基于粒子群优化和LightGBM的情景感知多式联运推荐

孙全明,曲志坚,任崇广
(山东理工大学计算机科学与技术学院,山东淄博 255049)

摘 要: 针对交通推荐服务中推荐的出行方式单一、忽略用户出行偏好以及多分类任务中样本类别不平衡等问题,本文提出一种基于粒子群优化和LightGBM的情景感知多式联运推荐方法.该方法综合考虑用户在时间、空间以及出行成本上的出行偏好,利用数理统计和表示学习方法捕捉用户出行与各要素之间的内在关系.同时,为了缓解样本类别不平衡带来的负面影响,利用基于粒子群优化算法的指标优化方法为每个类别搜索最优权重,对模型的预测结果进行修正,以实现最大化评价指标的目的.实验结果表明,与传统算法相比,本文提出的模型在时空特征提取、缓解类别不平衡和推荐准确性上均有较好的表现.

关键词: 多式联运;个性化推荐;网络表示学习;粒子群算法;特征工程
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2021)05-0894-10
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200952

Context-Aware Multi-modal Transportation Recommendation Based on Particle Swarm Optimization and LightGBM

SUN Quan-ming, QU Zhi-jian, REN Chong-guang
(*School of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong* 255049, *China*)

Abstract: In order to solve the problems of considering only one transportation mode and neglecting user preference in transportation recommendation problem, and class imbalance problem in multi-class task, a context-aware multi-modal transportation recommendation method based on particle swarm optimization and LightGBM is proposed. This method comprehensively considers the user's travel preferences in terms of time, space and travel cost, and uses mathematical statistics and representation learning methods to capture the internal relationship between user travel and various elements. At the same time, in order to alleviate the negative impact caused by the imbalance of sample class, the index optimization method based on particle swarm optimization algorithm is used to search for the optimal weight for each class, and the prediction results of the model are modified to achieve the purpose of maximizing the evaluation index. Experimental results show that compared with traditional algorithms, the model proposed in this paper has better performance in spatio-temporal feature extraction, alleviating class imbalance and recommendation accuracy.

Key words: multi-modal transportation; personalized recommendation; network representation learning; particle swarm optimization; feature engineering

1 引言

随着用户生活水平的不断提高以及网约车、共享单车等新出行方式的涌现,用户的出行模式越来越多样,对低成本、高效率的出行模式需求也越来越高.因此,交通推荐也逐渐成为导航应用中一项重要的地图服务.多式联运推荐是指根据用户输入的出发地和目的地,向用户推荐一种单式(如公交、出租车、自行车)或多式(如公交+出租车、公交车+地铁+自行车)联合运输的出行方案.

对于具有复杂性和不确定性的交通系统,如何推荐一种既符合用户偏好,又能应对不同时空背景的出

收稿日期:2020-08-31;修回日期:2021-03-17;责任编辑:梅志强
基金项目:山东省高等学校优秀青年创新团队支持计划(No. 2019KJN048);山东省自然科学基金(No. ZR2017LF004)

行方案,是推动智能交通发展的重要因素. Du 等[1]提出一种基于公共交通多式联运的最优路径选择方法. 该方法以出行的最小成本作为优化原则,利用广度优先搜索得到成本最小的出行路线,并根据用户出行情况,制定单式(普通公交)或多式(快速公交 + 普通公交)联合运输的出行方案. Herzog 等[2]提出一种个性化、多模式的路线推荐系统. 该系统将协同过滤算法与基于知识的推荐方法相结合,规划最适合用户的出行路线,并向用户推荐该路线中多种单式运输的出行方案. Socharoentum 等[3]提出一种多准则步行的个性化路线推荐方法. 该方法考虑与用户出行行为、用户所在位置和环境相关的多个因素,根据用户的历史行为分析用户偏好,在不同的情境下规划步行路线.

虽然上述方法都取得了较好的推荐效果,但是都只考虑了一种交通方式(单式)的出行方案,而且在很大程度上忽略了用户出行偏好和时空背景信息,进而无法提供令人满意的用户体验. 为了改善这些问题,Liu 等[4]提出一种基于多式联运图(Multi-Modal Transportation Graph,MMTG)的交通推荐表示学习框架. 该框架首先从大规模地图查询数据中提取一个多式联运图,以描述用户、源-目的地(Origin-Destination,OD)对和出行方式的并发性. 然后,通过学习各出行方式的网络嵌入向量,结合用户相关性和源-目的地对相关性,使表征学习规则化. 利用该方法,可以从候选集中召回一批符合当前情景的出行方式,用于在线多式联运推荐. 本文对召回的出行方式进行排序,并将其抽象为多分类建模问题,为用户推荐一种最合适的出行方式.

基于 LightGBM[5]高效、快速等优点,以及在多分类、点击率预估和搜索排序等任务中的广泛应用,本文利用 LightGBM 进行多式联运推荐. 通过给定用户、源-目的地对,以及情景上下文情况,向用户推荐一种最合适的单式或多式联合运输的出行方式. 为了捕捉用户与源-目的地对之间的上下文关系,采用融合 Word Embedding[6]和 Graph Embedding[7]的方法,将高维稀疏特征映射为低维稠密向量,进行嵌入表示学习. 同时,为了解决样本类别不平衡问题,提出了一种基于粒子群优化算法(Particle Swarm Optimization,PSO)[8]的指标优化方法,对 LightGBM 的预测结果进行权重修正,实现最大化评价指标的目的.

2　用户出行特征分析

2.1　用户出行的空间特征

利用 KDD Cup 2019 数据集,通过数据可视化分析,探索用户在时间和空间上的出行模式,挖掘用户的多模态出行规律,构建特征提取思路.

利用经纬度信息,对用户出行的空间特征进行可视化分析,如图 1 和图 2 所示. 从图中可以看出,出发地和目的地大部分集中在市区,且出发地比目的地更为集中,这表明大多数查询依赖特定的 POI(Point of Interest)[9]. 而且,通过进一步分析出发地和目的地的查询次数发现,查询次数排名前三的出发地与查询次数排名前三的目的地位置基本相同,这表明出发地和目的地查询存在热点地区. 根据出发地和目的地的空间分布情况,可以通过计算兴趣点分布和地区查询热度,表示区域历史模式.

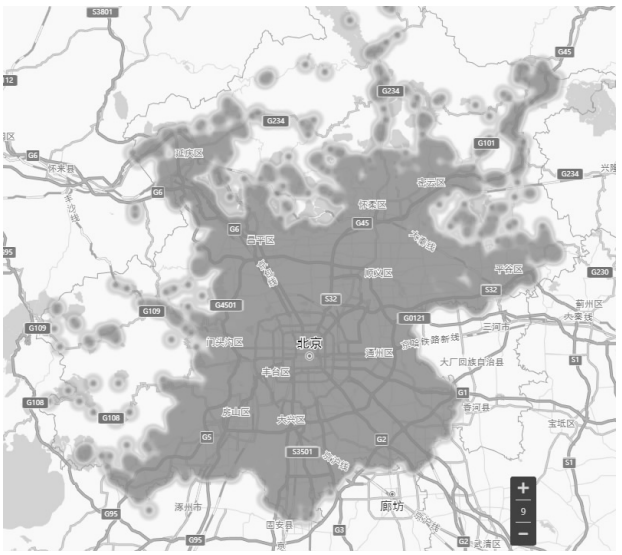


图1　出发地的空间分布

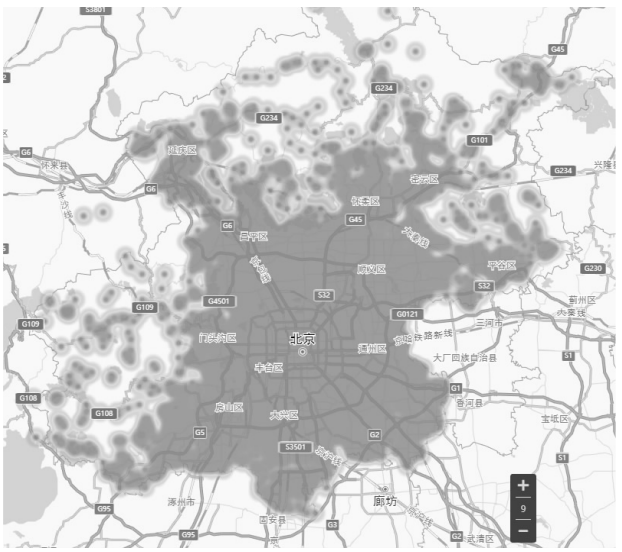
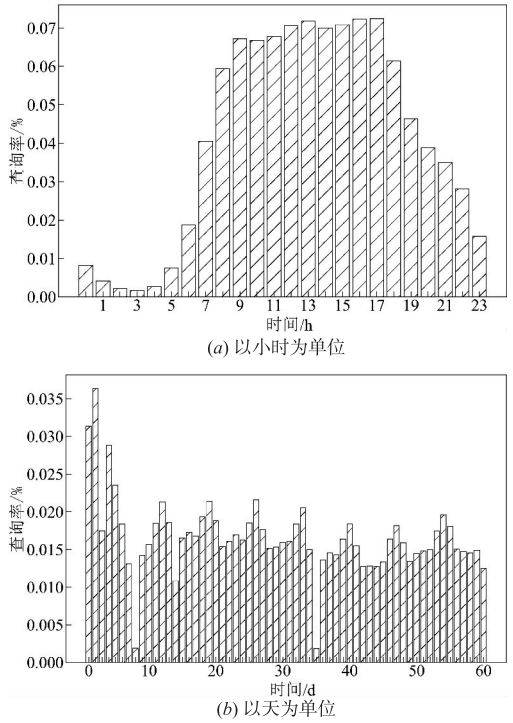


图2　目的地的空间分布

2.2　用户出行的时间特征

利用用户的出行记录,对用户出行的时间特征进行可视化分析,如图 3 所示. 图 3(a)描绘了查询记录以小时为单位的时间分布. 该时间分布存在明显的潮汐现象,高峰多集中在上午 9 点、中午 12 点以及下午 5

点,符合人们正常的出行规律. 图3(*b*)描绘了查询记录以天为单位的时间分布,该时间分布具有很强的周期性,高峰多集中在节假日和周末. 例如由于国庆节的原因,10月1日至10月7日的查询总量远高于其他时间. 根据用户行为的时间分布情况,可以通过提取周期特征或时间节点特征,表示用户出行的时间依赖关系.



(*a*) 以小时为单位

(*b*) 以天为单位

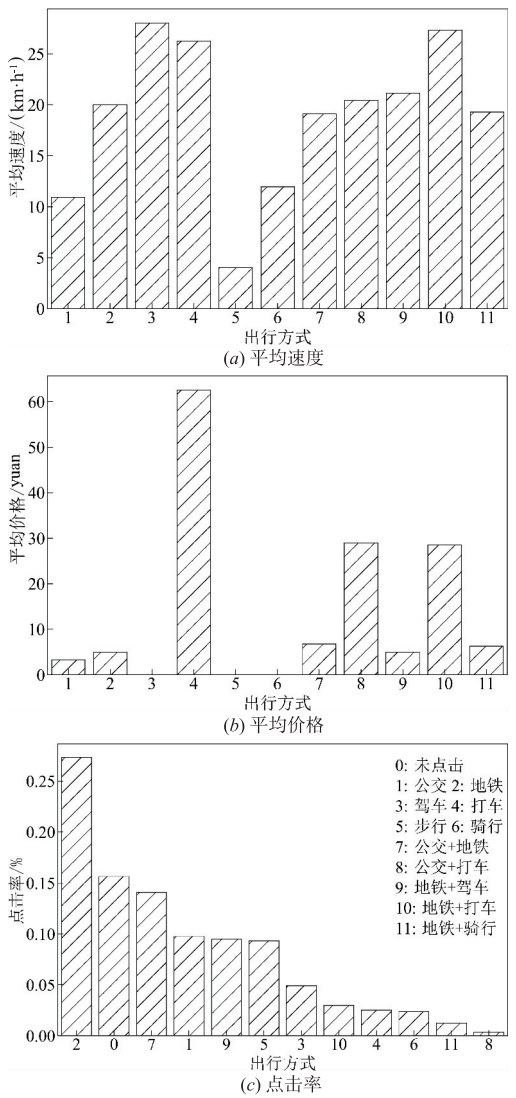
图3 用户出行的时间分布

2.3 各出行方式的点击率分布

由于不同的出行方式在速度和价格上存在较大的差异,因此导致用户在出行选择上也会存在明显的不同,如图4所示. 图4(*a*)和图4(*b*)分别描述了不同出行方式的平均速度和平均价格,图4(*c*)描述了不同出行方式的用户点击率. 由图中可以发现,虽然公交和地铁的速度较慢,但是由于其价格优势,成为超过50%的用户首选. 而驾车和打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,耽误行程时间,而且打车价格较其他方式较高,因此只有6%的用户选择这两种方式. 这些数据表明大多数用户出行时,更倾向于选择公共交通工具,这样既省钱又环保,同时在高峰期时还可以获得较好的出行体验. 此外,图4(*c*)中的数据呈现长尾分布,说明数据存在样本类别不平衡问题. 因此,在设计推荐模型时,还需充分考虑数据不均衡带来的负面影响.

2.4 各出行方式与出行距离的关系

通过数据分析还发现,用户的出发地与目的地之间的球面距离与各出行方式的点击率具有强相关性,



(*a*) 平均速度

(*b*) 平均价格

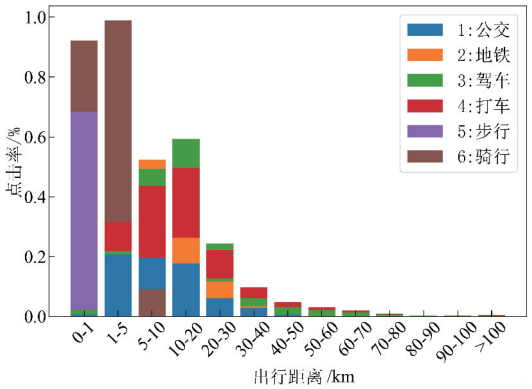
(*c*) 点击率

图4 不同出行方式的平均速度、平均价格和点击率

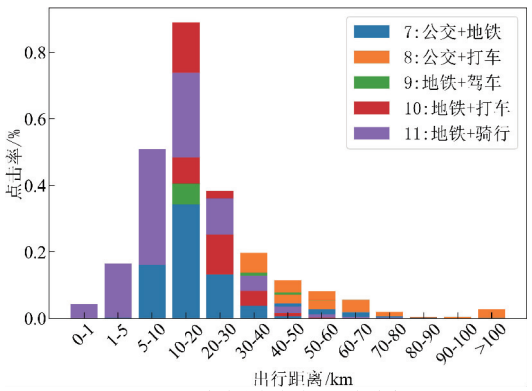
如图5所示. 图5(*a*)描绘了两地之间的球面距离与单式出行方案点击率的关系,其中步行和骑行是5km范围内出行的主要选择,公共交通和驾车是大于10km出行的主要选择. 当行程距离在5~10km范围内时,打车需求出现高峰. 图5(*b*)描绘了两地之间球面距离与多式出行方案点击率的关系,其中高峰多集中在10~20km范围内,地铁+打车的出行方案在10~20km范围内点击率最高,接近50%. 这些数据表明单式出行方案更适合中短距离的出行,而多式出行方案更适合中长距离的出行. 因此,可以通过提取出发地和目的地之间的地理特征,表示用户出行的空间依赖关系.

2.5 特征构建

路线特征. 用户出行会综合考虑每个路线的成本,例如花费和耗时. 实验过程中,提取了每个路线所对应的出行方式、预计距离、预计耗时、预计花费、路线列表中推荐的路线个数、路线列表中最优价格路线、最优时间路线、



(*a*) 单式出行方式的距离分布



(*b*) 多式出行方式的距离分布

图5　不同出行方式的距离分布

最优距离路线等特征. 除此之外,还根据各路线的展示顺序,提取了路线之间的排名特征,以及排名之间各路线的成本的数学统计信息,例如均值、方差、极值等.

时间特征. 根据图 3 中的数据分析,由于用户出行规律与时间具有强相关性,所以提取了用户当前查询时间的衍生特征来表示用户出行与时间的上下文关系,例如当前查询时间是周几、几时、是否为周末、是否为节假日等特征.

空间特征. 空间特征主要围绕出发地的经纬度和目的地的经纬度进行构造. 首先,提取经纬度的查询次数以表示用户出行的热点地区. 其次,由于各出行方式的点击率与球面距离具有强相关性,所以利用经纬度提取两地之间的球面距离 d_{OD} 和方位信息 b_{OD},具体计算如式(1)和式(2)所示:

$$d_{OD} = 2R \times \arcsin\left(\sqrt{\sin^2\left(\frac{\alpha_D-\alpha_O}{2}\right)+\cos(\alpha_O)\cos(\alpha_D)\sin^2\left(\frac{\beta_D-\beta_O}{2}\right)}\right) \tag{1}$$

$$b_{OD} = \arctan\left(\frac{\sin(\beta_D-\beta_O)\cos(\alpha_D)}{\cos(\alpha_O)\sin(\alpha_D)-\sin(\alpha_O)\cos(\alpha_D)\cos(\beta_D-\beta_O)}\right) \tag{2}$$

其中,R 表示地球半径,α_O 表示出发地的纬度,α_D 表示目的地的纬度,β_O 表示出发地的经度,β_D 表示目的地的经度.

再次,由于用户在相似源-目的地对之间出行具有相似的出行偏好,因此还提取了用户的地点查询记录,并将其转化为文本序列,然后利用 Word2Vec[10] 方法将用户与源-目的地对之间的关系映射到低维稠密向量进行表示学习.

另外,在推荐场景中,数据对象之间更多呈现的是图结构. 在本文数据中,用户在不同源-目的地对之间的出行,形成了用户行为数据和地点之间的全局关系图,而此时 Word Embedding 无法很好的展现这层关系,所以选择引入 Graph Embedding 信息. 根据用户查询的记录构建特征之间的有向带权图,利用 Node2Vec[11] 学习用户、源-目的地对之间的高阶协作关系.

具体训练过程如图 6 所示. 在 Word2Vec 训练过程中,首先提取用户查询记录并按时间进行排序;然后,以天为单位构建每个用户的行为序列并将其转化为文本,生成对应出发地 O、目的地 D 和二者组合 OD 的三个不同的文本 documents;最后,分别对三个文本进行训练,最终生成用户与源-目的地对之间的词向量. 在 Node2Vec 训练过程中,根据查询记录构建用户与地点之间的有向带权图,其中用户与地点之间的出行关系为边,用户与地点之间的共现次数为边的权重. 然后,利用结合 DFS 和 BFS 的随机游走算法对顶点序列进行采样,生成最终的行为序列. 最后对行为序列进行训练,生成用户与源-目的地对之间的嵌入向量.

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,例如用户编号与地点的共现次数,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(Factorization Machine, FM)[12] 进行特征提取. 该方法利用多项式模型将特征 x_i 和 x_j 进行交叉组合,以表述特征之间的相关性,具体计算如式(3)所示:

$$y = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{ij} x_i x_j \tag{3}$$

3　方法描述

3.1　LightGBM 模型

LightGBM 是对 GBDT 的高效实现框架,针对 GBDT 在迭代过程中多次遍历整个数据集,导致训练速度慢、内存消耗大等缺点,提出了基于直方图(Histogram-based)的决策树算法、基于梯度的单边采样(Gradient-based One-Side Sampling, GOSS)算法和互斥特征捆绑(Exclusive Feature Bundling, EFB)算法,在不损失预测准确率的前提下,加快了模型的训练速度,同时也降低了内存的消耗.

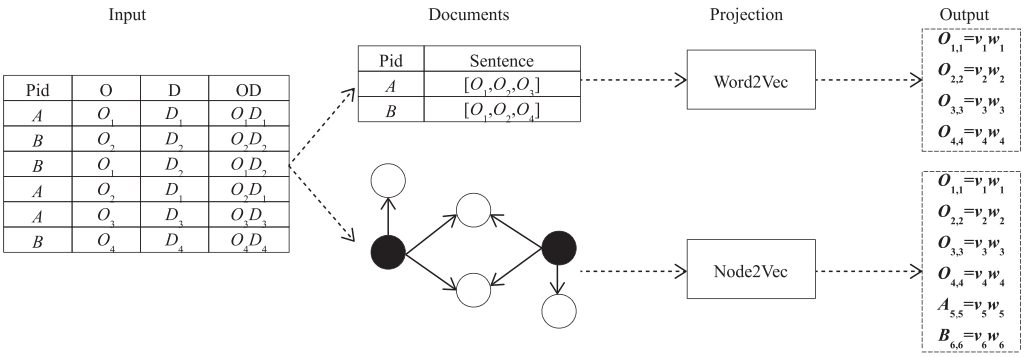


图6 Embedding训练过程

首先,为了解决传统预排序方法时间消耗大的问题,利用基于直方图的决策树算法将连续值离散化为 k 个值,并进行宽度为 k 的直方图统计. 在计算每个特征的信息增益时,由每遍历一次特征就计算一次变为只需计算 k 次,大大加快了训练速度. 而且由于做了离散化操作,内存开销也大大减少.

其次,由于梯度大的样本对计算信息增益具有更大的影响,因此 LightGBM 在每次迭代时,采用单边梯度采样算法保留小部分大梯度样本,对大部分小梯度样本进行随机采样,以获取更精准的信息增益. 该算法在获得精准信息增益的同时,也从减少样本量的角度提升了模型的训练速度.

最后,基于高维数据的稀疏性和稀疏特征的互斥性,LightGBM 采用互斥特征捆绑算法对特征进行降维,从而达到降低构建直方图特征时间复杂度的目的,以提高模型的训练速度. 该算法由贪心捆绑(Greedy Bundling)与合并互斥特征(Merge Exclusive Features)两部分组成. 通过贪心捆绑算法选择融合效果好的特征,然后利用合并互斥特征算法对这些特征进行合并,这样直方图特征构建的时间复杂度从 $O(\text{data}\times\text{feature})$ 直接降为 $O(\text{data}\times\text{bundle})$,显著加快了模型的训练速度.

3.2 基于粒子群优化的指标优化算法

在分类任务中,经常会遇到样本类别不平衡问题. 如图 4(c)所示,本文中的分类样本呈现长尾分布,这会导致分类模型的训练出现偏差,模型对于尾部类别的分类准确率不高. 常用的解决样本类别不平衡问题的方法有欠采样、过采样以及代价敏感学习.

欠采样通过减少一部分多数类样本,使得正例和反例数目接近,达到平衡样本分布的目的. 该方法可以减少内存开销,提升模型的泛化能力,但是由于抛弃了大部分数据,改变了原始数据的分布,会对模型造成偏差. 过采样通过增加一部分少数类样本,达到平衡样本分布的目的. 该方法最大程度的保留了原始数据的信息,但是由于过分强调少数类样本,增加了模型过拟合的风险. 同时,由于人为的数据扩增,加大了内存的开销.

采样算法是从数据层面解决类别不平衡问题,而代价敏感学习是从算法层面解决该问题. LightGBM 通过反复迭代更新样本权重,还可以通过调整 class weight 参数设置类别权重,这些措施都可以在一定程度上改善类别不平衡带来的影响. 但是,这些方法都依赖具体的分类器,且模型的训练与调参代价高. 本文将代价敏感学习作为分类结果的后处理,首先按照传统方法训练一个分类器,然后以实现最大化评价指标为目标,利用基于线性递减权值的粒子群优化算法[13]搜索最优权重,对模型预测结果进行修正. 优化目标如式(4)所示:

$$H=\sum_{i=1}^{k}w_iF_1(i) \tag{4}$$

式中,$F_1(i)$ 表示模型预测类别 i 的 F_1 值,w_i 表示类别权重. 该方法的优点是不依赖具体的分类器,且寻优成本低. 结合本文的类别不平衡问题,基于粒子群的指标优化方法主要包括以下 4 个步骤.

步骤 1 首先设置种群数量为 Q,随机生成维度为 12(类别数量)的初始粒子 $\boldsymbol{P}_0^q=(x_0^q,\cdots,x_{11}^q)$,其中 $x_0^q,\cdots,x_{11}^q$ 表示第 $q(q=1,\cdots,Q)$ 个粒子的初始位置. 其次,对每个粒子的历史最佳位置进行初始化,第 q 个粒子的初始历史最佳位置为 $\boldsymbol{P}_{\text{best}}^q=\boldsymbol{P}_0^q$,其初始速度为 $\boldsymbol{V}_0^q=(v_0,\cdots,v_{11})$;设置种群历史最佳位置 $\boldsymbol{P}_{\text{best}}^Q=\boldsymbol{0}_{12}$,$\boldsymbol{0}_{12}$ 表示长度为 12 的零向量;最后,对惯性权重 ω 进行初始化,初始惯性权重 $\omega_{\text{ini}}=9$,迭代终止时的惯性权重 $\omega_{\text{end}}=4$,线性递减权重的计算如式(5)所示:

$$\omega=\frac{(\omega_{\text{ini}}-\omega_{\text{end}})(K_{\max}-k)}{K_{\max}}+\omega_{\text{end}} \tag{5}$$

其中,$K_{\max}$ 表示最大迭代次数,k 表示当前迭代次数.

步骤 2 计算适应度并更新历史最佳位置. 根据每个粒子的初始位置计算适应度 f_0^q 作为每个粒子的初始适应度,并选择适应度中的最优值作为群体初始最优适应度 f_0^Q. 选择最优适应度的粒子位置作为种群初始

最佳位置 $\boldsymbol{P}_{\text{best}}^{Q}$. 适应度函数的计算如式(6)所示:

$$f_t^q = F_1(y_{\text{true}}, \text{argmax}(\boldsymbol{P}_t^q \times \boldsymbol{y}_{\text{proba}})) \tag{6}$$

其中,y_{true}表示样本标签,$\boldsymbol{P}_t^q$表示第 q 个粒子在第 t 次迭代是的位置,y_{proba}表示样本的类别预测概率.

步骤 3　更新种群中每个粒子的速度及位置,具体计算如式(7)和式(8)所示:

$$\boldsymbol{V}_t^q = V_{t-1}^q \times \omega + c_1 \times r_1 \times (\boldsymbol{P}_{\text{best}}^q - \boldsymbol{P}_t^q) + c_2 \times r_2 \times (\boldsymbol{P}_{\text{best}}^Q - \boldsymbol{P}_t^q) \tag{7}$$

$$\boldsymbol{P}_t^q = \boldsymbol{P}_{t-1}^q + \boldsymbol{P}_{t-1}^q \tag{8}$$

其中,ω 表示线性递减惯性权重,c_1、c_2为学习因子,其值为 2. r_1、r_2为[0,1]区间内的随机数.

步骤 4　检查当前迭代次数是否达到预设迭代次数 I;若 $t = I$,输出全局最优解 $\boldsymbol{P}_{\text{best}}^{Q}$和全局最优适应度 f_t^Q. 若 $t < I$,重复步骤 2 至步骤 4.

3.3　模型整体结构

多式联运推荐模型的整体结构如图 7 所示,模型的输入包括稀疏特征、稠密特征、用户行为序列和有向带权图四部分. 其中,稀疏特征包括与用户出行相关的类别特征,如用户编号,地点经纬度特征、时间特征等,通过因子分解机提取类别特征之间的交叉统计信息,用来表示用户的出行偏好. 稠密特征包括原始数据中的稠密特征和人为提取的统计特征,如路线个数、路线的行程时间、行程花费等. 该部分特征不经过二次处理,直接参与模型训练. 用户行为序列和有向带权图分别通过 Word2Vec 和 Node2Vec 方法,学习用户与源-目的地对之间低维稠密向量,表征用户出行的空间上下文关系. 最后通过拼接操作,将四部分特征向量输入 LightGBM 模型进行训练,输出类别概率,最后经过基于粒子群的指标优化算法,对预测结果进行修正,用于预测用户可能选择的出行方式.

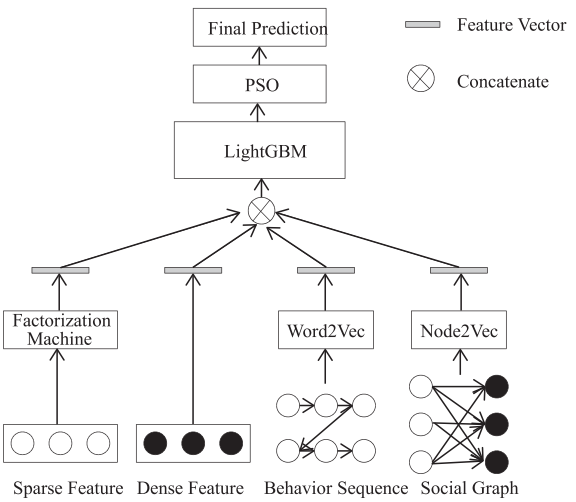


图7　多式联运推荐模型的整体结构

4　实验与结果分析

4.1　数据集信息与评价指标

实验采用北京和上海两个城市的数据集进行验证,选择2018 年 10 月 1 日至2018 年 11 月 23 日的数据为训练集,选择 2018 年 11 月 24 日至 2018 年 11 月 30 日的数据为测试集. 训练集与测试集的统计信息如表 1 所示.

表 1　训练集与测试集的统计信息

数据集	分类	样本数	特征数
北京	训练集	445340	208
	测试集	54660	208
上海	训练集	439557	215
	测试集	60443	215

原数据集中存在部分未点击(即查询之后未选择出行方式)的数据,我们将该部分数据的标签设置为 0,则本文实验为 12 分类(0 ~ 11)建模.

为了避免极端结果的影响,本文所有实验的预测结果取 5 次的平均值. 由于本文数据存在样本类别不平衡问题,为了能更好的衡量模型的表现,使模型同时兼顾较高的精准率和召回率,故采用加权的 F_1 score 作为不平衡学习的评价指标,具体计算如式(9)和式(10)所示:

$$F_1(i) = 2 \times \frac{p_i \times r_i}{p_i + r_i} \tag{9}$$

$$F_{1,\text{weighted}} = \sum_{i=1}^{k} w_i F_1(i) \tag{10}$$

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

4.2　不同模型的效果比较

首先,在 PSO 算法中,惯性权重 ω 会对最后的优化结果会产生一定影响. 较大的惯性权重有利于提高算法的全局搜索能力,而较小的惯性权重会增强算法的局部搜索能力. 为了确定粒子群算法的惯性权重参数,图 8 分别对固定惯性权重、线性递减惯性权重、随机惯性权重[14]之间的性能进行了比较,所有结果取 5 次实验的均值. 实验中,寻优终止条件为加权的 F_1 值不再提高,固定权重的值取 0.1 ~ 0.9 中的寻优时间和寻优精度的最优值,为 $\omega = 0.1$,线性递减权重的最大值和最小值分别为 $\omega = 0.9$ 和 $\omega = 0.4$,随机权重的最大值为 $\omega = 0.9$,最小值为 $\omega = 0.1$,最大值和最小值的方差为 0.2. 由图 8 可知,线性递减权重和固定权重的收敛速度均优于随机权重,线性递减权重在迭代 100 次左右开始趋于平稳,固定权重在迭代 200 次左右开始趋于平稳,而

随机权重在寻优初期出现较大波动,直到迭代350次左右才开始趋于平稳. 线性递减权重由于其初期的全局搜索能力,加之随着迭代次数的增加,局部搜索能力的增强,在寻优精度和收敛稳定性方面均优于固定权重. 所以指标优化方法采用基于线性递减权重的粒子群方法.

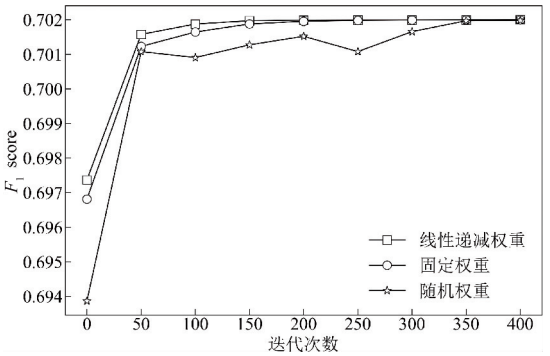


图8 不同惯性权重ω之间的比较

其次,为了验证本文模型的推荐准确度,选取 Logistic Regression[15], C4.5[16], Random Forest[17], XGBoost[18]四种模型与本文模型在北京、上海两个城市的数据集上进行比较. 表2和表3分别给出了指标优化前后的五种模型在两个城市数据集上的实验结果.

由表2可知,除了 Recall 以外,LightGBM 在所有指标上的性能都优于其他四个算法,这表明本文算法在多式联运推荐问题上的有效性. 与集成算法相比,LR 和 C4.5 性能较差,这符合我们的预期,即多模型集成学习在高维数据中更具优势. 虽然 XGBoost 的 Recall 指标优于 LightGBM,但是后者的 F_1 值更优,说明了模型在 Precision 和 Recall 指标上取得了更好的平衡.

由表3可知,除了 C4.5,本文提出的基于粒子群算法的指标优化方法在其他四个算法上都取得了较好的表现. 虽然 LR,RF,XGBoost 和 LightGBM 在 Precision 指标上都有所下降,但是在 Recall 指标上都有所提升,在北京数据集上,Recall 指标分别提升了4%、2%、0.1%、0.3%,在上海数据集上,Recall 指标分别提升了5%、3%、1%、2%. 而且,除了 C4.5,其他算法的 F_1 值都有了明显的提升,这表明本文提出的指标优化方案可以在 Precision 和 Recall 指标上取得更好的平衡.

为了验证本文提出的指标优化方案在解决样本类别不平衡问题上的有效性,分别与上采样方法 SMOTE[19]和欠采样方法 Random under sample[20]进行了比较,实验结果如表4所示. 由表4可知,在处理样本类别不平衡问题中,欠采样效果最差,这是因为欠采样丢弃了大量数据,导致模型效果不佳. 过采样可以取得较好的 Precision 和 Recall,但是 F_1 值不如粒子群指标优化方法,这表明本文提出的指标优化方法对处理样本类别不平衡问题是有效的.

为了验证本文构造的嵌入特征对模型推荐的有效性,分别对不同嵌入的特征在未指标优化的模型上进行了比较,实验结果如表5所示. 由表5可知,同时引入 Word Embedding 和 Graph Embedding 的模型在所有指标上都优于其他的模型,这表明本文构造的时空上下文特征提高了多式联运推荐的准确性.

表2 指标优化前模型的推荐效果

数据集	算法	Precision	Recall	F_1
北京	LR	0.493	0.454	0.453
	C4.5	0.578	0.594	0.586
	RF	0.726	0.702	0.683
	XGBoost	0.731	**0.718**	0.691
	LightGBM	**0.733**	0.716	**0.693**
上海	LR	0.491	0.437	0.442
	C4.5	0.572	0.591	0.581
	RF	0.725	0.700	0.684
	XGBoost	0.733	**0.725**	0.692
	LightGBM	**0.734**	0.717	**0.695**

表3 指标优化后模型的推荐效果

数据集	算法	Precision	Recall	F_1
北京	LR + PSO	0.492	0.491	0.468
	C4.5 + PSO	0.578	0.594	0.586
	RF + PSO	0.718	**0.721**	0.695
	XGBoost + PSO	0.725	0.719	0.700
	LightGBM + PSO	**0.728**	0.719	**0.702**
上海	LR + PSO	0.495	0.485	0.468
	C4.5 + PSO	0.572	0.591	0.581
	RF + PSO	0.720	0.729	0.695
	XGBoost + PSO	**0.726**	**0.735**	0.700
	LightGBM + PSO	0.725	**0.735**	**0.701**

表4 指标优化方案与采样算法的结果比较

数据集	方法	Precision	Recall	F_1
北京	Random under Sample	0.667	**0.721**	0.664
	SMOTE	0.723	0.719	0.695
	PSO	**0.728**	0.719	**0.702**
上海	Random under Sample	0.691	0.729	0.677
	SMOTE	**0.726**	0.722	0.698
	PSO	0.725	**0.735**	**0.701**

表 5　嵌入特征对推荐结果的影响

数据集	嵌入特征	Precision	Recall	F_1
北京	No Embedding	0.720	0.703	0.691
	Only Word	0.722	0.710	0.692
	Only Graph	0.718	0.711	0.691
	Word + Graph	**0.733**	**0.716**	**0.693**
上海	No Embedding	0.722	0.708	0.692
	Only Word	0.731	0.715	0.693
	Only Graph	0.727	**0.717**	0.693
	Word + Graph	**0.734**	**0.717**	**0.695**

4.3　模型的推荐结果分析

为了说明算法模型的有效性,分别对不同出行方式的推荐结果进行了分析. 图 9 描述了模型对不同出行方式的推荐精度(查准率). 由图 9 可知,模型对方式 2(地铁)、方式 5(步行)和方式 7(公交 + 地铁)的推荐精度较高,达到 0.8 左右. 而对于方式 3(驾车)、方式 4(打车)、方式 6(骑行)和方式 8(公交 + 打车)的推荐精度较差,只有 0.3 左右. 模型对于未点击的样本预测精度也达到了 0.8 以上,这表明模型对于多数类样本的预测精度较高,即查准率较高,而对于少数类样本的查准率较低.

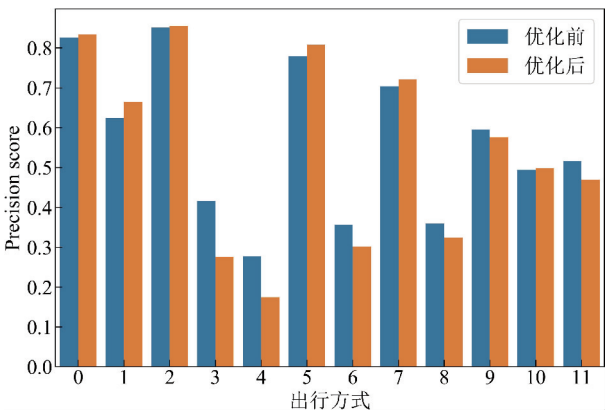


图9　模型对不同出行方式的推荐精度

图 10 描述了模型对不同出行方式的召回率(查全率). 由图 10 可知,方式 2(地铁)、方式 5(步行)和方式 7(公交 + 地铁)在获得较高精度的同时,也获得了较高的召回率,可以达到 0.9 左右. 对于方式 3(驾车)、方式 4(打车)、方式 6(骑行)和方式 8(公交 + 打车)的召回率也有了明显的提升. 虽然优化后的模型对于少数类样本的推荐精度下降了,但是大大提高了该类样本的召回率,从而使得优化后的模型在精度和召回率上取得了更好的平衡,如图 11 所示. 图 11 描述了模型对不同出行方式的 F_1 值. 由图 11 可知,优化后的模型在少数类样本上的推荐效果提升明显,且在查准率和查全率方面的整体水平较好.

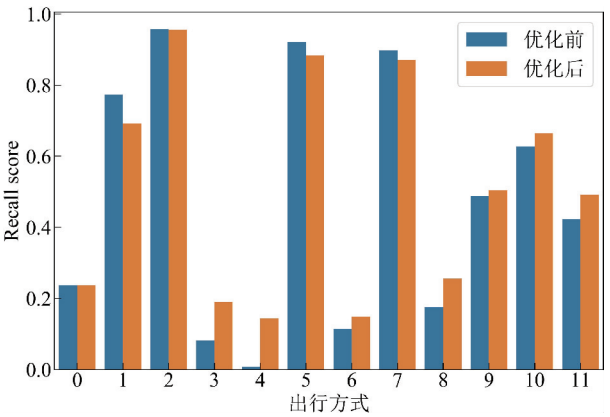


图10　模型对不同出行方式的召回率

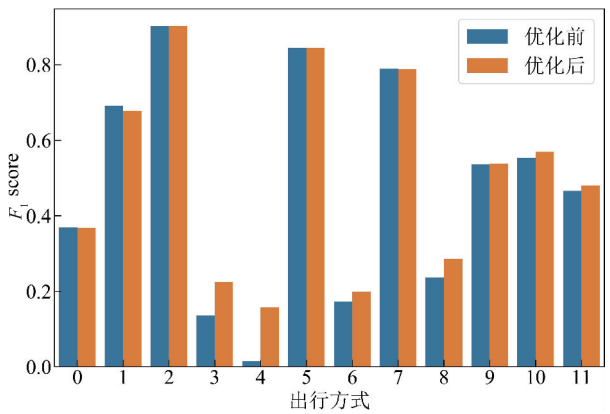


图11　模型对不同出行方式的F_1

4.4　特征重要度分析

树模型中,特征重要度特性有助于模型的可解释性,同时也可以帮助我们进行特征选择,剔除一些不重要的特征,降低模型的复杂度,提高模型的泛化能力. 为了评估 2.5 节中特征构造的有效性,本文根据信息增益对特征进行排序,图 12 给出了信息增益排名前 10 的特征. 由图 12 可知,排名第一的是时间特征 hour_minutes,该特征表示当前查询时间是一天当中的几时几分,即出行时间是影响出行方式选择的主要原因之一,符合预期. 排名 2 ~ 4 的特征代表路线排名特征,即路线

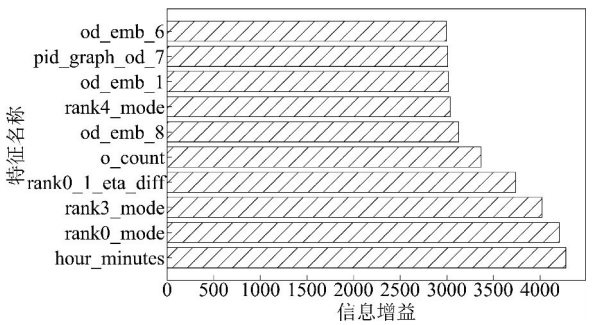


图12　信息增益排名前10的特征

排名次序会对用户点击行为产生影响,这也符合预期.此外,还观察到,出发地和目的地的嵌入特征对多式联运推荐也有显著贡献,这表明本文构造的 Embedding 特征在多式联运推荐中是有效的.

5 结论

为了向用户推荐一种个性化单式或多式联合运输的出行方式,本文提出一种基于粒子群优化和 LightGBM 的情景感知多式联运推荐方法.该方法通过构造数理统计特征、行为序列特征和行为图特征,能够更好地捕捉用户出行的时空上下文关系,提高了特征的表达能力;通过粒子群优化算法构建的模型指标优化方法,从预测结果的角度缓解了样本类别不平衡带来的负面影响.利用两个不同城市的数据对模型的推荐性能进行评估,实验结果证实,本文算法相较于对比算法,在推荐准确度和推荐稳定性上均有较好的表现,能够根据用户的历史出行记录和情景上下文信息,准确挖掘用户的出行偏好,向用户推进最合适的出行方案.

参考文献

[1] Du R J, Zhang N, Gao X F, et al. Optimal path choice based on multi-modal public transport—A case study of the Chengdu qinghua road area [A]. Proceedings of the 15th International Conference on Transportation Engineering [C]. Dalian, China: ASCE, 2015. 1682 – 1688.

[2] Herzog D, Massoud H, Wörndl W. RouteMe: A mobile recommender system for personalized, multi-modal route planning [A]. Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization [C]. Bratislava, Slovakia: Association for Computing Machinery, 2017. 67 – 75.

[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 H, Li T, Hu R J, et al. Joint representation learning for multi-modal transportation recommendation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 1036 – 1043.

[5] Ke G L, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree [A]. Proceedings of the 31st International Conference on Neural Information Processing Systems [C]. Long Beach, California, USA: Curran Associates Inc, 2017. 3146 – 3157.

[6] Lai S W, Liu K, He S Z, et al. How to generate a good word embedding [J]. IEEE Intelligent Systems, 2016, 31 (6): 5 – 14.

[7] 祁志卫,王笳辉,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 Z W, Wang J H, Yue K, et al. Methods and applications of graph embedding: A survey [J]. Acta Electronica Sinica, 2020, 48(4): 808 – 818. (in Chinese)

[8] 游思晴,周丽,赵东杰,等. 基于粒子群优化算法的协同过滤推荐并行化研究[J]. 北京邮电大学学报,2018,41(6):115 – 122.
You S Q, Zhou L, Zhao D J, et al. Research on parallelization of collaborative filtering recommendation algorithm based on particle swarm optimization [J]. Journal of Beijing University of Posts and Telecommunications, 2018, 41(6): 115 – 122. (in Chinese)

[9] Liu H, Tong Y X, Zhang P P, et al. Hydra: A personalized and context-aware multi-modal transportation recommendation system [A]. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C]. Anchorage, AK, USA: Association for Computing Machinery, 2019. 2314 – 2324.

[10] 潘博,于重重,张青川,等. 基于词性与词序的相关因子训练的 word2vec 改进模型[J]. 电子学报,2018,46(8):1976 – 1982.
Pan B, Yu C C, Zhang Q C, et al. The improved model for word2vec based on part of speech and word order [J]. Acta Electronica Sinica, 2018, 46(8): 1976 – 1982. (in Chinese)

[11] Grover A, Leskovec J. node2vec: scalable feature learning for networks [A]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C]. San Francisco, California, USA: Association for Computing Machinery, 2016. 855 – 864.

[12] Sun H X, Wang W J, Shi Z Z. Parallel factorization machine recommended algorithm based on MapReduce [A]. Proceedings of the 10th International Conference on Semantics, Knowledge and Grids [C]. Beijing, China: IEEE, 2014. 120 – 123.

[13] Nkwanyana T B, Wang Z H. Improved particle swarm optimization base on the combination of linear decreasing and chaotic inertia weights [A]. Proceedings of the 12th International Conference on Computational Intelligence and Communication Networks [C]. Bhimtal, India: IEEE, 2020. 460 – 465.

[14] Lin M J, Wang Z Y, Wang F. Hybrid differential evolution and particle swarm optimization algorithm based on random inertia weight [A]. Proceedings of the 34rd Youth Academic Annual Conference of Chinese Association of Automation [C]. Jinzhou, China: IEEE, 2019. 411 – 414.

[15] Zhang S L, Zhang L L, Qiu K M, et al. Variable selection in logistic regression model [J]. Chinese Journal of Elec-

tronics,2015,24(4):813 - 817.

[16] Xu W H, Qin Z. Constructing decision trees for mining high-speed data streams[J]. Chinese Journal of Electronics,2012,21(2):215 - 220.

[17] 潘剑飞,曹燕,董一鸿,等. 基于 Attention 深度随机森林的社区演化事件预测[J]. 电子学报,2019,47(10):2050 - 2060.
Pan J F, Cao Y, Dong Y H, et al. The community evolution event prediction based on attention deep random forest[J]. Acta Electronica Sinica, 2019, 47(10):2050 - 2060. (in Chinese)

[18] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system[A]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. San Francisco, California, USA: Association for Computing Machinery, 2016. 785 - 794.

[19] Gameng H A. A modified adaptive synthetic SMOTE approach in graduation success rate classification[J]. International Journal of Advanced Trends in Computer Science and Engineering, 2019, 8(6):3053 - 3057.

[20] Anand A, Pugalenthi G, Fogel G B, et al. An approach for classification of highly imbalanced data using weighting and undersampling[J]. Amino Acids, 2010, 39(5):1385 - 1391.

作者简介

孙全明　男,1995 年生于山东省潍坊市. 现为山东理工大学计算机技术专业硕士研究生. 主要研究方向为机器学习与数据挖掘.
E-mail:quanming_sdut@163.com

曲志坚(通信作者)　男,1980 年生于山东省青岛市. 现为山东理工大学计算机科学与技术学院副教授、硕士生导师. 主要研究方向为机器学习与数据分析.
E-mail:zhijianqu@sdut.edu.cn

任崇广　男,1982 年生于山东省临沂市. 现为山东理工大学计算机科学与技术学院教授、硕士生导师. 主要研究方向为机器学习与智能信息处理.
E-mail:renchg@sina.com