ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 3,May 2022,pp. 609－620
DOI:10. 16337/j. 1004-9037. 2022. 03. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于排序学习的城市设施选址方法

韩文军[1]，张亚平[1]，陈　红[2]，陈　丹[2]，孙婉婷[3]，赵　斌[3]

(1.国家电网经济技术研究院有限公司，北京 102209;2.江苏省电力有限公司经济技术研究院，南京 210008;
3.南京师范大学计算机与电子信息学院/人工智能学院，南京 210023)

摘　要：提出一种采用排序学习技术解决城市设施选址问题的方法，并引入人类移动性特征提升选址的质量。首先对人类移动行为进行特征提取与分析，使用双流自编码器融合人类移动性特征与其他特征，提取表征向量；然后基于候选集的表征向量与排序学习网络进行地块排序；最后，基于真实的多源数据集进行实验，结果验证了本文提出的排序学习选址方法的有效性。

关键词：排序网络；人类移动性；多源数据；神经网络；选址方法

中图分类号：TP391.3；TP183　　**文献标志码**：A

Urban Facility Locating Method Based on Ranking Learning

HAN Wenjun[1], ZHANG Yaping[1], CHEN Hong[2], CHEN Dan[2], SUN Wanting[3], ZHAO Bin[3]

(1. State Grid Economic and Technological Research Institute Co., Ltd., Beijing 102209, China;2. Economic Research Institute, State Grid Electric Power Co. Ltd., Nanjing 210008, China; 3. School of Computer and Electronic Information/Artificial Intelligence, Nanjing Normal University, Nanjing 210023, China)

Abstract: A locating method based on learning to rank is proposed to solve the location of urban facilities and introduce the features of human mobility to improve the effectiveness. First, representation vector is extracted with two stream autoencoders, fusing the features of human mobility with others. Then the plots are sorted based on representation vector of the candidate sets and the ranking network. Extensive experiments based on real multi-source dataset verify the effectiveness of the proposed locating method.

Key words: ranking networks; human mobility; multi-sourcedata; neural network; locating method

引　　言

选址问题是一个经典的运筹学问题，传统的选址模型通常以最小成本或最大利润为优化目标建立数学模型，根据模型特征选用优化算法在侯选集中选取最优位置[1]。近年来，随着云计算、大数据和人工智能技术的快速发展，数据规模持续积累，计算能力不断提升，为传统的选址研究提供了新的机遇与挑战。采用新技术的选址方法，利用精细化、高质量的真实数据对复杂场景建模，弥补传统选址模型无法贴合现实场景的不足，逐渐应用于城市规划[2]、能源开采[3]以及应急抢险[4]等领域。本文以城市地理空间为场景，将排序学习技术应用于城市设施选址问题。

机器学习技术为选址问题提供了新思路与新方法。传统选址模型的求解思路是针对优化目标确

基金项目：国家电网有限公司科技项目(5200-201956106A-0-0-00)。
收稿日期：2021-05-12；**修订日期**：2021-10-23

定约束条件建立数学模型。但是,当前现实的选址问题不仅决策标准复杂,而且需求目标往往多样,这导致传统数学模型难以建模,实际效果差强人意。当前,研究者采用机器学习方法解决选址问题,已经取得了一些研究成果,并用于城市公共设施的选址应用中。Xu等[5]针对酒店选址问题,采用聚类方法挖掘供需差距从而获取候选位置,使用监督回归模型预测排序,发现更吸引消费者的酒店位置。Quan等[6]在商业选址问题中,同样以空间聚类发现最佳候选集,使用基于图的半监督学习方法预测候选点的销售数量,推断出最优的商业区域。Karamshuk等[7]研究了各类特征对零售商店流行程度的预测能力,通过支持向量机、决策树等监督回归方式对地理区域进行排序来预测最佳地理位置。Chen等[8]在物流仓库选址问题中,以最小化仓库网络的运输成本为目标,使用神经网络算法预测销售分布,进而帮助选址决策。Liu等[9]研究了新店选址问题,提出了基于深度神经网络的DeepStore模型,分别从密集特征和稀疏特征中学习低阶特征之间和高阶特征之间的交互关系,用以预测用户的消费水平,从而选择最佳的新店位置。Wang等[10]在商店选址问题中研究空间拓扑关系,通过构建POI(Point of interest)的关联关系图对地块间的空间结构和连通性进行建模,建立模型学习表征,发现活跃社区,帮助商店的选址决策。然而,上述研究工作存在两方面的不足。首先,现有研究工作主要采用经典的机器学习方法实现选址建模,在已发现的候选集基础上完成位置排序。此类建模过程繁琐,排序效果有限。其次,现有研究工作的选址建模通常忽略了城市地理空间中人类移动行为的重要作用。鉴于上述问题,本文采用排序学习技术[11]解决城市设施选址问题,该方法可以对众多排序特征进行组合优化,直接得到精准高效的排序模型,排序效果突出。此外,本文将引入人类移动性特征,全面考虑影响城市设施选址的关键因素,提升选址质量。

本文将人类移动性特征纳入特征提取的范围,采用排序学习技术解决城市设施选址问题,主要有两方面的挑战:(1)城市区域内人类活动难以建模。与之前的选址模型不同,本文考虑了人类活动对城市设施选址的影响,但城市内人类活动行为十分复杂,对城市人群在各个区域间的流动和交互关系进行建模存在较大的难度。(2)多源异构数据难以学习。排序学习根据不同的输入数据调节神经网络的权值,这些数据之间存在一定的关联性,同时也包含一些特异性信息,这使得多源异构数据的学习非常具有挑战性,例如人类移动性数据大多为轨迹数据,一般分析方式无法体现其时空特性。

为了应对上述挑战,本文提出了基于排序学习的选址方式来解决城市设施选址问题。该方法针对实际选址场景,全面分析了城市地块的各类特征,不仅包括城市地块原有的单一属性特征,还考虑了人地交互过程与用户动态活动特征,将人类移动性特征作为待排序地块的特征之一。针对城市地块的各种特征,使用双流自编码器对多源特征进行特征融合,使用排序学习模型对学习到的地块特征进行排序,最后获得所有地块的价值排序。该方法可以对众多排序特征进行组合优化,直接得到精准高效的排序模型;引入人类移动性特征[12],全面考虑影响城市设施选址的关键因素,提升选址质量。

1 问题描述

本文要解决的问题是在城市地理空间中选出一个或多个满足特定需求的区域用于城市设施的建造。基本思想是采用空间网格将城市地理空间划分为多个局部区域,然后根据选址需求对各个区域的“可用价值”进行评估,最后选取价值最高的区域作为最佳选址。

假设在城市地理空间中,给定时间区间 $T=\langle t_1, t_2, \cdots, t_m \rangle$,由空间网格划分的所有地块集合 $\{g_{i,j}\}_{i,j=1}^{n}$,地块属性集合 $A=\{a_1, a_2, \cdots, a_l\}$,移动对象集合 $O=\{o_1, \cdots, o_{|O|}\}$,其中移动对象 o 的轨迹为 o.traj。选址的问题是对候选地块进行排序,根据选址需求输出地块集合的排序预测结果 D_{rank}。

定义 1(城市地块) 城市地块记为 $g_{i,j}$,二维空间坐标为 (x_i, y_i),城市地块是采用空间网格划分的

二维平面的各个矩形区域,其中二维平面代表城市地理空间。如图1所示,空间网格是一个由$n\times m$个方格组成的网状结构,它将城市空间划分为$n\times m$个区域,每个区域被称为1个城市地块(简称地块),通过其所在的行和列进行标记,例如$g_{a,b}$表示网格中第a行、第b列的城市地块。

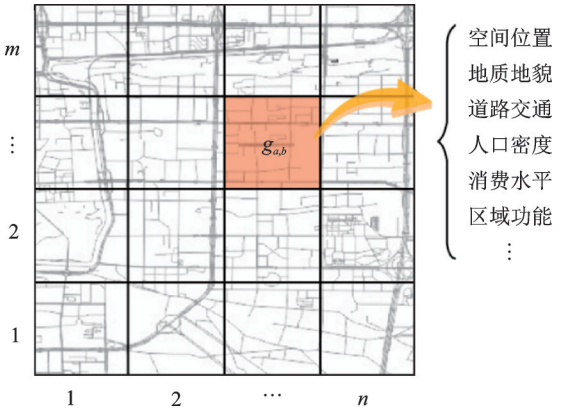


图1 城市地块及其属性

Fig.1 City block and its attributes

定义2(可用价值) 城市地块$g_{i,j}$的可用价值记为$g_{i,j}\cdot v$,用来评估城市地块对选址需求的满足程度,可用价值越高代表地块越满足选址的需求。

定义3(地块属性) 城市地块$g_{i,j}$的属性集合$A=\{a_1,a_2,\cdots,a_l\}$,其中集合元素$a_1\sim a_m$代表自然属性,包括空间位置、地质地貌、气候水文等;集合元素$a_{m+1}\sim a_l$代表人文属性,包括道路交通、人口密度、消费水平和区域功能等。

除此之外,城市空间中不同区域间还存在着一些显著的交互特性,这对大多数选址任务中地块的可用价值有着举足轻重的影响。例如在对公交站台进行选址时,一个区域与周边区域的交通流量状况直接影响着该区域能够服务的乘客数量。然而,地块与地块间的交互性是无法通过每个地块所单独具备的地块属性反映出来的,人类在城市空间中的移动特性恰好能够弥补这一不足。

定义4(时空轨迹) 给定城市空间中的移动对象集合$O=\{o_1,\cdots,o_{|O|}\}$,$|O|$表示集合中的对象个数,移动对象o_i的时空轨迹记为o_i.traj,则空间中的时空轨迹数据集可以表示为$\mathrm{Traj}_{DB}=\{o_1.\mathrm{traj},o_2.\mathrm{traj},\cdots,o_{|O|}.\mathrm{traj}\}$。移动对象的时空轨迹本质上是一个带有时间标记的空间位置序列,形如$\langle(p_1,t_1),(p_2,t_2),\cdots,(p_k,t_k)\rangle$,其中$p_i=(\mathrm{lng},\mathrm{lat})$表示移动对象在$t_i\in[t_1,t_k]$时刻的经纬度坐标。

定义5(地块区域流量) 本文将地块间人员的交互定义为地块间的流量,记为f_t,即在t时刻下,若时空轨迹数据集中存在w个移动对象从地块$g_{a,b}$移动到地块$g_{c,d}$,则$g_{a,b}$到$g_{c,d}$的流量为$f_t(g_{a,b},g_{c,d})=w$。图2中的时空轨迹造成地块$g_{5,4}$到$g_{5,5}$的流量$f_t(g_{5,4},g_{5,5})$加1。显然,跨越相邻地块的时空轨迹数量越多,地块间的流量越大,说明其人员交互性就越强。

图2 城市空间中的时空轨迹与地块间的流量

Fig.2 Trajectory in city space and the flow between city blocks

2 求解方法

本文提出的城市设施选址整体框架如图3所示,其中包含3个阶段:数据预处理、选址模型训练和模型测试。

数据预处理阶段的任务是对来自多种类型的地块属性数据进行集成,构建出一个统一的数据整体作为选址模型训练阶段的输入。该阶段主要包括对多源地块属性数据的筛选和对地块属性与人类移动数据的集成。本文采用时空图结构集成所有数据,其中图的节点为地块,包含地块属性数据,图中节点之间的边记录地块之间的流量。

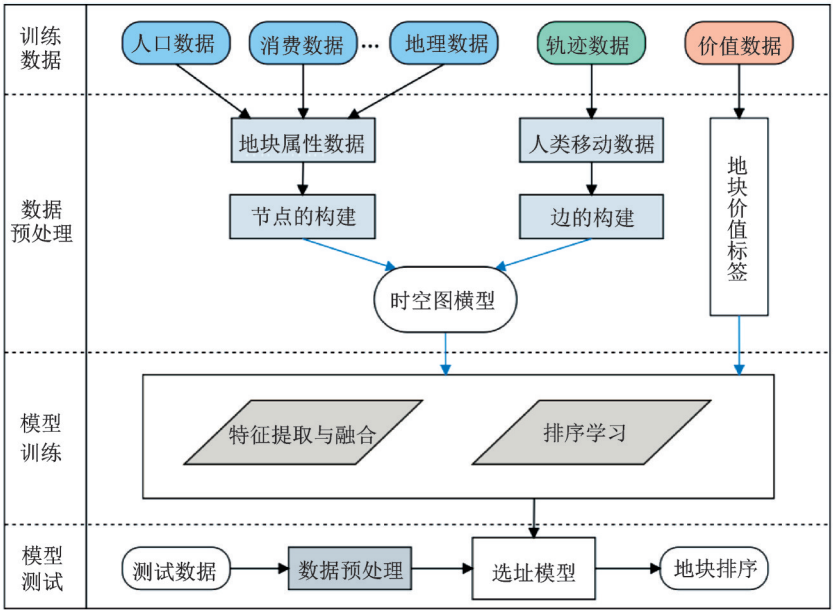


图3 城市设施选址框架

Fig.3 Framework of city facility locating

选址模型训练阶段以集成的多源数据和地块的价值标签数据为输入对选址模型进行训练。该阶段主要包括特征的提取与融合和排序学习两个模块。本文采用1个双流自编码器对时空图中的特征进行提取,并融合为1个特征向量,然后训练1个基于神经网络的排序学习模型用于对地块可用价值的评估与排序,最后在排序选址阶段使用训练好的选址模型从所有候选地块中进行选址。

2.1 数据预处理

2.1.1 多源地块属性数据筛选

由于不同选址任务对地块价值的度量方式不同,因此在面对不同的选址需求时,也应选择不同的地块属性参与对可用价值的评估。为了提高价值评估的准确率,本文根据选址需求选取对价值影响更大的那些地块属性参与计算。因此地块价值与地块属性之间的相关性成为进行地块属性数据筛选的重要依据。

皮尔逊相关系数广泛用于度量两个变量X和Y之间的线性相关程度,可以用于本文对地块价值与地块属性之间的相关性度量中。该系数的变化范围为-1到1,相关系数越接近1,代表两个变量的正相关性越强。计算方式如式(1)所示,两个变量X和Y之间的皮尔逊相关系数$\rho(X,Y)$定义为两个变量之间的协方差$\mathrm{cov}(X,Y)$与标准差的商,μ_X,σ_X为变量X的均值和标准差,μ_Y,σ_Y为变量Y的均值和标准差。本文计算所有地块属性与地块价值的相关性,选取相关系数大于0的地块属性参与地块价值的评估和排序。

$$\rho(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]}{\sigma_X\sigma_Y} \tag{1}$$

2.1.2 地块属性数据与人类移动性数据集成

属性数据反映了地块内部特征,人类移动数据反映了地块之间的交互特征,两者共同影响选址决策。然而这两类数据具备不同的数据类型和存储格式,在进行模型训练前需要对其进行集成。地块属

性数据一般只有空间维度,在短时间内不会轻易变化,而人类移动性数据由多条轨迹数据组成,一条轨迹数据由各个时刻下的轨迹点组成,且不同的轨迹数据暗含了用户的不同需求倾向,同时反映地块间的交互特性。因此简单的数据统计无法反映人类移动性数据的时空特性。

本文选取时空图模型对地块属性数据与人类移动性数据进行集成。时空图(Spatial temporal graph, STG)模型用于分析各个地块间的流量交互及其随时间的变化情况。1个时空图由有限的边和点组成,其中,节点对应1个城市地块,包含地块的各种属性;连接节点的边记录,地块间的流量,在时间维度上,图的结构及边的权重不断发生变化,形成1个三维模型。如图4所示,时空图是1个三维的图结构,它在二维空间中是1个加权有向图,详细定义如下所述。

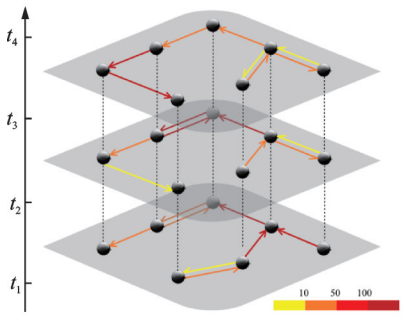


图4 时空图模型

Fig.4 Model of the spatial temporal graph

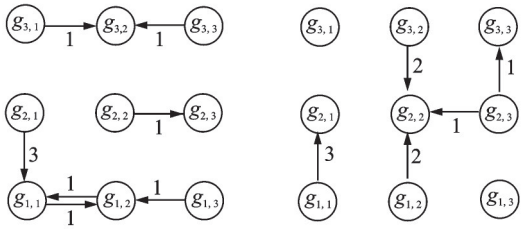
定义6(时空图) 集合$N=\left\{n_{i,j}\right\}_{i,j=1}^{n}$代表图中的点集,其中$n_{i,j}$对应地块$g_{i,j}$,集合$E=\left\{e_{i,j}\right\}_{i,j=1}^{n}$代表图中的边集,对应不同时间区间内的地块区域流量,时空图$G=(N,E)$,由点集N和边集E组成。

本文对时空轨迹中相邻两个时刻间的位置变化进行统计,将跨越两个地块的移动对象数量作为地块间的流量,从而构建出反映城市空间人类移动特性的时空图模型。若地块间的流量超过一个数量阈值μ,则时空图中将生成1条地块之间的边,其权重为流量值。由于人类的移动是有方向性的,故时空图中的边为有向边。随着时间的变化,城市空间中不同地块间的流量不断发生变化,因此时空图的结构和边的权重不断发生改变。

以图5为例介绍时空图的构建过程。城市空间中移动对象$\{o_1,o_2,\cdots,o_{10}\}$在$t_1$至$t_3$时刻下的位置数据如图5(a)所示,$g_{i,j}$表示移动对象所在的地块。根据移动对象在相邻时刻间所在地块的变化,可以统计得到任意两个地块之间发生位置移动的对象数量,在大于流量阈值μ的地块间建立有向边。例如在t_1至t_2时间段内,移动对象o_2, o_7, o_8从$g_{2,1}$移动到$g_{1,1}$,假设在当前示例中的流量阈值$\mu=1$,则在$[t_1,t_2]$的时空图中存在1条$g_{2,1}$指向$g_{1,1}$的边,其权重为3。图5(b, c)分别为时间段$[t_1,t_2]$和$[t_2,t_3]$的时空图。

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
t_1	$g_{1,1}$	$g_{2,1}$	$g_{3,3}$	$g_{3,1}$	$g_{2,3}$	$g_{1,2}$	$g_{2,1}$	$g_{2,1}$	$g_{2,2}$	$g_{1,3}$
t_2	$g_{1,2}$	$g_{1,1}$	$g_{3,2}$	$g_{3,2}$	$g_{2,3}$	$g_{1,1}$	$g_{1,1}$	$g_{1,1}$	$g_{2,3}$	$g_{1,2}$
t_3	$g_{2,2}$	$g_{1,1}$	$g_{2,2}$	$g_{2,2}$	$g_{2,2}$	$g_{2,1}$	$g_{2,1}$	$g_{2,1}$	$g_{3,3}$	$g_{2,2}$

(a) Locations of moving objects



(b) STG in $[t_1$-$t_2]$ (c) STG in $[t_2$-$t_3]$

图5 时空图的构建过程示例

Fig.5 Example of the STG construction

2.2 基于排序学习的选址模型训练

本文提出了一个基于人类移动性特征的排序学习选址模型。排序模型由双流自编码器模块和排序模块组成,其中,双流自编码器从时空图中地块内部属性特征与人类移动性特征中分别学习出各自的表征,再融合已学习到的表征作为候选地块的表征向量,而排序模块使用有标签的数据进行监督学习,更新模型参数的同时学习候选地块的表征向量,从而反映出候选地块的潜在价值。本文排序学习选址模型的总目标函数是两个模块的损失函数的加权之和,定义为式(2),其中L_{rec}为特征提取融合模块的损失函数,L_{rank}为排序学习模块的损失函数,α和β为对应的权重参数,取值范围为(0,1),L是选址模型的总目标函数。

$$L=\alpha L_{\text{rec}}+\beta L_{\text{rank}} \tag{2}$$

2.2.1 特征提取与融合模块

本文使用自编码器融合地块特征,并重新设计了双流自编码器。在本文的选址任务中,候选地块的属性特征与人类移动性特征来自多种数据集,且结构各不相同,因此需要从多源数据中提取地块特征并融合,以便于后续的排序学习。而在实际求解中,数据集来源于现实世界,数据存在冗余,数据源的不确定度较高且种类繁多,因此需要对数据进行降噪和降维。出于以上需求,本文将地块内的属性特征与人类移动性特征进行分开编码,然后在中间层对各自的表征向量进行融合,最后使用2个解码器分开还原,使得各自的输入和输出尽可能相似。模型输入定义为

$$\begin{cases}\boldsymbol{a}_i=\left[F_1^i,F_2^i,\cdots,F_n^i\right]\\ \boldsymbol{p}_i=\left[e_{i1}^t,e_{i2}^t,\cdots,e_{iN}^t\right]\\ \boldsymbol{a}_j=\left[F_1^j,F_2^j,\cdots,F_n^j\right]\\ \boldsymbol{p}_j=\left[e_{j1}^t,e_{j2}^t,\cdots,e_{jN}^t\right]\end{cases} \tag{3}$$

式中:地块v_i和v_j的属性特征为$\boldsymbol{a}_i$和$\boldsymbol{a}_j$;人类移动性特征为$\boldsymbol{p}_i$和$\boldsymbol{p}_j$。人类移动性特征中的e_{ik}^t表示在时间片t内,地块v_i到v_k之间的人类移动数量;F_n^i表示地块v_i的第n个属性数据;F_n^j表示地址v_j的第n个属性数据。

在中间层,表征向量利用神经网络映射进行融合。首先对两部分的表征向量进行简单的拼接,再通过1层神经网络对拼接后的表征向量进行映射,最后得到新的表征向量,有

$$\begin{cases}\boldsymbol{z}_i=g\left(W_{\text{c}}\left[f_{\text{e}}^1\left(a_i\right),f_{\text{e}}^2\left(h_i\right)+b_{\text{c}}\right]\right)\\ \boldsymbol{z}_j=g\left(W_{\text{c}}\left[f_{\text{e}}^1\left(a_j\right),f_{\text{e}}^2\left(h_j\right)+b_{\text{c}}\right]\right)\end{cases} \tag{4}$$

式中:f_{e}^1和f_{e}^2为2个不同的编码器;$g(\cdot)$为激活函数;W_{c}和b_{c}为神经网络参数。

最后将得到的表征向量通过解码器进行还原,还原之后的参数记为$\widehat{a_i}$、$\widehat{p_i}$、$\widehat{a_j}$和$\widehat{p_j}$。双流自编码器模块的损失函数定义为

$$L_{\text{rec}}=\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\left\|\widehat{\boldsymbol{a}_i}-\boldsymbol{a}_i\right\|_2^2+\left\|\widehat{\boldsymbol{p}_i}-\boldsymbol{p}_i\right\|_2^2+\left\|\widehat{\boldsymbol{a}_j}-\boldsymbol{a}_j\right\|_2^2+\left\|\widehat{\boldsymbol{p}_j}-\boldsymbol{p}_j\right\|_2^2\right) \tag{5}$$

2.2.2 排序学习模块

本文排序模块的构建方法参考了RankNet[12]排序网络。城市被划分为多个地块,通过地块的表征向量对地块的价值进行排序。根据不同的选址场景,地块的多个特征的重要性也各不相同,因此本文需要根据不同的问题选取相关性较高的多个特征,选择基于神经网络的排序模型,得出与排序标签相关性最高的特征组合方式。排序学习利用机器学习方法在数据集上对大量的排序特征进行组合训练,自动学习参数,优化评价指标以产生排序模型[13],方法包括感知机、神经网络、支持向量机和极限学习机等。按照输入数据的样例区分,排序学习的主要方法有3种:单文档排序(Pointwise)、文档对排序(Pairwise)和文档列表排序(Listwise)。

RankNet是基于神经网络的Pairwise排序模型。与其他排序学习方法不同,Pairwise方法考虑了文档对的偏序关系,更接近排序问题的实质。该方法基于神经网络,定义了基于概率的损失函数,通过输入调节神经网络的权值,衡量多源特征的重要性,而基于感知机与支持向量机等方法更关注输入样本的正反例的区分,不适用于本文选址问题。将双流自编码器模块中学习的地块表征向量作为输入,计算排序模块损失。排序学习模块为

$$\begin{cases} o_i = W_p^{(2)}\left(g\left(W_p^{(1)}\boldsymbol{z}_i + b_p^{(2)}\right)\right) \\ o_j = W_p^{(2)}\left(g\left(W_p^{(1)}\boldsymbol{z}_j + b_p^{(2)}\right)\right) \\ o_{ij} = o_i - o_j \\ p_{ij} = \dfrac{\exp(o_{ij})}{1+\exp(o_{ij})} \end{cases} \tag{6}$$

式中:W_p和b_p为神经网络参数;p_{ij}为排序学习模块的输出值,表示地块v_i比v_j在排序列表中位置更靠前的概率;$\boldsymbol{z}_i$和$\boldsymbol{z}_j$分别为地块v_i和v_j的表征向量;o_i和o_j为对应地块的排序等级;$g(\cdot)$为激活函数。排序学习网络使用交叉熵作为损失函数,计算方式为

$$L_{\text{rank}} = \sum_{i=1}^{N}\sum_{j=i+1}^{N} \boldsymbol{I}_{ij}\left(-Y_{ij}\log\left(p_{ij}\right) - (1-Y_{ij})\lg\left(1-p_{ij}\right)\right) \tag{7}$$

式中:Y_{ij}为地块v_i比v_j的输入标签;若地块v_i比v_j在排序列表中位置更靠前,则Y_{ij}等于1,否则等于0。$\boldsymbol{I}_{ij}$为标签指示矩阵,若地块v_i和v_j都有标签,则$\boldsymbol{I}_{ij}$等于1,否则等于0。

2.2.3 模型训练算法

本文设计的排序学习选址模型主要分为两个部分:双流自编码器模块和排序学习模块。通过该模型,本文可以对选址问题中的候选地块进行排序,帮助进行选址决策。该模型不仅考虑到候选地块的属性特征与人类移动性特征,也能借助部分有标签的地块进行表征学习,再进行统一排序。

整个排序模型算法的过程概述见算法1。

算法1 排序学习选址算法

输入:地块属性特征集合:$(F_1, F_2, \cdots, F_n)$;

人类移动性特征(E_{ij}^t);

候选地块和部分带标签的地块集合$\left(V, V_l\right)$

输出:候选地块的排序标签D_{rank}

(1) Prepare pair $\{P\}$; //数据预处理

(2) $D_{\text{train}} \leftarrow \varnothing$;

(3) for each pair$\left(v_i, v_j\right)$ in P do

(4) if both v_i, v_j have labels then

(5) $I_{ij} \leftarrow 1$, $Y_{ij} \leftarrow$rank laber;

(6) else

(7) $I_{ij} \leftarrow 0$, $Y_{ij} \leftarrow 0$;

(8) for each t in STG do

(9) $e_{ij}^t \leftarrow n_{ij}^t$;

(10) 训练样本$\left(\left(v_i, v_j, I_{ij}, \{e_{ij}^t\}\right), Y_{ij}\right)$放入$D_{\text{train}}$;

(11) end

(12) 初始化模型参数α、β;

(13) repeat //模型训练

(14) 从D_{train}随机选取一批训练数据;

(15) 损失函数最小化,更新模型的参数;

(16) until stopping criteria is met

(17) $D_{\text{rank}} \leftarrow \varnothing$; //使用已训练的模型预测元素的排序顺序

(18) for each pair $\left(v_i, v_j\right)$ in P do

(19) $Y_{ij} \leftarrow \text{resule}(u_i, u_j)$;

(20) 预测结果$\left(\left\{v_i, v_j\right\}, Y_{ij}\right)$放入$D_{\text{rank}}$;

(21) end

(22) return D_{rank}

3 实验与分析

为了验证本文排序学习选址模型的有效性,以北京市为例,选取合适的地块作为大型商场的建造地点,即对北京市的地块商业价值进行排序。地块属性特征包括地理特征、人口特征和消费特征,其中消费特征由房价数据反映,人类移动性特征从出租车轨迹数据中提取。实验选择的平台是Windows操作系统,CPU为AMD Ryzen 536006-Core Processor 3.60 GHz,RAM为8 GB。实验使用的数据包Tensorflow版本为1.4.0, Python版本为3.6.0,Conda的版本为4.9.2。

3.1 数据集

本文在商场选址任务中主要使用到5个数据集,数据集的信息具体描述如下:

(1)POI数据集。该数据集为2013年北京市的POI数据,包含北京市各类建筑(例如学校、医院和车站等)的类别信息和地理位置。

(2)人口数据集。该数据集包含2000年北京市各个区域的人口规模和各年龄段的人口分布数据,其中各年龄段包括0~14岁、15~64岁、65岁及以上。

(3)房价数据集。该数据集包括北京市2013年7 832个住房的房价。

(4)轨迹数据集。该数据集包括北京市10 357辆出租车在2008年2月2日到8日共计7天内行驶过程中由GPS定位系统采集的轨迹数据。

(5)商场评分数据集。该数据集包括当前北京市598个大型商场的百度总评分及评分人数,本文使用地块内商场的总评分与评分人数的乘积作为标签对建有商场的地块进行排序,构造出带标签的训练数据样本。

3.2 评价方法

实验将从两个方面进行比较,分别是方法有效性和特征有效性。为了验证本文选址模型的有效性,将本文模型与基准排序模型进行效果比较。为了验证人类移动性对选址模型准确率提升的作用,本文进行了特征有效性实验,将人类移动性特征用于本文模型和所有基准模型的排序学习中,与未加入人类移动性特征的排序结果进行对比。为了进一步验证本文提出的时空图模型对人类移动性数据集成和特征提取中发挥的作用,本文将人类移动性数据分别以统计数据和时空图模型两种形式进行表示,分别用于所有排序学习选址模型中,比较选址结果。本文的实验将证明,考虑时空图人类移动性特征的排序学习选址模型比一般的排序模型表现更出色。

3.3 基准模型

本文将设计的排序学习选址排序模型与已有的3个排序方法进行对比,下面将简要介绍3个基准模型及其实验参数设定。

(1)RankNet[13]:该模型是基于神经网络的经典排序学习方法,属于文档对排序方法,运用神经网络和梯度下降法最小化损失函数,本实验将模型的网络参数设定为3,中间层的节点数为32。

(2)RankingSVM[14]:该模型是基于支持向量机的变种模型,也属于文档对排序方法,定义在文档对

上的正则化Hinge损失是该方法的损失函数。

(3)PRank[15]:该模型是基于感知机的经典排序学习方法,通过感知机模型将全部训练数据准确映射到一个自定义的数值区间中。

3.4 评价指标

模型预测的排序列表定义为L_{p},$L_{\mathrm{p}}=\{l_1,l_2,\cdots,l_m\}$,其中$m$为列表样本总数,真实的标签排序列表为$L_{\mathrm{r}}$。本文将使用两个常用的排序指标对比本文的排序模型与基准模型,分别是归一化折损累计增益(Normalized discounted cumulative gain, NDCG)和Tau(肯德尔等级相关系数)。

(1)NDCG:用来衡量更靠前的样本是否被正确排序。若要衡量前K个样本,则记为NDCG@K。对于每一个样本l_i,首先通过样本在排序列表L_p的位置计算它的相关性分数rel_i,计算方式如式(8)所示,其中m为样本总数量。样本的相关性分数之和称为累计增益(Cumulative gain, CG)。考虑到越靠前的样本价值越高,因此将每个增益除以对应的损益值,样本的排序分数即为元素的折损累计增益(Discounted cumulative gain, DCG),计算方式如式(9)所示。当模型预测的排序列表与真实排序列表一致时,DCG达到理性情况下的最大值,即IDCG(Idea disconnted cumulative gain)。为了更好地展现模型的排序效果,本文对DCG做归一化得到NDCG,计算方式为

$$\mathrm{rel}_i=\frac{(m-\mathrm{rank}(l_i)+1)}{m} \tag{8}$$

$$\mathrm{DCG}=\sum_{i=1}^{k}\frac{\mathrm{rel}_i}{\log_2(i+1)} \tag{9}$$

$$\mathrm{NDCG}=\frac{\mathrm{DCG}}{\mathrm{IDCG}} \tag{10}$$

(2)Tau:关注模型预测的排序列表与真实的排序列表是否一致。假设地块v_i预测的排序位置为$L_p(i)$,真实的排序位置为$L_r(i)$,那么对于一对地块v_i和v_j,如果两者预测排序位置的相对位置与真实排序位置的相对位置一致,即$L_p(i)>L_p(j)$且$L_p(i)>L_p(j)$(或$L_p(i)<L_p(j)$,$L_p(i)<L_p(j)$),这对样本是一致的,否则是不一致的。计算方式如式(11)所示,其中,Conc为一致的样本对数,Disc为不一致的样本对数。

$$\mathrm{Tau}=\frac{\mathrm{Conc}-\mathrm{Disc}}{m(m-1)/2} \tag{11}$$

3.5 实验结果分析

3.5.1 方法有效性

本文在北京数据集上分别从5个指标将本文模型与3个基准模型进行比对,其中指标为NDCG@5、NDCG@10、NDCG@15、NDCG@20和Tau,指标值越大越好。基准模型与本文模型采用相同特征进行实验,实验对比结果如图6所示。

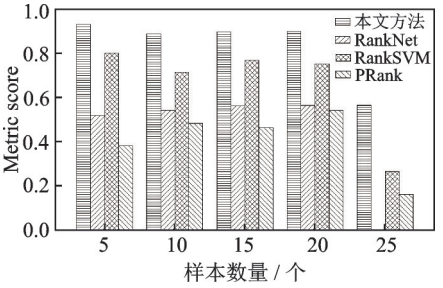


图6 本文模型与基准模型的方法有效性比较

Fig.6 Comparison of method validity of proposed model and baseline models

从结果图可以看出,本文方法优于3个基准模型,排序top3的商场分别是龙湖长楹天街、北京世纪金源购物中心和北京新奥购物中心。对于排序学习模型,直接在少量带有标签的数据集上进行学习,会导致排序效果不好。在NDCG指标上,本文方法结果最高为92.89%,最低为88.65%,优于其他3个基准模型。相较于最好的

模型，本文方法在NDCG@5指标上带来了16.16%的提升，在NDCG@10指标上带来了24.00%的提升，在NDCG@15指标上带来了16.82%的提升，在NDCG@20指标上带来了19.30%的提升；在Tau指标上，相较于最好的模型，本文方法带来了1.12倍的提升。

3.5.2 特征有效性

针对北京市大型商场的选址问题，本文选取的属性特征分别为地理特征、人口特征和消费特征，人类移动性特征来自出租车数据集，其中，人类移动性数据使用时空图进行表示。

首先，对人类移动性特征有效性进行验证。本文将特征删减版本的变种模型效果与原模型效果进行对比，删减版本的特征仅包含地块的属性特征，即地理特征、人口特征与消费特征。对比结果见图7。可以看出，融合了人类移动性特征的多源特征表现效果很好，NDCG指标最高达到了92.89%。如果仅使用地块内部属性特征进行特征提取，本文的模型效果最高仅能达到69.50%，3个基准模型的最好效果分别为48.90%、51.77%和74.14%。可以看出，选址问题需要全方面的选取多源特征，部分特征不足以达到很好的排序结果。

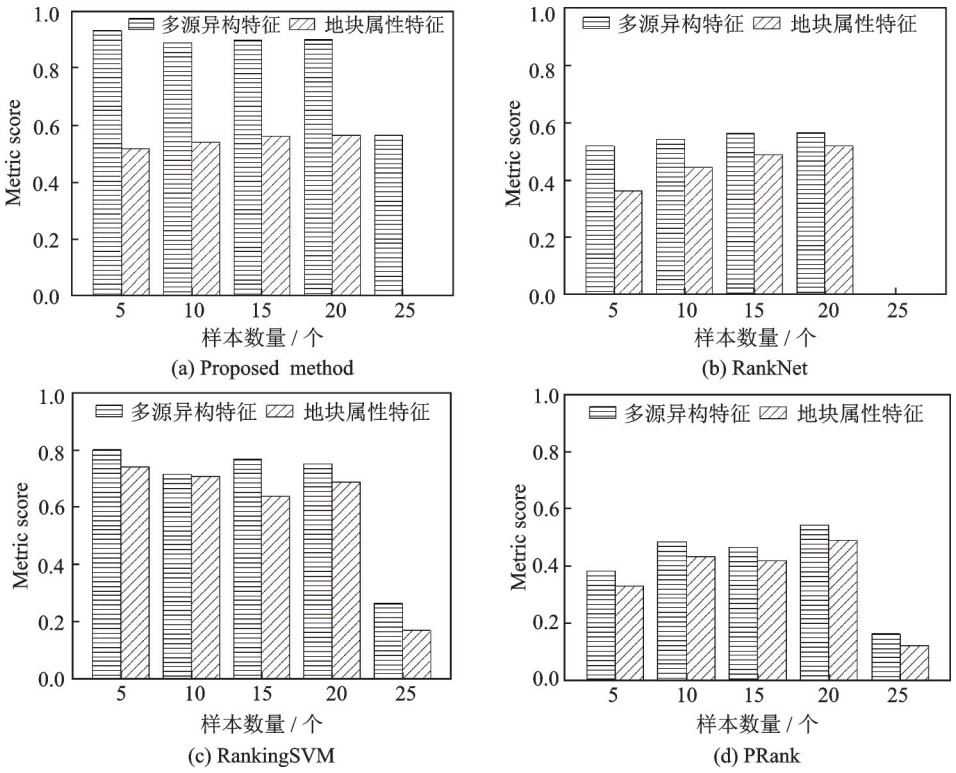


图7 多源异构特征与地块属性特征的有效性比较

Fig.7 Comparison of validity of multi-source heterogeneity features and land parcel attribute features

其次，对时空图模型有效性进行验证。本文分别用统计模型与时空图数据模型提取出租车数据，结合北京市地块属性特征，带入排序学习选址模型，实验对比结果见图8。从结果图可以看出，采用了时空图数据模型的人类移动性特征表现很好，这是因为人类移动性数据隐含了时空分布与演化规律，仅用统计模型无法体现这些特性。对于本文的方法，时空图数据模型在NDCG指标上最高提升了62.93%，Tau指标也从6.64%提高到56.53%。3个基准模型也有不同程度的提高。

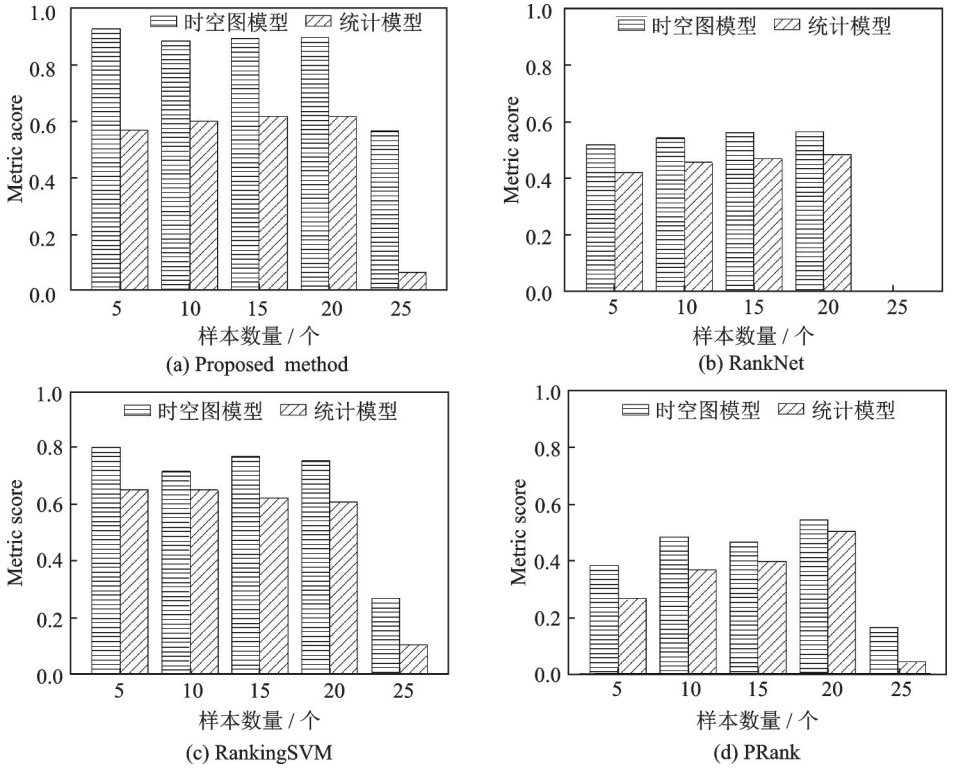


图8 时空图模型与统计模型的有效性比较

Fig.8 Comparison of validity of spatio-temporal graph model and statistical model

4 结束语

本文提出了一种基于人类移动性特征的排序学习选址模型,该模型可以有效地解决数学模型难以对应实际选址场景的问题,从地块价值排序的角度求解选址问题。该模型首先对多种复杂数据进行特征提取与分析,并引入了人类移动性特征,然后使用双流自编码器对多源特征降维降噪,提取有效的表征向量。最后使用排序学习模块对地块的表征向量进行排序,获取所有地块的价值排序。以北京市各大商场的真实评分数据与各类参考数据进行实验,实验结果表明本文提出的基于人类移动性特征的排序学习选址模型在排序效果上明显优于现有的选址排序模型。

参考文献:

[1] 王非,徐渝,李毅学.离散设施选址问题研究综述[J].运筹与管理, 2006(5): 64-69.
WANG Fei, XU Yi, LI Yixue. Review on facility location models [J]. Operations Research and Management Science, 2006 (5): 64-69.

[2] NIU H T, LIU J M, FU Y J, et al. Exploiting human mobility patterns for gas station site selection [C]//Proceedings of Database Systems for Advanced Applications—21st International Conference, DASFAA 2016. Dallas, TX, USA: Springer, 2016: 242-257.

[3] MUSTAFA S G, FATIH K, KEMAL K, et al. Suitable site selection for offshore wind farms in Turkey's seas: GIS-MCDM based approach [J]. Earth Sci. Informatics, 2021, 14(3): 1213-1225.

[4] LI Y H, ZHENG Y, JI S G, et al. Location selection for ambulance stations: A data-driven approach [C]// Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. Bellevue, WA, USA: ACM,

2015: 1-4.

[5] XU M, WANG T, WU Z, et al. Demand driven store site selection via multiple spatial-temporal data [C]// Proceedings of the 24th ACM SIGSPATIAL. Burlingame: ACM, 2016 : 1-10.

[6] QUAN X, LIU W, DOU W, et al. Link graph analysis for business site selection [J]. Computer, 2012, 45(3) : 64-69.

[7] KARAMSHUK D, NOULAS A, SCELLATO S, et al. Geo-spotting: Mining online location-based services for optimal retail store placement [C]//Proceedings of The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD, Chicago: ACM, 2013: 793-801.

[8] CHEN C, LIU J, LI Q, et al. Warehouse site selection for online retailers in inter-connected warehouse networks [C]// Proceedings of 2017 IEEE International Conference on Data Mining, ICDM 2017. New Orleans: IEEE Computer Society, 2017: 805-810.

[9] LIU Y, GUO B, LI N, et al. Deepstore: An interaction-aware wide deep model for store site recommendation with attentional spatial embeddings [J]. IEEE Internet of Things Journal, 2019, 6(4): 7319-7333.

[10] WANG P, ZHANG J, LIU G, et al. Ensemble-spotting: Ranking urban vibrancy via poi embedding with multi-view spatial graphs [C]// Proceedings of the 2018 SIAM International Conference on Data Mining. SanDiego: SIAM, 2018: 351-359.

[11] 李金忠, 刘关俊, 闫春钢, 等. 排序学习研究进展与展望 [J]. 自动化学报, 2018, 44(8): 1345-1369.
LI Jinzhong, LIU Guanjun, YAN Chungang, et al. Research advances and prospects of learning to rank [J] .Acta Automatica Sinica, 2018, 44(8): 1345-1369.

[12] BURGES C, SHAKED T, RENSHAW E, et al. Learning to rank using gradient descent [C]// Proceedings of the Twenty-Second International Conference ICML 2005. Bonn: ACM, 2005: 89-96.

[13] 陆锋, 刘康, 陈洁. 大数据时代的人类移动性研究 [J]. 地球信息科学学报, 2014, 16(5): 665-672.
LU Feng, LIU Kang, CHEN Jie. Research on human mobility in big data era [J]. Journal of Geo-information Science, 2014, 16(5): 665-672.

[14] JOACHIMS T. Training linear svms in linear time [C]// Proceedings of the 12th ACM SIGKDD. New York: ACM, 2006: 217-226.

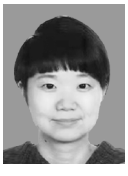
[15] CRAMMER K, SINGER Y. Pranking with ranking [C]// Proceedings of the 2011 Advances in Neural Information Processing Systems. La Jolla, Canada: NIPS, 2001: 641-647.

作者简介:

韩文军(1965-),男,高级工程师,研究方向:输变电工程勘测设计、信息系统开发及咨询研究,E-mail:hanwenjun@chinasperi.gcc.com.cn。

张亚平(1973-),女,高级工程师,研究方向:工程数据管理、基建数字化应用与建设。

陈红(1990-),女,工程师,研究方向:输变电工程变电土建技术评审。



陈丹(1977-),女,高级经济师,研究方向:电网企业项目投资分析研究与管理。

孙婉婷(1998-),女,硕士研究生,研究方向:数据挖掘。

赵斌(1978-),**通信作者**,男,博士,副教授,研究方向:数据挖掘,E-mail: zhaobin。

(编辑:刘彦东)