DOI: 10.3979/j.issn.1673-825X.2020.04.020

基于多源大数据融合的银行网点选址方法

邓　轲[1,2],冯辉宗[1],许国良[2],雒江涛[2]
(1.重庆邮电大学 软件工程学院,重庆 400065; 2.重庆邮电大学 电子信息与网络工程研究院,重庆 400065)

摘　要: 针对传统银行网点选址方法中存在的人为主观因素较大、数据量支撑不够、考虑因素理想化等问题,提出一种基于多源大数据融合的银行网点选址方法。该方法通过多源数据构造人流量、交通拥堵指数、用户价值、周边竞争网点数和人均收入 5 个基础特征,并利用协同训练的半监督学习方法扩充训练集。基于基础特征与机器学习算法构建多个子模型,将子模型的输出概率作为特征,构建基于逻辑回归的集成算法,作为银行网点选址模型,同时提出一种优化银行网点权重的损失函数,以保证模型预测中更佳的银行网点具有更高的权重。通过实验分析表明,该算法相较于传统算法预测评估更为准确,能够很好地解决银行网点选址问题。

关键词: 多源大数据; 银行网点选址; 机器学习; 逻辑回归
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673-825X(2020)04-0664-09

Site selection method of banking facility location based on multi-source big data fusion

DENG Ke[1,2], FENG Huizong[1], XU Guoliang[2], LUO Jiangtao[2]
(1. School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, P.R. China;
2. Electronic Information and Networking Research Institute, Chongqing University of Posts and Telecommunications, Chongqing 400065, P.R. China)

Abstract: In order to solve the problems in the traditional method of bank facility location, such as subjective human factors, a single source of data, and idealized factors, this paper proposes a bank location selection method based on multi-source big data fusion. The method constructs five characteristics by multi-source data, namely, human traffic, traffic congestion index, user value, surrounding competitive network points and per capita income, and uses semi-supervised learning method of collaborative training to expand the training set. Based on the basic features and machine learning algorithm to build a number of sub models, the output probability of the sub model is taken as the feature, and the integrated algorithm based on logical regression is constructed as the location model of bank outlets. At the same time, a loss function is proposed to optimize the weight of Bank outlets, so as to ensure that the better bank outlets in the model prediction have higher weights. The experimental analysis shows that the algorithm is more accurate than the traditional algorithm, and it can be well applied to the bank location problem.

Keywords: multi-source data; bank facility selection; machine learning; logistic regression

收稿日期: 2019-02-21　**修订日期:** 2020-05-20　**通讯作者:** 雒江涛　luojt@cqupt.edu.cn
基金项目: 教育部-中国移动科研基金(MCM20170203);重庆市基础与前沿研究计划重点项目(cstc2015jcyjBX0009);重庆邮电大学人才引进项目(A2017-10)
Foundation Items: The Ministry of Education-China Mobile Research Fund Project (MCM20170203); The Chongqing Municipal project under GRANT cstc2015jcyjBX0009; The Talents Introduction Project of Chongqing University of Posts and Telecommunications (A2017-10)

0 引　言

随着社会经济与互联网技术的发展，各大商业银行的业务飞速扩增，面对市场环境日趋复杂的社会经济形势，为抢占更多的用户与市场，其选址是否合理直接影响到银行自身的经济效益和竞争力。因此，一个科学合理的银行网点选址决策对银行未来规划及发展有着重要的意义。

传统的选址方法主要可以分为定性分析法和定量分析法，定性分析法包括有层次分析法、模糊综合评价法、灰色关联法、中心法[1-4]等，这些方法在操作上较为简单，具有一定的可行性，但这些方法在选址过程分析中过于简单，且通常无足够多的数据支撑，选址结果受人为主观判断影响较大，选址结果缺乏足够理论支撑，缺乏说服力。定量分析法主要可分为启发式方法和模拟方法，如粒子群算法、遗传算法、蚁群算法[5-8]等，此类方法通常考虑选址影响因素较少，考虑不够全面，且不便对各种多源数据进行相关分析。

近年来，国内外针对选址问题展开了大量研究，在银行选址方面，文献[9]将影响银行网点选址的 7 种因素作为输入因子，并分析样本数据，构建 7 种输入，3 层网络的 BP 神经网络算法，并引入惯性冲量和动态因子，构建一种银行网点选址与评价体系。文献[10]将熵权法与理想点法(technique for order preference by similarity to ldeal solution，TOPSIS)相结合，利用熵权法提高 TOPSIS 模型的精度，构建组合模型进行分析验证，以此来提高银行网点选址决策的可靠性。文献[11]将卷积运算方法应用于银行选址问题，并通过启发式算法对模型求解。

在其他行业选址方面，文献[12]将遗传算法与支持向量机算法相结合，优化相关参数，应用于城市车库选址。文献[13]提出一种基于人工神经网络与混合整数非线性规划的仓库选址问题。文献[14]将统计学和社会科学理论相结合，提出一种基于社会选择的综合评估模型，以提高避难所选址的可靠性。文献[15]引用模糊随机理论，设计了一种基于优先级的粒子群算法，文献[16]采用均差排序法应用于配送回收中心的选址问题。

随着大数据技术受到广泛关注并应用于各行各业，各大企业也将大数据相关技术融入自身业务中，在选址方面，美团点评餐饮平台商业智能部开发了基于大数据的选址工具，腾讯云基于大数据对位置与人群进行深入洞察，为企业与个人提供智能选址服务。因此，为解决传统银行网点选址方法中数据量过少，且难以全面考虑当代社会银行网点选址中复杂、抽象的选址要素及其之间的复杂关系的问题，本文将海量数据与机器学习相结合，结合经典模型和预测算法，提出一种基于多源大数据融合的银行网点选址方法，从海量数据中提取其潜在价值，为银行网点选址提供理论和数据支撑。

1 选址特征分析

1.1 选址目标

本文的选址目标是对给定目标区域，给出合理科学的选址方案，同时对选址方案的优先级进行排序，确保选址方案有效可用。由于在银行网点选址应用过程中，需要考虑的因素十分复杂，难以准确地构建数学模型或目标函数进行求解。因此，本文将利用大数据技术及机器学习算法，对多源数据进行分析，挖掘其内在价值，为银行网点选址提供可靠有效的解决方案。

1.2 选址影响因素分析

影响银行网点选址的因素比较多样和复杂，不仅需要考虑经济效益，还需要考虑环境效益和社会效益。从总体上来说，影响银行网点选址的主要因素包括人口因素、市场竞争因素、经济基础因素以及道路交通因素[16]。

1.2.1 人口因素

影响银行网点选址的人口因素是指银行点覆盖范围内的人口总量、年龄分布、购买力指数、人口增长率、流动特点、学历、家庭特点等。作为银行服务用户的物理载体，尽可能地服务更多更优质的客户是银行的主要目标之一。因此，人口因素是银行网点选址中需要考虑的重要因素之一。

1.2.2 市场竞争因素

市场竞争因素也是银行网点在选址时优先考虑的因素之一。随着市场的成熟、用户的选择也更加多样化，在同一个区域内，如果服务水平相当的银行网点分布密度较大，就会导致竞争加剧、用户分散的局面，对发展有不利影响。

1.2.3 经济基础因素

经济基础因素主要是指选址区域内的经济环境、周边的基础设施等因素。经济环境主要是指选址区域内的经济状况、当前利率水平、就业率等。一个好的经济环境不仅可以大幅度缩减营销成本，对

网点业务开展也提供了极大的便利。同时,如果一个区域内的经济水平较高,相应地消费水平也会较高,对金融服务需求会较大,因此,更适合营业网点的业务开展。

1.2.4 道路交通因素

城市道路交通是连接用户和银行设施的载体。在选址过程中,应该综合考虑道路自身路况、公共交通设施的分布、道路连接状况等因素。一个区域的道路交通状况,其决定了能否给客户提高最大化的便捷度。

1.3 特征概述

本文从银行网点的商业性质及自身定位出发,充分考虑其影响因素。针对人口因素,利用移动用户位置数据,分析区域内的人流量。针对交通因素,利用本地公交轨迹数据分析道路交通的拥堵状况。针对经济基础因素,通过移动信令数据,从用户上网行为习惯出发,分析用户价值,利用房价与其他数据分析区域内人均收入。针对市场竞争因素,分析周边竞争网点数。其特征与数据来源如表 1。

表 1 选址方法基础特征

Tab.1 Basic features of location method

特征	数据
人流量特征	移动用户位置数据
交通拥堵指数特征	公交轨迹数据
用户价值特征	移动信令大数据
竞争网点数特征	网络爬取数据
人均收入特征	房价与其他数据

1.4 数据处理

本文以重庆市区内银行网点作为研究对象,利用百度地图 API(application programming interface) 获取重庆市市区地图数据,如图 1,并对重庆市区地图进行 500 m×500 m 的栅格划分,并将对应数据映射至栅格范围内。每个栅格坐标可以用(p,l)进行表示。

本文通过重庆某移动运营商平台,对地图进行栅格划分,同时获取栅格范围内的基站信息,基站位置信息由位置区识别码(lac,ci)进行表示,根据基站信息,对用户手机信令数据、用户位置数据进行匹配。将其映射至所在栅格范围内。其公式可表示为

$$f(c)_{lac,ci}=c\in C(p,l) \tag{1}$$

(1) 式中: $C(p,l)$为坐标(p,l)对应栅格内的所有基站信息; c 为通过基站匹配到的手机信令数据、用户位置数据等。

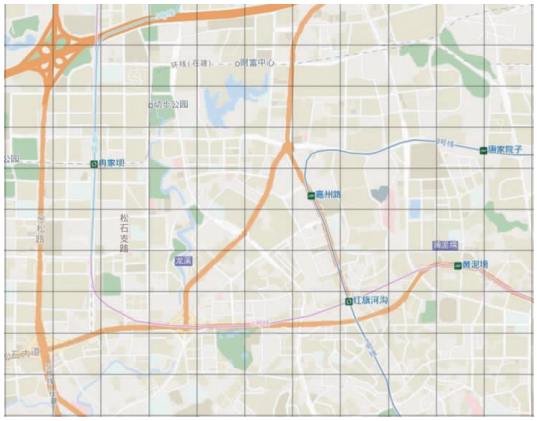

图 1 地图栅格划分

Fig.1 Map raster division

通过经纬度坐标,将公交运行轨迹数据映射至对应栅格内,对于栅格内公交运行轨迹数据,都有$x_1\leqslant lon\leqslant x_2$, $y_1\leqslant lat\leqslant y_2$,其中, lon, lat 为公交轨迹 GPS 数据, x, y 表示所在栅格的经纬度上下限。

同时,通过专家对栅格内银行现有经营现状、发展潜力、市场竞争等多因素分析,对少部分数据进行标记并排序,作为训练样本,未标记数据作为测试集样本。

1.5 特征提取

针对已有的基础数据展开研究,充分挖掘数据与银行网点选址的相关联系,选择与网点选址相关的特征参数作为选址模型的输入。利用用户位置数据与基站数据分析区域内的人流量,并对人流量进行划分; 利用公交轨迹数据与路段长度,构建基于路段长度作为权重的交通拥堵指数特征; 利用移动信令数据、用户上网数据,基于层次分析法,确定区域内的用户价值。利用网络爬取数据,分析区域内的竞争网点数与人均收入,其流程如图 2。

1.5.1 人流量特征

本文基于重庆某运营商平台,确定栅格范围内基站信息,利用移动信令数据统计基站范围内移动用户数,通过去重处理,统计记录条数,获取该地人流量。同时,人流量包括固定人流量和非固定人流量,不同类型人流量对银行网点选址也有着不同作用,因此,本文在确定人流量的同时,划分出该区域居住人数、工作人数以及流动人数。

移动用户位置数据表示为$(time,lac,ci)$,由(1)式可得出其所在栅格坐标,因此,用户位置数据可以转换为$(time,p,l)$。单位时间内人流量计算公式为

$$f(s)_{time}=S\in F(time,p,l) \tag{2}$$

(2) 式中,$S \in F(time, p, l)$ 为单位时间内落在栅格 (p, l) 内的用户数量。

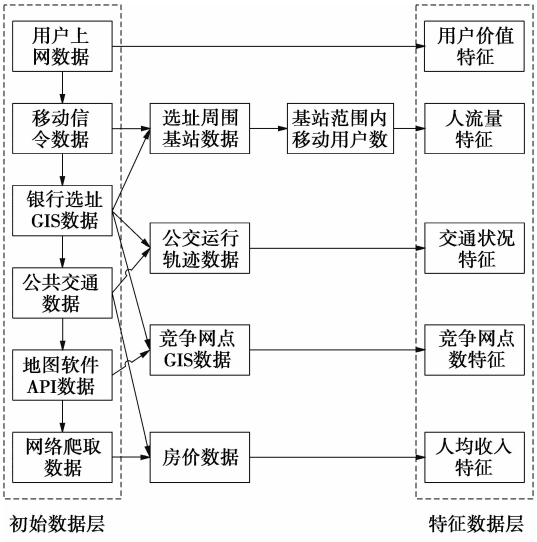

图 2　数据分析流程图

Fig.2　Data flow chart

同时,根据用户滞留时间,设定阈值 t,划分出流动人数和固定人数,根据工作时间区间、休眠时间区间将固定人数划分出工作人数和居住人数。

1.5.2　交通拥堵指数特征

本文将公交轨迹数据映射至对应栅格区域内,从交通状况来判断该区域是否适宜建立银行网点。因此,需要计算该区域的交通拥堵状况来对该区域交通进行评估,本文结合本地公交大数据,基于路段权重计算交通拥堵状况指数特征。其中,本地公交运营数据如图 3。

伪IP	纬度	经度	高度	速度	方向	累计里程	时间	上一站点编号	站间里程	限速值
0004B8A2	29.52136	106.3616	126	12	79.3	255256	2018/3/6 9:21	16	18.99	50
0000B698	29.56788	106.489	236	11	70.21	255256	2018/3/6 9:21	17	4.53	35
19989	29.58058	106.5265	19	11	166.75	255256	2018/3/6 9:21	2	1.21	35
0005AB88	29.60199	106.2897	141	32	184.73	255256	2018/3/6 9:21	11	8.55	50
000289A8	29.81214	106.4034	246	0	0	255256	2018/3/6 9:21	29	9.7	40

图 3　部分公交轨迹数据

Fig.3　Partial bus trajectory data

基于路段权重计算交通拥堵状况指数的详细步骤如下。

根据本地公交数据实时上传间隔,路段 i 的拥堵指标 A_{ij} 的计算公式为

$$A_{ij} = \left(\frac{RS_{ij}}{CS_{ij}} - 1\right) \times 100\% \tag{3}$$

(3) 式中: RS_{ij} 表示该路段 i 第 j 个间隔参考速度; CS_{ij} 表示该路段 i 第 j 个间隔的计算速度,即实际的运行速度。

根据拥堵指标值 A_{ij},计算统计间隔内选址范围内指标值 B_j,以路段的长度 L_i 为权重系数,对各个路段的指标值 A_{ij} 进行加权求平均,B_j 计算公式为

$$B_j = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} L_i \cdot A_{ij}}{\sum_{i=1}^{N} L_i} \tag{4}$$

最后求出交通拥堵状况指数,即每月数据仅统计该月的高峰时段(07: 00—19: 00) 共 12 h 的 n 条数据记录,该指标值作为交通状况指数 C,其计算公式为

$$C = \frac{\sum_{j=1}^{n} B_j}{n} \tag{5}$$

1.5.3　用户价值特征

确定范围内所有用户数,结合移动手机信令数据,用户上网使用习惯,构建用户价值计算模型,其步骤如下。

建立阶梯层次结构,底层为随机选取的该区域的 n 个用户,下层为影响价值的 k 个指标,本文 k 为 3,分别为金融 app 使用次数、支付行为信息、是否属于该银行,其中,金融 app 使用次数与支付行为信息为归一化后的值,是否属于该银行取值为 0,1。上层为通过计算后所得出的区域用户价值,其流程如图 4。

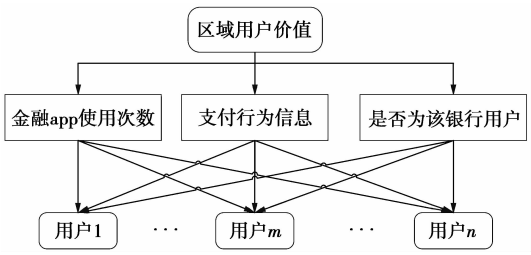

图 4　建立阶梯层次结构

Fig.4　Establish a step hierarchy

建立判断矩阵: 计算 3 个指标对用户评分的影响权重,构建对比矩阵,对各要素进行两两比较,w 为各指标对应权重,两两比较后可以得到矩阵 $\boldsymbol{A}$ 为

$$\boldsymbol{A} = \begin{bmatrix} w_1/w_1 & w_1/w_2 & w_1/w_3 \\ w_2/w_1 & w_2/w_2 & w_2/w_3 \\ w_3/w_1 & w_3/w_2 & w_3/w_3 \end{bmatrix} \tag{6}$$

对矩阵 $\boldsymbol{A}$,通过一致性检验方法,求其最大特征根与其特征向量 $\boldsymbol{a} = (a_1, a_2, a_3)^{\mathrm{T}}$,通过对应的影响因素 x_{ij},求得区域用户价值为

$$val = \frac{\boldsymbol{a} \cdot \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}}{n} \tag{7}$$

1.5.4 竞争网点数特征

竞争网点数对银行网点选址有着重要影响,网点数多则意味着该区域竞争比较激烈,挖掘新客户难度将会加大,网点数少则意味着该区域不具备相应的市场环境,本文通过地图 API,查询栅格范围内竞争银行网点数,作为模型的输入参数之一。

1.5.5 人均收入特征

以银行所在栅格区域内人流量进行划分,将人流量划分为居住人数、工作人数,即居住在该区域用户的人均收入以及工作在该区域用户的人均收入。

居住地用户人均收入 X 主要通过爬取周边 n 个小区的房价 x_i,各小区居住人数 p_i,以及该城市房价收入比 ω 计算得出,其计算公式为

$$X = \frac{\omega \sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i} \tag{8}$$

工作地用户主要分为身处写字楼、公司的用户以及其他用户,因此,本文将结合网络爬取数据与本市人均收入数据,计算栅格区域内工作地用户的人均收入。工作地用户人均收入 Y 主要通过网络爬取所在栅格区域内的 n 个公司,包括公司的人数规模 p_i 和公司人均工资 x_i,其数据主要来源包括看准网、智联招聘等。对于该区域内的其他工作地用户以本市平均收入 y 表示,α,β 为对应所占比例,工作地用户人均收入 Y 计算公式为

$$Y = \alpha \cdot \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i} + \beta \cdot y \tag{9}$$

2 银行网点选址模型

2.1 模型架构

本文模型主要分为 3 个部分: 数据采集部分、特征工程部分和模型训练部分。

数据采集部分通过重庆市某移动运营商大数据平台获取移动信令大数据,公交轨迹大数据,通过地图 API 爬取银行网点周边数据,如房价数据、道路数据、竞争网点数据等。

特征工程部分对数据采集部分获取的数据进行清洗,并结合数据与银行之间的相关联系,构建影响银行网点选址评定的相关特征。

模型训练部分利用已有标签数据,通过协同训练方法扩充训练集,以此增加模型训练的可靠性,最后构建银行网点选址模型。协同训练方法如图 5。

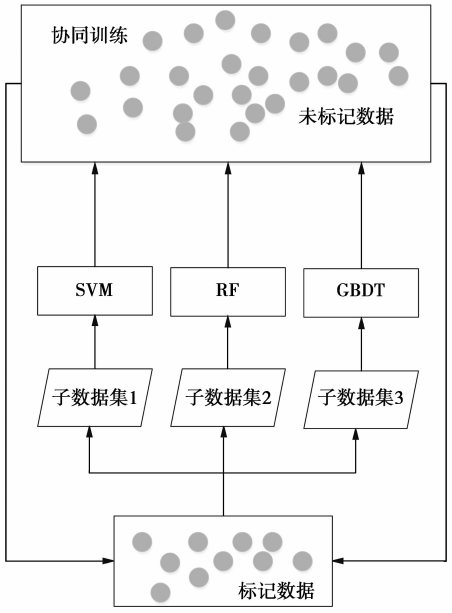

图 5 协同训练

Fig.5 Co-training

首先,从标记数据中随机选取 70%数据作为子数据集,一共选取 3 次,即 3 个子数据集。利用不同数据集与不同算法构建 3 个差异度较大的模型,3 个模型分别对未标记数据进行标记,并将置信度高的数据加入标记数据集作为训练集一起训练,重复此过程使模型不断更新,以此迭代使得更多未标记数据加入到标记数据集中。具体步骤如下。

算法: 协同训练

输入: 子数据集 $F=\{F_1,F_2,F_3\}$

标记数据集 G_1

未记数据集 G_2

迭代次数阈值 θ

输出: 标记数据集 G_1

方法:

$i=0$

DO

$SVM \leftarrow SVM.Learning(F_1)$;

$RF \leftarrow RF.Learning(F_2)$

$GBDT \leftarrow GBDT.Learning(F_3)$

将未标记数据集 G_2 中每条数据输入到 3 个模型中进行预测,并将每个模型预测到的置信度最高的部分样本加入到 G_1 中,并将标记数据 G_1 重新划分构建 3 个子数据集 $F=\{F_1,F_2,F_3\}$。

i++

Until $i>\theta$ or $G_2=\varnothing$

Output: G_1

本文模型构建分为以下几个步骤: ①引入移动信令数据、公交轨迹数据、网络爬取数据、用户上网数据等多源数据构建人流量、交通拥堵指数、用户价值、竞争网点数、人均收入等特征; ②基于已构建特征，构建银行网点选址模型; ③基于多模型的输出概率进行模型融合，并对损失函数进行优化。整体框架如图 6。

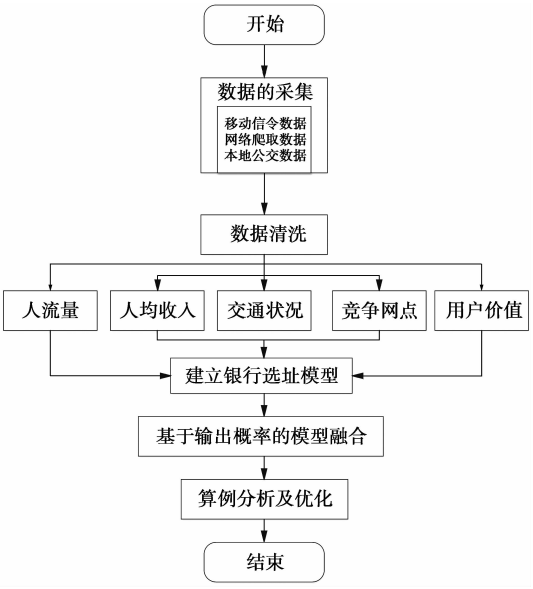

图 6　模型框架

Fig.6　Model framework

2.2　选址推荐

2.2.1　问题定义

银行网点选址问题定义为对给定区域栅格划分的数据，将已标记数据作为训练集样本，利用协同训练方法对余下部分未标记数据进行标记，扩充标记数据集。从已标记的数据集中随机选取正负样本作为训练集，通过对待选区域人流量、交通拥堵状况、用户价值、竞争网点数、人均收入等因素进行综合分析，为银行选址提供选址决策。即对给定区域 A 和候选选址集合 R，为区域 A 提供最优选址方案 $r(r \in R)$。同时，从正负样本中选取部分数据作为验证，以评估模型效果。

2.2.2　选址流程

本文选址模型算法为一个多模型融合算法。其构建 2 层模型，第 1 层模型为预测精度高、差异度较大的多个模型，主要包括梯度提升树、支持向量机、K 近邻，逻辑回归等。并将第 1 层模型的输出概率作为第 2 层模型的输入特征，并构建一种基于优化损失函数的逻辑回归算法。其实现方式如图 7。

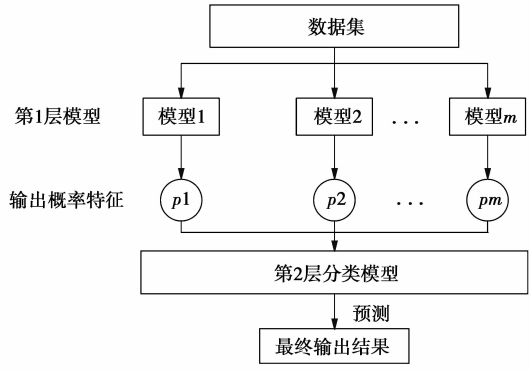

图 7　模型构建

Fig.7　Model building

其详细步骤如下。

1) 选取 m 个不同的模型作为第 1 层分类器，本文使用不同的数据集方法以及不同的模型方法获得 m 个不同分类器，所用算法为支持向量机、梯度提升树以及 K 近邻算法等。

2) 对第 1 层分类器的基础模型进行 5 折交叉验证，即将训练集分成 5 等分，选取其中 1 等分以及测试集作为预测集，另外 4 等分作为训练集，则一共需要进行 5 次预测，最后可以得出训练集所有记录的预测概率，以及 5 次测试集预测的概率。

3) 将训练集预测概率以及 5 次测试集预测的概率的平均值作为特征，一共有 m 个不同模型，即有 m 列特征，基于逻辑回归构建第 2 层分类器，对测试集进行预测。

基于逻辑回归的预测函数表示为

$$h_\theta(x) = 1/e^{-g(x)} \tag{10}$$

$$g(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n \tag{11}$$

(10)—(11) 式中: $h_\theta(x)$ 为选址决策模型，其输出值为 $h_\theta(x) \in [0,1]$，x 为对应特征数据，将预测函数计算得到的值进行归类，表示为

$$y = \begin{cases} 1, & \text{如果 } h_\theta(x) \geqslant 0.5 \\ 0, & \text{其他} \end{cases} \tag{12}$$

通过极大似然法，并加大优秀银行网点权重，构建如下损失函数，即

$$L(\theta) = \prod_{i=1}^{m} h_\theta(x_i)^{y_i} \cdot (1 - h_\theta(x_i))^{1-y_i} \cdot \frac{p^i}{\sum_1^m p^j} \tag{13}$$

$$p^i = \begin{cases} h_\theta(x_i), & \text{如果 } y^i = 1 \\ 1 - h_\theta(x_i), & \text{其他} \end{cases} \tag{14}$$

(13)—(14) 式中: p 为训练样本真实标签对应输出概率; y 为对应标签值，以此保证模型预测中优秀银

行网点所对应权重更高。

4) 将模型的预测概率作为网点选址方案的总评,并对其进行排序,设定阈值,低于该阈值的则不适合作为银行网点选址地点,选取排序前 k 个候选地点作为最终的银行选址地点。

3 实验结果分析

3.1 数据预处理

实验为消除因数据质量对选址推荐模型的影响,本文通过对数据清洗提供数据质量,主要方法为用平滑噪声方法替换极值和离散值,用均值填补缺失值等,对缺失值较多的数据,采用构建模型预测的方法填补缺失值。为消除数值大小和量纲的影响,并提高模型训练速度,对部分数据进行标准化处理,其标准化公式为

$$\hat{x}_i = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{15}$$

(14) 式中: x_i 为实际值; $\bar{x}$ 为均值; $\hat{x}$ 为标准化的值。

3.2 评价指标

本文通过 2 个方面评价体系评价模型的性能。

1) 精确率、召回率以及 $F1$ 值。以二分类问题为例,预测结果包含以下 4 种情况,4 种情况出现的总数分别记作:

TP(真正例)—将正类预测为正类数;

FN(假反例)—将正类预测为负类数;

FP(假正例)—将负类预测为正类数;

TN(真反例)—将负类预测为负类数。

则精确率也称查准率可定义为

$$precision = \frac{TP}{TP + FP} \tag{16}$$

召回率也称查全率可定义为

$$recall = \frac{TP}{TP + FN} \tag{17}$$

F_1 值是精确率和召回率的调和均值,其计算公式为

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{18}$$

由此可得分类模型的混淆矩阵,矩阵的每一列表示模型预测的样本情况; 每一行表示样本的真实情况。

2) 排序预测准确性 P_{Rank},即优秀网点对应权重更大,预测结果是否正确重要性更高。其通过专家对已有网点区域进行排序,并用以下公式计算预测准确性。

$$P_{\text{Rank}} = \sum_{i=1}^{N} w_i o_i \Big/ \sum_{i=1}^{N} w_i \tag{19}$$

$$w_i = \frac{1}{1 + \ln(i)} \tag{20}$$

(18)—(19) 式中: i 为网点的排序顺序; o_i 为预测的第 i 个网点的正确标志; 当预测正确时 o_i 为 1,否则 o_i 为 0。

3.3 结果分析

本实验数据集为通过重庆某移动运营商平台提供的移动信令数据、本地公交数据等构造的 500 条数据记录。其模型分为 2 层,第 1 层对应多个模型,其使用特征为通过数据分析工作获取的人流量、交通拥堵指数、人均收入、竞争网点数、区域用户价值等特征。第 1 层所使用模型为通过不同机器学习算法和不同数据集构造的差异度较大的模型算法,其中第 1 层所使用模型个数为 5 个,所使用机器学习算法包括梯度提升树、支持向量机、K 近邻等。表 2 为分模型的结果对比。

表 2 第 1 层分模型结果对比

Tab.2 Results of the first-level model

分模型	Precision	Recall	F_1
SVM	0.798 5	0.727 8	0.761
GBDT	0.805 1	0.784 8	0.795
KNN	0.773	0.734	0.753
LR	0.786 8	0.727 9	0.756
XGBOOST	0.801 2	0.791 1	0.796

第 2 层模型所使用特征为第 1 层分模型的输出概率,采用方法为基于优化损失函数的逻辑回归算法,并与传统逻辑回归算法,其他预测算法进行对比。表 3 为不同模型的结果对比。

表 3 模型结果对比

Tab.3 Results of the model

模型	*precision*	*recall*	F_1	S
本文模型	0.916 6	0.897 9	0.907	0.937
LR	0.919 5	0.921 9	0.925	0.903
GBDT	0.898 6	0.904 8	0.901	0.909

由表 2 与表 3 可知,基于分模型输出概率构建的融合算法模型相比直接使用基础特征的算法模型,预测精度有大幅提升。本文算法为基于优化损

失函数的逻辑回归算法，相较于传统逻辑回归算法，在牺牲一部分精确率和召回率的情况下，大幅提升了排序预测的准确性。

为进一步验证本文算法的有效性，本文通过使用不同概率特征个数的方法，与传统机器学习算法进行对比验证，其结果见图 8。由图 8 可知，当概率特征为 5 个时，模型趋于稳定，且本文基于损失函数优化后的逻辑回归算法相较于传统逻辑回归、其他机器学习算法，预测排序性有大幅提升。

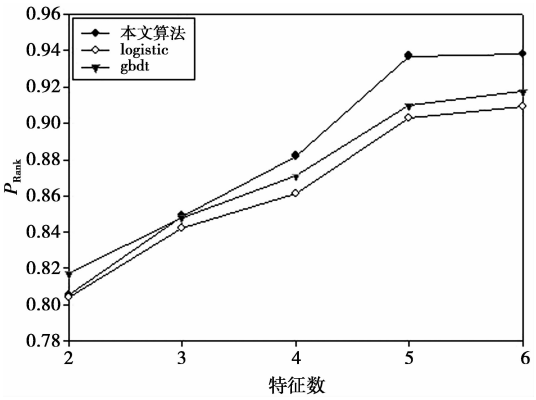

图 8　算法对比结果

Fig.8　Algorithmic comparison results

最后，本文以重庆市渝中区为例，以模型的输出概率作为选址方案总评，输出对应的选址方案。其结果如表 4，并将其可视化展示。选址结果可视化如图 9，其中，红色三角形对应区域为排序靠前的银行网点选址方案。

表 4　选址方案结果

Tab.4　Results of thesite selection

排序	坐标	人流量	用户价值	…	输出概率
Rank1	(29.573 79，106.532 26)	43 073	0.886	…	0.997
Rank2	(29.558 36，106.578 40)	81 323	0.745	…	0.986
Rank3	(29.557 36，106.567 97)	44 231	0.824	…	0.967
Rank4	(29.542 07，106.516 01)	38 320	0.847	…	0.942
Rank5	(29.562 19，106.590 58)	40 067	0.851	…	0.913

综上所述，本文算法在牺牲部分精确率和召回率的情况下，所得结果的排序预测准确性更高，在工程实际应用中，具备更高的价值意义，业务人员通过模型的预测结果，可以获得更加科学化、结构化的选址决策。

图 9　选址结果可视化

Fig.9　Visualization of site selection results

4　结束语

本文通过基于层次分析法构建区域用户价值特征，基于路权系数的方法构建交通拥堵指数特征。建立一种基于协同训练的半监督学习方法，利用大量未标记的样本作为训练数据，以提升模型预测性能。基于已有特征，构造 5 个分类模型作为第 1 层的分类器，并将第 1 层模型的输出概率作为第 2 层模型的特征，构建一种基于优化损失函数后的逻辑回归算法，使得优秀银行网点对应权重更高。通过实验表明，其具备更好的预测效果，具有一定的应用价值。本文认为，在模型融合与损失函数优化方面还有进一步的研究空间，未来期望在该方面展开进一步的研究。

参考文献：

[1] 丁建新.基于灰色关联分析法的商业银行网点选址研究[J].河北工业大学学报，2012，41(2)：110-113.
DING J X. Research on Commercial Bank Network Location Based on Grey Relational Analysis [J]. Hebei Journal of Industrial Science and Technology，2012，41(20)：110-113.

[2] SUN Z M，NIU L D，LI C S，et al. Site Selec-tion and Optimization Based on GIS for Old People's Homes in Urban Areaof Kunming City [C]//International Co-nference on Intelligent Transportation. Xiamen：IEEE，2018：368-372.

[3] WEN J J. Research on Site Selection of Logistics Park Based on Fuzzy Comprehensive Evaluation Method [C]// International Conference on Computer Engineering and Applications. Bali Island：IEEE，2010：44-47.

[4] THONGPUN A，NASOMWART S，PEESIRI P，et al. Decision support model for solar plant site selection [C]//International Conference on Smart Grid and Smart Cities. Singapore：IEEE，2017：50-54.

[5] LIU J，LI P，SHI T Y，et al. Optimal site selection of China railway data centers by the PSO algorithm [C]//

World Congress on Intelligent Control and Automation. Guilin: IEEE, 2016: 251-257.

[6] CHEN M Z, LIU J X, LI Z Z, et al. Research on site selection of rescue sites at sea based on NSGA II [C]//International Conference on Cloud Computing and Big Data Analysis. Chengdu: IEEE, 2017: 460-465.

[7] 胡伟,徐福缘,台德艺,等.基于改进粒子群算法的物流配送中心选址策略[J].计算机应用研究,2012,29(12):4489-4491.
HU W, XU F Y, TAI D Y, et al. Location Strategy of Logistics Distribution Center Based on Improved Particle Swarm Optimization [J]. Application Research of Computers, 2012, 29(12): 4489-4491.

[8] 邱晗光,李海南,宋寒.需求依赖末端交付与时间窗的城市配送自提柜选址—路径问题[J].计算机集成制造系统,2018,24(10):2612-2621.
QIU H G, LI H N, SONG H. Location-Routing Problem of Demand-Dependent End Delivery and Time Window for Self-Container in Urban Distribution [J]. Computer Integrated Manufacturing Systems, 2018, 24(10): 2612-2621.

[9] 陈久龙.基于BP神经网络的银行营业网点选址与评价研究[D].长春:长春工业大学,2016.
CHEN J L. Research on Location and Evaluation of Bank Business Network Based on BP Neural Network [D]. Changchun: Changchun University of Technology, 2016.

[10] 郑双怡,冯琼,郑智维.商业银行网点选址的熵权法[J].学习与实践,2017,23(12):75-81.
ZHENG S Y, FENG Q, ZHENG Z W. Entropy Weight Method for Commercial Bank Site Selection [J]. Study and Practice, 2017, 23(12): 75-81.

[11] 马江山,李昱莹.基于卷积的自助银行选址分析[J].深圳大学学报(理工版),2018,35(01):92-98.
MA J S, LI L Y. Location Analysis of Self-help Bank Based on Convolution [J]. Journal of Shenzhen University (Science and Engineering), 2018, 35(01): 92-98.

[12] TANG M N, REN E. Site selection of mechanical parking garage in high density vehicle urban area based on genetic algorithm-support vector machine [C]//International Symposium on Knowledge Acquisition and Modeling. Wuhan: IEEE, 2009: 100-102.

[13] CHEN C, LIU J M, LI Q, et al. Warehouse Site Selection for Online Retailers in Inter-Connected Warehouse Networks [C]//International Conference on Data Mining. New Orleans: IEEE, 2017: 805-810.

[14] PENG C. Suitability Evaluation of Urban Emergency Shelter Site Based on TOPSIS Evaluation Method [J]. Technology for Earthquake Disaster Prevention, 2017, 12(03): 700-709.

[15] 康凯,王小宇,马艳芳.不确定条件下配送回收中心选址配送问题研究[J].计算机工程与应用,2018,54(18):242-249.
KANG K, WANG X Y, MA Y F. Research on Location and Distribution of Distribution and Recycling Center under Uncertain Conditions [J]. Computer Engineering and Applications, 2018, 54(18): 242-249.

[16] 田立新,崔晓红,唐焕超.均差排序法在配送中心选址中的应用[J].江苏大学学报(自然科学版),2009,30(5):532-535.
TIAN L X, CUI X H TANG H C. Application of average subtraction arrangement method in distribution center location [J]. Journal of Jiangsu University (Natural Science Edition), 2009, 30(5): 532-535.

[17] 徐峰.银行网点选址因素的实证研究[D].杭州:浙江工业大学,2012.
XU F. An Empirical Study on the Location Factors of Banking Networks [D]. Hangzhou: Zhejiang University of Technology, 2012.

作者简介:

邓 轲(1993—),男,湖北武汉人,硕士研究生,研究方向为数据挖掘。E-mail: 645066606@qq.com。

冯辉宗(1972—),男,四川邻水人,教授,博士,主要研究方向为汽车电子控制系统建模仿真、ABS、发动机控制、电动汽车动力电机控制等。

许国良(1973—),男,浙江金华人,教授,博士,主要研究方向为光电传感与检测、通信网络设计与规划、大数据分析挖掘。

雒江涛(1971—),男,河南郑州人,教授,博士生导师,研究方向为移动大数据、新一代网络技术、通信网络测试与优化等。

(编辑:刘 勇)