-32.