计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(2):57−68 [doi: 10.15888/j.cnki.csa.008317]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

省域房产大数据热力图人工智能预测系统①

杨海涛[1], 孙庆辉[2], 吕建明[2], 阮镇江[1], 夏兰亭[1], 徐　飞[1]

[1](广东省建设信息中心, 广州 510055)
[2](华南理工大学 计算机科学与工程学院, 广州 510006)
通信作者: 杨海涛, E-mail: 992000247@qq.com

摘　要: 省域范围房产交易与登记大数据可视化呈现的建模分析预测对于研究我国城乡建设、区划经济的布局趋势, 呈现城镇建设发展指标的时空演化, 辅助支持科学决策、宏观调控等具有重要意义. 考虑到这些经济活动数据的预测建模涉及到尚无明确数学表达的、因素作用复杂的事物状态演变过程, 受近代人工智能深度神经网络技术在类似复杂场景成功应用的启发, 我们采用相关的长短时记忆网络模型 (LSTM) 与全连接层 (FC) 技术等 AI 技术, 建立起宏观可视化的省域房产大数据热力图预测系统. 本文的主要系统建设实践是, 利用所获的广东省域 (东沙群岛除外) 历年积累的房产法定业务大数据, 基于各市房屋建成年份时序, 实现对区域房产套数和面积等基本指标的年末地理热力图建模预测功能. 本文创造性提出“网格累计量预测+市域增量预测修正”的总体预测建模计算框架, 为省域房地产大数据人工智能建模预测增加了网格粒度调选和局部结合全局预测修正的调优途径, 提高了预测模型的适用性. 应用分析表明, 建模预测系统的计算结果具有较高的合理性和实用性.

关键词: 大数据热力图预测系统; 深度神经网络; 时序数据建模; 房产数据网格处理

引用格式: 杨海涛,孙庆辉,吕建明,阮镇江,夏兰亭,徐飞.省域房产大数据热力图人工智能预测系统.计算机系统应用,2022,31(2):57–68. http://www.c-s-a.org.cn/1003-3254/8317.html

Artificial Intelligence Prediction System of Big Data Heat Map for Provincial Realty

YANG Hai-Tao[1], SUN Qing-Hui[2], LYU Jian-Ming[2], RUAN Zhen-Jiang[1], XIA Lan-Ting[1], XU Fei[1]

[1](Guangdong Construction Information Center, Guangzhou 510055, China)
[2](School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract: The analysis and prediction modeling for the visual presentation of provincial big data concerning realty transaction and registration is of great significance for studying the layout trend of China’s urban-rural construction and regional economy. It shows the temporal and spatial evolution of urban construction and development indicators and supports scientific decision-making and macro-control. The prediction modeling of these economic activity data involves the understanding of the state evolution of things with complex factors and without clear mathematical expression. Thus, inspired by the successful applications of modern artificial intelligence-deep neural network technology in similar complex scenes, we intend to establish a macro visual prediction system of heat maps of provincial realty big data by related long short-term memory (LSTM) model and fully connected layer (FC) technology. The main system construction practice of this paper is that we utilize the big data from the legal business of realty which are accumulated in Guangdong Province (not including Dongsha Islands) over the years to implement the functions of modeling and predicting regional year-end geographic heat maps of two basic indicators, i.e., the number and the total area of existing realty units, regarding the temporal years when the realty was built in each city. This study creatively puts forward the overall prediction modeling and calculation framework of “grid cumulative prediction + incremental prediction correction for a

① 基金项目: 广东省应用型科技研发专项资金重点项目 (2015B010131012); 广东省自然科学基金 (2018A0303130022)
收稿时间: 2021-04-02; 修改时间: 2021-04-29, 2021-05-28; 采用时间: 2021-06-02; csa 在线出版时间: 2022-01-17

city". It increases the optimization options of grid granularity adjustment and local-global prediction correction for the artificial intelligence modeling and prediction of provincial realty big data and improves the applicability of the prediction model. Application analyses show that the calculation results of the modeling prediction system are reasonable and practical.

Key words: prediction system of big data heat map; deep neural network; temporal data modeling; realty data grid processing

1 背景和目标

1.1 课题背景

省域地理范围房产权利登记和交易活动中所形成的历年数据记录是反映我国经济社会活动的重要大数据基础资源. 我们在承担广东省应用型科技研发专项资金重点项目"省域房地产交易数据资源云同步及大数据规模化应用"过程中, 获得了海量的广东省域房地产交易法定业务实录大数据资源. 然而, 这些过往业务所产生的历史和现状大数据的直接使用, 只能发挥其档案查询、数据统计和现态监控的数据支持作用, 对于省级房地产主管部门进行房地产市场预警预报, 开展住房和房地产管理政策、产业发展和住房建设规划的研究和制定等并无前瞻性的帮助. 就我国房地产大数据应用意义而言, 省域区划是我国社会治理和政治经济特色的最大综合管治 (包括监管服务与行业调控) 单元. 特别是, 广东省作为我国第一经济大省, 2019 年全省实现地区生产总值 107 671.07 亿元 (仅低于世界排行第 12 名的韩国), 其中全年新增房地产开发投资 15 852.16 亿元[1], 加上城镇与房产为依托的各行各业的经济产值则总量更为巨大. 因此在宏观层面研究广东省域房产大数据并深化其应用具有重要的现实意义. 本文拟通过建立基于人工智能的系统平台, 全程实现对既有积累的海量房产法定业务大数据资源做可视化呈现并面向未来进行建模预测, 以探索实现直观地显示预示广东省域城乡建设、城镇发展的某些重要指标 (如房产或房屋建筑面积和套数) 的时空演化过程, 为研究广东省域城乡建设、区划经济的布局趋势, 科学有效地辅助支持各相关城市开发建设管理决策和省域房地产市场宏观调控等工作服务.

1.2 前期工作

在"十三五"期间, 我们一直从事广东省域房产大数据相关工作, 具备了进一步开展房产大数据深度智慧应用的基础.

1.2.1 房产大数据基础平台及数据资源开发建设

建立"(房产) 行业数据云同步枢纽平台系统": 实现了可覆盖全省各市房地产交易登记数据的同步归集. 提供了包括同步系统节点规划管理、安装配置, 分块流水线处理、单和双向同步 (全量/增量, 乐观/谨慎校验策略, 同步块及分组调适)、并发控制、指标映射、多属性主键归一、敏感字段 Hash, 以及同步正确性保障等较齐全的同步枢纽平台功能.

建立"HBDP (housing big data platform) 省房屋大数据计算集群及作业调度系统"[2]: 集群采用 Hadoop 分布式文件存储架构, 选用 Hive 管理元数据, 供用户利用 Spark SQL 进行房屋大数据分布式交互分析, 参见图 1.

归集省域房产交易登记数据资源: 形成全省房产交易与产权管理数据大字典 1 137 页 (省和各市卷合编), 入库各市 4 108 个原始表、92 871 列字段、9.07 亿条记录, 约 203.7 GB 数据量. 析出房产单元约 1 551.4 万套 (其中, 住宅 831.8 万套, 有房屋建成年份的约 1 200 万套)、产权人 827 万. 梳理 8 市网签和预售系统楼盘表共 935.4 万户和 22 790 个房地产项目的数据.

1.2.2 海量地址关联数据的热力图渲染优化研发

地址关联数据的地理分布热力图呈现通常是依托基于位置服务 (location-based service, LBS) 公共平台进行二次开发实现. 现有方案是将所有地址关联数据在本地进行整理后按照固定的格式全部上传至 LBS 服务提供商的远程服务器, 由远程服务器处理后返回本地进行呈现. 但是, 在处理海量数据时, 该方案由于互联网带宽和 PC 浏览器处理能力的限制, 实际响应慢, 用户体验差. 对此, 我们通过实践探索, 创造性实现了"一种地址关联数据处理方法、用户终端和服务器"[3], 它依据用户选择的需求参数和应用的地图属性, 将输入的海量地址关联数据划分成即时渲染 (第 1 组) 和延缓渲染 (第 2 组) 两组数据, 使得第 1 组的数

据能迅速在客户端呈现, 同时第 2 组数据的落图渲染计算在本地服务器 (集群) 同步进行. 这种前后两组计算结合能显著地提高海量地址关联数据落图渲染呈现的响应速度, 改善了相关应用的用户体验, 解决了一般 PC 端网页浏览器在交互式播放海量地址关联数据渲染所普遍遇到的显示滞卡问题. 具体参见图 2.

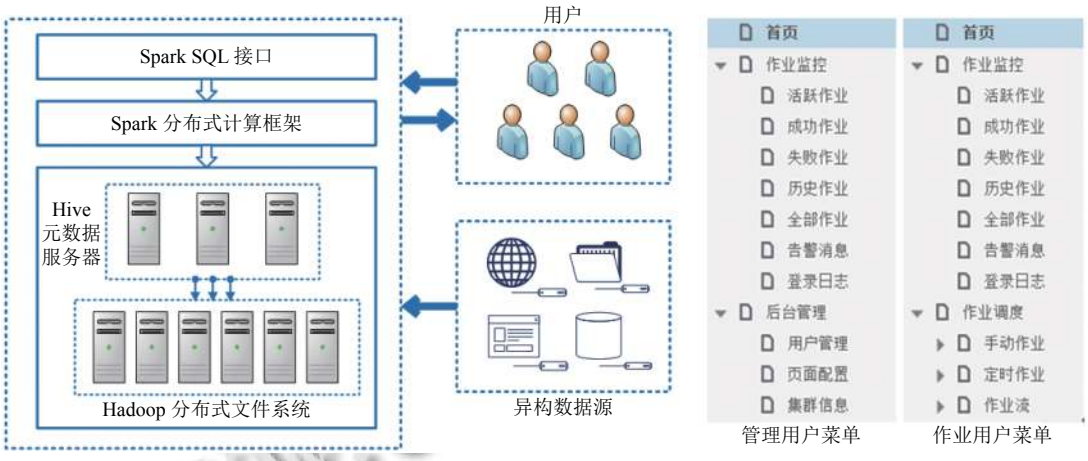


图 1 HBDP 省房屋大数据计算集群及作业调度系统

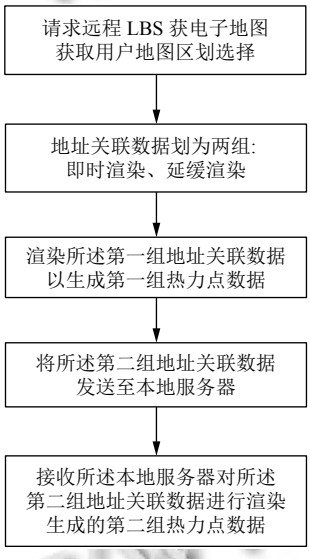


图 2 海量地址关联数据渲染优化处理框图

用, 具体沿继使用习惯.

第二部分是热力图播放和预报部分, 它由“热力图播放系统”和“人工智能预测系统”构成. 其中, “人工智能预测系统”为本文研发的重点, 它提供人工智能建模预测功能, 负责使用已有的热力值数据训练模型并产生预测输出.

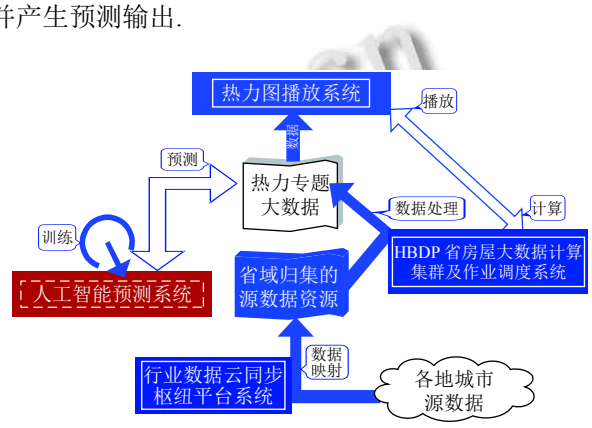


图 3 省域房产大数据热力图系统总体构成

1.3 课题目标

1.3.1 省域房产大数据热力图系统总体构成

系统总体由两大部分构成, 参见图 3.

第一部分是基础设施部分, 它由负责数据归集和专题数据生产的系统构成: 首先由“(房产) 行业数据云同步枢纽平台系统”同步归集广东省域各地城市源数据, 然后, 经由“HBDP 省房屋大数据计算集群及作业调度系统”计算生成适用于热力图渲染播放的热力值专题大数据对象. 本文术语“房屋”与“房产”可相互通

1.3.2 关键研发任务—人工智能预测系统

在前述总体框架, 待研发实现的是“人工智能预测系统”. 而该系统的关键在于如何有效地实现“省域房产大数据热力值预测计算”的核心课题任务上. 基于我们所能掌握的广东省域房产大数据资源限于 2018 年之前的原始数据, 我们将本文主要课题任务的实质性研发内容具体地明确为: 如何利用人工智能算法对既

获得的已知的广东省各市截至 2017 年的房产登记和交易业务等历史记录的指标数据 (房产套数和面积) 的热力值, 建立可重复使用的时序预测模型系统, 特别是, 近期实现预测后来 2018–2023 年指标数据热力值; 据此, 为将来具备条件 (获得后继年份数据资源) 时, 滚动地推广应用至更多的未知数据年份.

2 技术问题归结与预测建模设计

针对上节提出的主要课题任务, 我们首先从计算机系统开发建设与应用的角度出发, 细化和明确所拟要研究解决的实质技术问题及其技术路线. 具体地, 本文课题的核心研发内容主要包括以下 3 方面: 1) 关键指标及其计算问题的归结; 2) 总体计算处理框架设计; 3) 时序预测模型设计与实现.

2.1 问题归结

(1) 房产热力值的定义

所谓热力图就是关于地理区域单元上的计算指标的值即“热力值”的地图渲染. 具体地, 本课题所研究的房产热力值是关于一个地理区域内的房产指标 (套数和面积) 的统计量的数值.

(2) 时序数据计算任务定义

本文处理的房产指标热力值是具有时间属性的, 从而可构成时序序列. 相应的时序数据计算任务包括: 一是由原始房产业务记录导出可直接计算房产热力值的房产单元记录; 二是由已知的一系列年份的房产单元记录数据集计算出对应地理区域的房产指标 (套数和面积) 的热力值 (可直接在百度地图上标识渲染成热力图); 三对地理区域后继年份的未知房产指标热力值进行预测计算. 为简化起见, 只对套数和面积两个基本指标进行预测.

鉴于房产数据的变化规模与频度的实际情况, 本文所考虑的时序单位明确为“年份”.

原始房产业务记录化为如下房产单元记录格式:

<房产单元>(经度, 纬度, 建成年份, 套数|面积, 城市)

其中, 经度和纬度坐标是用房产单元坐落地址调用“百度开放平台”Web 服务 API 的地理编码服务 (又名 Geocoder)[4] 获得. 例如, 以房产单元坐落 “广州市越秀区豪贤路 102 号”调用百度 Geocoder, 可获得经纬度坐标{113.281 270 035 554 5, 23.136 617 015 096 8}.

我们所讨论的房产热力值是关于具体地理区域内的所有<房产单元>个数 (称作“套数”) 或面积的统计值的数量指标. 欲将房产热力值落在百度地图上渲染显示, 就必须将其与地理区域的坐标相关联. 为简单起见, 我们采取地理区域中心点的经度和纬度来标识地理区域, 从而有如下的房产指标热力属性关系:

<区域房产热力>(经度, 纬度, 年份, 热力值, 城市)

其中, “热力值”是该 (经度, 纬度) 所标识区域截止于“年份”的期末实有房产单元的“套数|面积”统计数 (假设房产房屋建成后一直存在, 则它代表历年直至该“年份”期末的累计数, 以下均采用此假设).

(3) 时序计算区域的网格化处理

为确定最基本的地理区域单元, 也为了细化计算处理、减少随机噪音影响, 我们将广东省全域 (东沙群岛除外) 分成$M\times M$个矩形区域 ($M>0$), 称作“M 分网格”. 每个网格区域可用经其左下角点的网格线的行号和列号来唯一标识: 网格(x,y)代表以第 x 列、第 y 行 ($0\leqslant x<M$, $0\leqslant y<M$) 网格线的交点为其左下角的矩形网格区域.

网格(x,y)区域的年房产热力值 (套数|面积) 按房产房屋建成年份排列构成如下时序:

1) 网格(x,y)房产套数累计时序:

$$\boldsymbol{T}_{x,y}=\{\boldsymbol{T}_{x,y,1},\ \boldsymbol{T}_{x,y,2},\ \cdots\}$$

其中, $\boldsymbol{T}_{x,y,i}$为第 i 年(x,y)区域累计房产套数.

2) 网格(x,y)房产套数增量时序:

$$\Delta\boldsymbol{T}_{x,y}=\{\Delta\boldsymbol{T}_{x,y,1},\ \Delta\boldsymbol{T}_{x,y,2},\ \cdots\}$$

其中, $\Delta\boldsymbol{T}_{x,y,i}$为第 i 年(x,y)区域新增房产套数, 可由下式计算:

$$\Delta\boldsymbol{T}_{x,y,i}=\boldsymbol{T}_{x,y,i}-\boldsymbol{T}_{x,y,i-1}$$

3) 网格(x,y)房产面积累计时序:

$$\boldsymbol{A}_{x,y}=\{\boldsymbol{A}_{x,y,1},\ \boldsymbol{A}_{x,y,2},\ \cdots\}$$

其中, $\boldsymbol{A}_{x,y,i}$表示第 i 年(x,y)区域累计房产面积.

4) 网格(x,y)房产面积增量时序:

$$\Delta\boldsymbol{A}_{x,y}=\{\Delta\boldsymbol{A}_{x,y,1},\ \Delta\boldsymbol{A}_{x,y,2},\ \cdots\}$$

其中, $\Delta\boldsymbol{A}_{x,y,i}$表示第 i 年(x,y)区域新增房产面积, 可由下式导出:

$$\Delta\boldsymbol{A}_{x,y,i}=\boldsymbol{A}_{x,y,i}-\boldsymbol{A}_{x,y,i-1}$$

进一步, 在更大的尺度上, 对于某个城市 c, 我们得到市级房产热力值统计值序列如下:

1) 市级房产套数累计时序:

$$\boldsymbol{T}^{c}=\{\boldsymbol{T}^{c}_{1},\ \boldsymbol{T}^{c}_{2},\ \cdots\}$$

其中, $\boldsymbol{T}_i^c$表示第 i 年城市 c 累计房产套数.

2) 市级房产套数增量时序:

$$\Delta\boldsymbol{T}^c = \{\Delta\boldsymbol{T}_1^c, \Delta\boldsymbol{T}_2^c, \cdots\}$$

其中, $\Delta\boldsymbol{T}_i^c$表示第 i 年城市 c 新增房产套数, 可由下式导出:

$$\Delta\boldsymbol{T}_i^c = \boldsymbol{T}_i^c - \boldsymbol{T}_{i-1}^c$$

3) 市级房产面积累计时序:

$$\boldsymbol{A}^c = \{\boldsymbol{A}_1^c, \boldsymbol{A}_2^c, \cdots\}$$

其中, $\boldsymbol{A}_i^c$表示第 i 年, 城市 c 累计房产面积.

4) 市级房产面积增量时序:

$$\Delta\boldsymbol{A}^c = \{\Delta\boldsymbol{A}_1^c, \Delta\boldsymbol{A}_2^c, \cdots\}$$

其中, $\Delta\boldsymbol{A}_i^c$表示第 i 年城市 c 新增房产面积, 可由下式导出:

$$\Delta\boldsymbol{A}_i^c = \boldsymbol{A}_i^c - \boldsymbol{A}_{i-1}^c$$

本课题的基本预测计算任务可归结为两个:

预测 1: 给定网格区域(x,y)的房产套数 n 年时序数据$\{\boldsymbol{T}_{x,y,1}, \boldsymbol{T}_{x,y,2}, \cdots, \boldsymbol{T}_{x,y,n}\}$, 预测下一年 (第 n+1 年) 该区域累计房产套数$\boldsymbol{T}_{x,y,n+1}$.

预测 2: 给定网格区域(x,y)的房产面积 n 年时序数据$\{\boldsymbol{A}_{x,y,1}, \boldsymbol{A}_{x,y,2}, \cdots, \boldsymbol{A}_{x,y,n}\}$, 预测下一年 (第 n+1 年) 该区域累计房产面积$\boldsymbol{A}_{x,y,n+1}$.

显见, 通过逐年向前移动时序数据, 就可实现: 利用过去一段时间房产数据的变化, 对未来一段时间内的房产数据进行预测.

2.2 总体处理框架设计

根据上节的分析, 课题的基本科学技术问题在于建立时序预测模型: 利用过去一段时间内某事件 (房屋建成事件) 的时间特征来预测未来一段时间内该事件的特征 (房产套数或面积)—这种时间序列数据预测.

鉴于房产套数和建筑面积的预测的建模都是类同的, 本文仅需阐述房产套数的预测模型.

实践上, 我们先直接对广东省域 (由于东沙群岛无房地产项目, 本文研究的广东省域不包括东沙群岛) 各网格单元应用深度神经网络建立预测模型. 结果所产生的预测误差普遍过大, 且无法调优模型将误差降到合理程度. 这是因为, 全省各市房地产源数据集的房产热力指标值的地理空间和时间区间的分布相当不均匀, 并且在时间和空间上存在不同程度的数据样本不足. 因此, 我们提出“网格累计量预测+市域增量预测修正”的总体预测建模计算框架. 具体工作思路如下.

1) 对广东省域地图做 M 分网格, 将其中的任意网格看作独立单元, 运用基于深度神经网络的预测模型对其房产套数时序数据进行独立预测, 获得其下一时序点套数的预测值.

2) 考虑到实际上同一城市的不同区域间受共同的城市发展内在关系影响, 彼此间应存在某些关联或约束, 我们在模型中进一步引入同城数据约束修正来提高预测结果的合理性和准确性.

总体处理框架如图 4 所示. 具体过程如下.

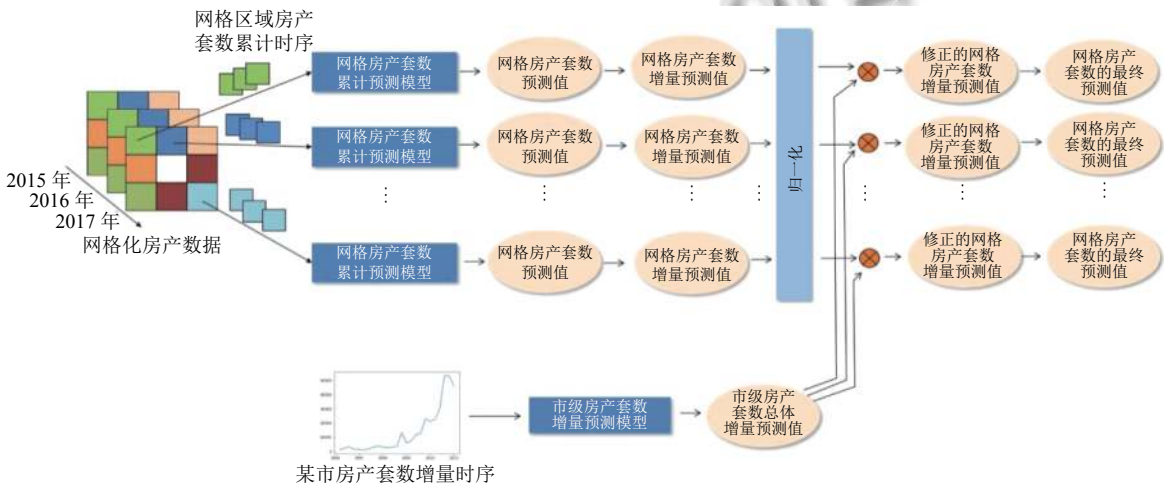


图 4 房产套数预测总体处理框架

1) 市域网格预测

将同属一个城市的所有网格区域筛滤出来, 然后将每个网格区域(x,y)的房产累计套数年份时序$\{\boldsymbol{T}_{x,y,1}, \boldsymbol{T}_{x,y,2}, \cdots, \boldsymbol{T}_{x,y,n}\}$输入到基于深度神经网络的网

格房产套数累计预测模型. 该预测模型输出下一个时序年份该区域房产套数累计$\boldsymbol{T}_{x,y,n+1}$的预测值$\overline{\boldsymbol{T}}_{x,y,n+1}$. 根据$\overline{\boldsymbol{T}}_{x,y,n+1}$, 计算出对应的房产套数增量的预测值:

$$\overline{\Delta \boldsymbol{T}}_{x,y,n+1} = \overline{\boldsymbol{T}}_{x,y,n+1} - \boldsymbol{T}_{x,y,n}$$

对同属城市 c 的每个区域(x,y)的房产套数增量预测值进行归一化, 得到:

$$Norm(\overline{\Delta \boldsymbol{T}}_{x,y,n+1}) = \frac{\overline{\Delta \boldsymbol{T}}_{x,y,n+1}}{\sum_{(u,v)\in\Omega_c} \overline{\Delta \boldsymbol{T}}_{u,v,n+1}}$$

其中, Ω_c表示所有属于城市 c 的网格区域(u,v)的集合.

2) 市域全局修正

用同城全域数据作为约束对$\overline{\Delta \boldsymbol{T}}_{x,y,n+1}$值进行修正. 为此, 先计算出该市全域房产套数增量时序$\{\Delta \boldsymbol{T}_1^c, \Delta \boldsymbol{T}_2^c, \cdots, \Delta \boldsymbol{T}_n^c\}$, 将该时序输入到基于深度神经网络的市级房产套数增量预测模型, 得到下一个时序年份该市房产套数增量$\Delta \boldsymbol{T}_{n+1}^c$的预测值$\overline{\Delta \boldsymbol{T}}_{n+1}^c$. 以此同城全域增量预测值$\overline{\Delta \boldsymbol{T}}_{n+1}^c$作为全局约束, 对该市网格区域$(x,y)$的房产套数增量预测值$\overline{\Delta \boldsymbol{T}}_{x,y,n+1}$做修正:

$$\overline{\overline{\Delta \boldsymbol{T}}}_{x,y,n+1} = \overline{\Delta \boldsymbol{T}}_{n+1}^c \cdot Norm(\overline{\Delta \boldsymbol{T}}_{x,y,n+1})$$

由此修正后的房产套数增量的预测值, 得到该网格区域的累计房产套数最终预测值:

$$\overline{\overline{\boldsymbol{T}}}_{x,y,n+1} = \overline{\overline{\Delta \boldsymbol{T}}}_{x,y,n+1} + \boldsymbol{T}_{x,y,n}$$

实际操作上, 整个工作路线包括如下基本步骤:

1) 数据预处理: 对原始数据进行清洗、网格映射、序列提取、数据规范化, 获得网格区域房产套数累计时序、网格区域房产面积累计时序、市级房产套数增量时序、市级房产面积增量时序.

2) 使用网格区域房产套数累计时序, 训练图 4 中的网格房产套数累计预测模型[5,6].

3) 使用市级房产套数增量时序, 训练图 4 中的市级房产套数增量预测模型.

4) 使用网格区域房产面积累计时序, 训练网格房产面积累计预测模型 (与套数预测模型类似, 省略).

5) 使用市级房产面积增量时序, 训练市级房产面积增量预测模型 (与套数预测模型类似, 省略).

6) 应用图 4 所示的预测过程对 2018–2023 年的房产套数进行预测, 输出规定格式的数据.

7) 对 2018–2023 年的房产面积进行预测, 输出规定格式的数据 (与步骤 6) 类似, 省略).

训练数据资源限于 2018 年之前的原始数据.

2.3 时序预测模型设计

前节的总体处理框架设计关键在于图 4 中的网格房产累计套数 (或面积) 预测模型和市级房产套数 (或面积) 增量预测模型设计—它们都是基于深度神经网络的时序预测模型设计. 众所周知, 在人工智能深度学习算法中[7,8], 正如卷积神经网络主要用于图像数据建模, 循环神经网络 (recurrent neural networks, RNN) 主要用于时序数据建模. 但是, 传统的 RNN 在长期依赖方面存在梯度消失的问题, 也就是会遗忘时间序列距离比较远的信息. 1997 年, Hochreiter 和 Schmidhuber 提出了传统 RNN 的一种变形: 长短时记忆网络 (long short term memory, LSTM)[9]. LSTM 通过引入 3 种门限 (遗忘门限、输入门限和输出门限) 而获得学习长期依赖的能力, 即具有学习时间序列距离较远信息的能力. 尽管 LSTM 相对于 RNN 更加复杂, 但因为它可以适应更长的时间序列数据, 我们用 LSTM对已知的过往年份房产热力数据进行时序特征抽取, 并对这些特征进行时序预测[10], 然后用全连接层 (fully connected layers, FC)[11] 神经网络再将时序预测得到的特征数据回归映射成房产指标热力数据.

图 4 中的网格房产累计套数 (或面积) 预测模型和市级房产套数 (或面积) 增量预测模型均采用如图 5 所示的 (LSTM→FC) 房产数据时序预测模型设计: 先用 LSTM 对输入的 n 年份时序数据$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$进行特征抽取和预测, 输出该时序数据隐含的时序特征的抽象表达向量的预测值 (不直接是房产指标热力值本身); 然后, 将 LSTM 模型的输出向量作为后面 FC 的输入, 经由 FC 做非线性变换后, 输出下一年份 (第 n+1 年) 的房产热力值$\boldsymbol{X}_{n+1}$的预测值$\hat{\boldsymbol{X}}_{n+1}$.

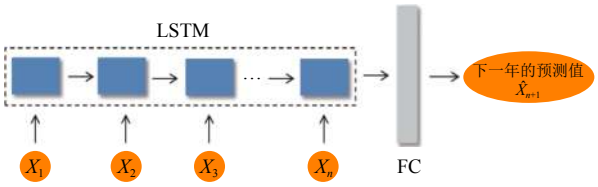


图 5 房产数据时序预测模型

2.4 预测建模技术实现

(1) 技术选型

我们选择在谷歌公司的 TensorFlow 开源机器学习平台上实现预测建模[12]. 在 TensorFlow 平台上, 我们

使用 Keras[13] 这种直观的高阶 API 来构建和训练机器学习模型, 这样能够快速迭代模型并轻松地调试模型.

鉴于为省域房产大数据热力图预测而构建和训练机器学习模型需要频繁地进行大规模 CPU-heavy 的张量数值计算, 一般 CPU 难以承担. 于是, 我们配置强力的加速计算协处理器, 即引入具有比 CPU 更加强大密集数据计算能力的 GPU 来参加计算. 但是, 传统的访问 GPU 模式 (如依赖图像 API 接口来实现 GPU 访问) 无法将 GPU 强大的密集数据计算能力用于图像处理之外的用途. NVIDIA 发明的 CUDA (compute unified device architecture) 编程模型采用了一种全新的计算体系结构来使用 GPU 硬件资源, 可让软件开发者在应用程序中能充分地利用 CPU 和 GPU 各自的优点, 特别是充分利用 GPU 强大的计算能力加速大规模密集数据计算[14]. NVIDIA 已将 CUDA 实现成一套实用编程环境, 并且可通过对应的 SDK 集成到更高级别的机器学习框架中. 其中, cuDNN 就是 CUDA 的一个专门用于 TensorFlow 神经网络运算加速的 SDK[15]. 因此, 我们采用 TensorFlow+CUDA+cuDNN 的开发环境来运行调试深度神经网络.

(2) 算法部署

TensorFlow 平台选择 Python 作为表达和控制模型训练的语言. Python 是数据科学家和机器学习专家用的最舒适的高级语言, 籍用其 NumPy[16] 库的丰富计算资源, 可以很容易地在 Python 中进行各种海量数据高性能预处理计算, 然后输送给 TensorFlow 进行真正 CPU-heavy 计算. 所以我们采用 Python 语言进行编程. 具体环境部署上, 使用 Anaconda 安装和管理 Python 相关包—Anaconda 提供最便捷的方式来使用 Python 进行数据科学计算和机器学习[17]. 算法部署主要包括模型训练和生成数据两部分.

模型训练部分的目录部署: data 文件夹: 存放预处理后的数据, model 文件夹: 存放训练好的模型; Python 可执行代码文件“网格累计 (套数|面积) 模型训练.py”用于网格累计 (套数|面积) 模型的训练, “登记单市模型训练.py”与“网签单市模型训练.py”分别用于对单个城市的登记与网签数据增量模型进行训练.

生成数据部分的目录部署: Python 可执行代码文件“登记面积数据生成.py”“登记套数数据生成.py”“网签面积数据生成.py”和“网签套数数据生成.py”分别对登记数据面积、登记数据套数、网签数据面积、网签数据套数进行预测, 并会将结果分别保存在 dj_area (登记面积)、dj_count (登记套数)、wq_area (网签面积)、wq_count (网签套数) 目录下.

(3) 训练预测模型

训练好的模型存放在 model 文件夹下. 如果一个市的数据发生大幅变动, 则需重新训练. 训练命令语法格式为:

Python + 运行脚本名称 + *argv*1 + *argv*2 + *argv*3

其中, *argv*1 代表预处理后的数据集名, *argv*2 代表城市名称, *argv*3 代表热力指标名称 (套数或面积). 若训练过程误差不下降, 则需重新运行脚本—可能是模型参数随机初始化或者训练样本过少造成的.

(4) 生成预测数据

在数据生成代码所在目录, 使用 Python 命令:

Python + 运行脚本名称 + *argv*1

其中, *argv*1 代表预处理后的数据集名. 运行结果除在指定的文件夹下生成套数|面积热力图预测数据的 JSON 格式文件外, 还在当前目录下生成相应的 CSV 格式文件. JSON 格式的输出数据可直接用于热力图渲染, CSV 格式的输出数据便于做大数据分析.

3 预测模型应用

3.1 处理模式和建模设定

3.1.1 数据压缩落图

原始房产数据极其庞大, 难以全部在百度地图上落图显示, 必须先进行压缩映射预处理并生成相应的训练数据集后, 才能应用预测模型对网格区域累计房产热力值 (套数、面积) 进行预测.

压缩映射: 将原始数据记录的经纬度坐标精度降低 (将经纬度小数点后 13 位有效数字四舍五入至小数点后仅保留 5 位有效数字), 然后去掉重复值. 本课题所有房产地址的经纬度坐标均取自百度地图公开数据, 原始精度为小数点后 13 位有效数字.

“房产登记簿数据”共有 12 476 111 条房屋建成年代记录. 压缩映射得到 353 464 条不同经纬度坐标的记录, 每条记录包含聚集计算出来的热力值属性 (套数|面积).

“房产网签数据”共有 6 076 778 条新建房屋年份记录. 压缩映射得到 50 242 条不同经纬度坐标的记录.

3.1.2 数据网格映射

基于经纬度网格的房产指标数据热力图必须先将

落在同一个网格中的各个房产原始数据记录归化成网格数据: 用网格中间点的经纬度坐标作为网格坐标, 网格区域中所有房产的指标统计值 (如总建筑面积或总房产套数) 作为网格的房产指标值.

再进一步将网格数据规范化, 即, 将所有网格数据映射成一个$Y\times M\times M$点的张量$\boldsymbol{Z}$ (Y为房屋建成年份的最大时间跨度, M为网格划分数):

对于任一条略去“城市”属性的网格房产数据记录$(lon_q, lat_q, year_q, count_q)|1\leqslant q\leqslant m$, m 为网格数据记录总数, (lon_q, lat_q)为对应网格的经纬度, $count_q$为该网格区域在年份$year_q$的房产统计指标增量值, 则其到张量$\boldsymbol{Z}$的映射由式 (1)–式 (3) 计算:

张量 $\boldsymbol{Z}$ 的第 1 维度坐标k可通过式 (1) 求得:

$$k = year_q - y_{\min} \tag{1}$$

其中, $y_{\min}$代表所有记录年份中的最小值.

张量 $\boldsymbol{Z}$ 的第 2 维度横坐标i可通过式 (2) 求得:

$$i = \frac{(lon_q - lon_{\min})\times M}{lon_{\max} - lon_{\min}} \tag{2}$$

其中, $lon_{\min}$和$lon_{\max}$分别代表广东省经度范围 (东沙群岛除外) 的最小值和最大值.

张量 $\boldsymbol{Z}$ 的第 3 维度纵坐标j可通过式 (3) 求得:

$$j = \frac{(lat_q - lat_{\min})\times M}{lat_{\max} - lat_{\min}} \tag{3}$$

其中, $lat_{\min}$和$lat_{\max}$分别代表广东省纬度范围 (东沙群岛除外) 的最小值和最大值.

最后, $\boldsymbol{Z}$任意点(k,i,j)的数值$\boldsymbol{Z}_{k,i,j}$由下式计算:

$$\boldsymbol{Z}_{k,i,j} = \begin{cases} \boldsymbol{Z}_{k-1,i,j} + count_q, & k>0 \\ count_q, & k=0 \end{cases}$$

这样, $\boldsymbol{Z}_{k,i,j}$ ($0\leqslant k\leqslant Y$, $0\leqslant i<M$, $0\leqslant j<M$) 代表网格(i,j)里累计至第k年末的房产套数或面积统计值.

3.1.3　网格数目选择

网格越细化 (网格划分 M 越大), 网格化后的坐标点越多, 反之亦然.

首先, 我们选定广东省全域 (东沙群岛除外) 经、纬度范围:

$[lon_{\min}, lon_{\max}]$= [110.177, 116.885]

$[lat_{\min}, lat_{\max}]$= [20.334, 25.313]

然后, 就省域房产登记簿数据集, 计算网格划分数 M 与房产数据集网格化后其坐标点数目的关系, 结果见图 6.

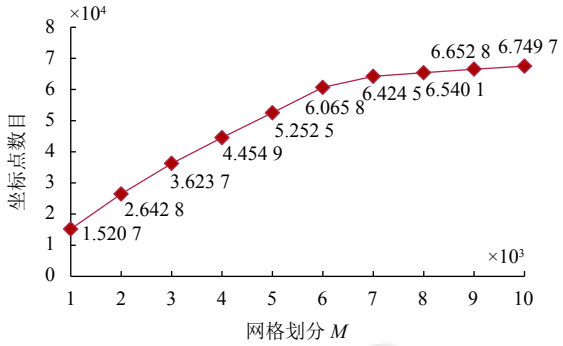


图 6　网格划分 M 与网格化后坐标点数目的关系

由图 6 可见, 网格化后数据集经纬度坐标点数目随着 M 的增大而增加, 但当 M 达到 7 000 后, 坐标点数目增长逐渐减弱, 说明: 此后网格划分继续细化对于网格化近似精度的提高其作用逐渐趋无. 若在获得较高的网格化近似精度的同时, 又兼顾算法的性能 (网格化后坐标点越少就越好), 推荐 $M=7000$ 的网格化.

3.1.4　滑窗切分处理

对于长的时序数据序列一般采用滑窗法进行数据切分预测, 即限定一个滑动窗口, 依次将其顺时序向前移动一定步长来切取数据子序列用以预测后继若干时序值: 对于已知数据序列$(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N)$, 取滑窗跨度 $L(L<N)$, 滑动步长 S, 初始切取$(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_L)$来预测$\boldsymbol{X}_{L+1}, \boldsymbol{X}_{L+2}, \cdots, \boldsymbol{X}_{L+P}$ $(P<L)$, 向前滑动第 1 步, 切取$(\boldsymbol{X}_{1+S}, \boldsymbol{X}_{2+S}, \cdots, \boldsymbol{X}_{L+S})$预测$\boldsymbol{X}_{L+S+1}, \boldsymbol{X}_{L+S+2}, \cdots, \boldsymbol{X}_{L+S+P}$, 接着滑动第 2 步, 重复前步类似的过程, 如此滑动预测, 直至第 k 步$(L+k\times S+1)\leqslant N\wedge(L+(k+1)\times S+1)>N$.

鉴于要预测的年末实有房产套数和面积是不断增长的数值, 为降低解空间的取值范围, 我们将滑窗内 L 个时点序列$(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_L)$各项数据规范化为相对于其第 1 时点的增量值: $(\boldsymbol{X}_1-\boldsymbol{X}_1, \boldsymbol{X}_2-\boldsymbol{X}_1, \cdots, \boldsymbol{X}_L-\boldsymbol{X}_1)$, 以加快神经网络学习过程收敛.

3.1.5　建模参数取值

根据以上设计建立模型后, 我们针对本课题应用实际进行优化. 经测试后确定:

1) “网格房产累计预测模型”(参见图 4 中的“网格房产套数累计预测模型) 的参数如表 1 所示.

2) “市级房产增量预测模型”(参见图 4 中的“市级房产套数增量预测模型) 的参数如表 2 所示—只对全市域总增量值进行预测, 不用做网格划分.

3.2　房产数据预处理

通常不是所有的原始数据都可直接用于神经网络

模型. 在此数据预处理是必须的.

3.2.1 数据清洗

数据清洗过程包括对缺失值、异常值进行处理.

表 1 网格房产累计数据建模参数

参数符号	参数意义	参数取值
Y	训练数据跨度 (年份)	24
M	网格划分数	7 000
L	滑窗跨度 (年份)	10
S	滑动步长 (年份)	1
P	预测年份长度	1
LSTM_out_dim	FC输入向量维度	128
Output	模型输出数据维度	1

表 2 市级房产增量数据建模参数

参数符号	参数意义	参数取值
Y	训练数据跨度 (年份)	各市取值范围不同
L	滑窗跨度 (年份)	3
S	滑动步长 (年份)	1
P	预测年份长度	1
LSTM_out_dim	FC输入向量维度	128
Output	模型输出数据维度	1

缺失值处理: 对于缺失房屋建成年份、坐落地址或面积数值的数据, 直接舍弃, 因为这些数据对于时序预测没有意义; 对于缺失城市属性的数据, 可以根据房产坐落确定地理位置, 不会对预测结果产生影响.

异常值处理: 主要针对面积数值为零的原始数据—对此类数据记录直接舍弃.

3.2.2 数据集准备

深度学习算法需要切取历史时序数据进行训练和测试. 全部房产数据记录经网格数据映射存储在三维张量$\boldsymbol{Z}$中. 令Ω_c为城市 c 所有网格的集合, 则从$\boldsymbol{Z}$中切取: 1) 矩阵序列$(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_Y) \mid \boldsymbol{Z}_k = \{\boldsymbol{Z}_{k,i,j} | (i,j) \in \Omega_c\}$ $(1 \leqslant k \leqslant Y)$, 2) 数据序列$(\boldsymbol{Z}_{1,i,j}, \boldsymbol{Z}_{2,i,j}, \cdots, \boldsymbol{Z}_{V,i,j}) | \boldsymbol{V} = \max(l)$, $\forall \boldsymbol{Z}_{l,i,j} \neq \text{null}(1 \leqslant l \leqslant Y), \forall (i,j) \in \Omega_c$作为预测模型训练与测试的候选数据集.

对于每一份数据集来说, 均需要划分训练集与测试集. 训练集是模型学习时用的数据集, 是确定模型参数用的; 测试集则是检验模型性能时用的数据集. 在时序数据预测问题中, 训练集与测试集不能交叉. 训练集就类似考生平常做的习题, 测试集类似考试的题目, 后者是衡量一个模型泛化能力的数据集.

若测试集序列太短, 将不足以评价和调校预测模型. 结合各城市房产单元数据集的房屋建成年份时序范围的实际 (详见表 3), 我们取每个市至少最后 5 个年份的数据作为其评价模型的测试集, 即要具备最近 5+3 年 (加上 3 年滑窗跨度) 的原始数据. 这样绝大多数城市数据满足要求, 只有个别城市例外—在表 3 中用*标注: 佛山市房产登记簿仅有 2012–2016 年的数据; 梅州市、惠州市和茂名市的网签数据分别只有 6、7 和 6 个年份的.

表 3 各市房屋建成年份数据范围

数据类型	城市	起始年份	终止年份	城市	起始年份	终止年份
登记簿数据	广州	1958	2017	中山	2001	2015
	深圳	1985	2017	江门	1958	2015
	珠海	1958	2013	阳江	1982	2013
	汕头	1986	2013	湛江	1958	2013
	佛山*	2012	2016	茂名	1986	2015
	韶关	1983	2011	肇庆	1958	2011
	河源	1981	2014	清远	1958	2012
	梅州	1958	2012	潮州	1958	2014
	惠州	1958	2012	揭阳	1980	2013
	汕尾	1962	2016	云浮	1984	2014
	东莞	1958	2014	顺德	1991	2016
网签数据	广州	2000	2017	惠州*	2009	2015
	深圳	1998	2017	汕尾	2007	2015
	珠海	1985	2013	江门	2006	2015
	汕头	1985	2015	茂名*	2011	2016
	佛山	1979	2017	肇庆	1999	2016
	韶关	1994	2011	潮州	2008	2017
	河源	1993	2013	云浮	2006	2015
	梅州*	2007	2012	顺德	2002	2014

具体以某市为例说明:

1) 用最近 1993–2017 年数据进行预测评价:

输入 2012 年以前数据, 预测 2013 年的环比增量.

输入 2013 年以前数据, 预测 2014 年的环比增量.

输入 2014 年以前数据, 预测 2015 年的环比增量.

输入 2015 年以前数据, 预测 2016 年的环比增量.

输入 2016 年以前数据, 预测 2017 年的环比增量.

2) 然后, 将预测结果与真实数据进行比较评价.

计算及评价结果见图 7. 图中, 预测数据序列在 2012 年及之前的年份直接使用真实数据.

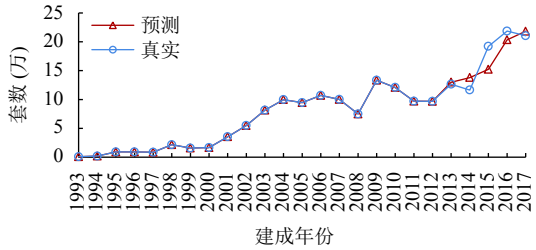


图 7 某市房产登记预测示例

3.3 预测数据集生成

房产登记簿数据: 其落图压缩后得到的 353 464 条不同经纬度坐标的数据记录网格化后, 落至 81 240 个经纬度网格上, 用于生成其预测模型的训练数据集.

房产网签数据: 其落图压缩后得到 50 242 条不同经纬度坐标的数据记录网格化后, 落到 3 514 个经纬度网格上, 用于生成其预测模型的训练数据集.

对于已知房产时序数据$seq=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_L)$, 将其规范为$(\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_L)|\boldsymbol{G}_k=\boldsymbol{X}_k-\boldsymbol{X}_1(1\leqslant k\leqslant L)$后, 放入房产数据预测模型预测出$L+1$年份相对于时序首年的热力值增量$\boldsymbol{G}_{L+1}$, 即可预测出至$L+1$年份的全量值:

$$\boldsymbol{X}_{L+1}=\boldsymbol{X}_1+\boldsymbol{G}_{L+1}$$

将$\boldsymbol{X}_{L+1}$加入已知序列得$seq:=seq+(\boldsymbol{X}_{L+1})$. 再从新的$seq$最后面切取长度为 L 的子序列, 并对此子序列重复上述过程, 就可逐年推进预测未知年份的数据. 我们将一个年份的预测数据存储成一个文件, 以 (经度, 纬度, 年份, 预测全量值) 格式输出到 JSON 和 CSV 文件中. 其中, “预测全量值”代表直到该“年份”年末, (经度, 纬度) 所标识区域的房产统计指标的历史累计值.

4 误差评价与学习校正

预测模型的评价指标采用平均绝对百分误差 (mean absolute percentage error, *MAPE*):

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-y_i'}{y_i}\right|$$

其中, n 表示序列样本的个数, y_i表示真实值, y_i'表示预测值. 平均绝对百分误差反映了预测值对真实值的偏离程度. 优化方法采用 Adam (adaptive moment estimation) 算法[18]. Adam 是一种自适应调节学习率的方法. 它利用梯度的一阶矩估计和二阶矩估计动态调整预测模型每个参数的学习率. Adam 的优点主要在于经过偏置校正后, 每一次迭代学习率都有个确定范围, 使得参数比较平稳. 我们设定学习率$\alpha=0.001$, 取每个市最后 5 年真实数据作为评价模型的测试集.

误差评价可按以下指标衡量:

1) 预测增量相对于真实增量的误差 *MAPE-I*.

2) 预测全量相对于真实全量的误差 *MAPE-T*.

对于深度学习过程校正: 增量和全量预测模型分别使用 *MAPE-I* 和 *MAPE-T* 进行偏置校正.

对于最终预测结果评价: 考虑到本课题是以预测各市年末实有房产数 (累计全量数) 为目标, 我们使用 *MAPE-T* 进行误差评价.

测试评价表明, 不同市的训练样本的数量和质量、环比变化梯度对于预测误差的影响是不同的, 详见表 4. 表中佛山市的登记簿和梅州、惠州、茂名 3 市的网签数据因真实数据序列太短无法进行深度学习, 不能进行预测, 误差评价不适用 (标记为 NA). 此外, 个别市的训练数据样本较小, 年份实际增量波动较大, 预测效果相比其他市较差 (例如, 湛江、茂名两市的登记簿数据预测, 以及韶关和汕尾两市的网签数据预测), 甚至出现极端误差情况 (例如云浮市的网签数据预测). 但从整体结果来看, 我们的模型还是能捕捉到各市的房产指标数据的基本变化, 平均误差大多数在 5% 以下, 绝大多数在 10% 以下. 这说明, 针对性建立和训练的预测模型是有效的, 达到预期的目的.

表 4 各市房产套数和面积预测的平均绝对百分误差 (%)

数据类型	城市	套数	面积	城市	套数	面积
登记簿数据	广州	2.13	2.50	中山	2.44	2.16
	深圳	1.11	4.61	江门	1.56	3.12
	珠海	0.30	0.28	阳江	2.84	2.25
	汕头	3.25	1.75	湛江	38.62	45.48
	佛山	NA	NA	茂名	14.27	42.8
	韶关	7.28	7.47	肇庆	10.80	9.43
	河源	6.12	2.96	清远	0.96	1.73
	梅州	5.16	1.96	潮州	2.20	0.46
	惠州	0.44	0.78	揭阳	7.00	7.46
	汕尾	0.44	1.82	云浮	5.68	0.94
	东莞	0.52	0.70	顺德	2.61	5.01
网签数据	广州	1.01	1.07	惠州	NA	NA
	深圳	6.70	3.21	汕尾	17.08	18.71
	珠海	4.35	11.33	江门	1.45	2.38
	汕头	1.28	2.38	茂名	NA	NA
	佛山	1.36	2.56	肇庆	4.77	2.33
	韶关	22.37	13.89	潮州	13.68	8.79
	河源	5.83	6.84	云浮	88.67	88.61
	梅州	NA	NA	顺德	1.19	1.18

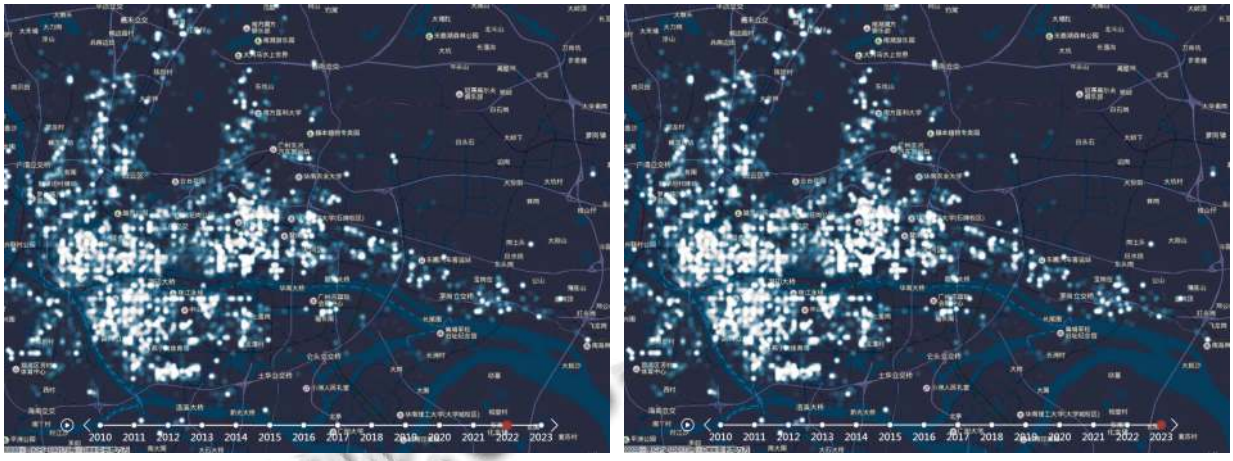
5 图效展示

以广州市房屋登记簿时序数据预测为例. 我们用已知的 1958–2017 年历史数据记录, 对其各网格单元区域未知的 2018–2023 年各年年末房屋套数热力值进行预测. 然后将广州市各网格区域所有年份的年末实有房屋套数热力值在百度地图落图呈现—调用百度地图 LBS 服务, 渲染成市域房屋统计指标数据地理分布热力值图[19], 并在此基础上提供正反时序的各年度热力图播放, 供观察比较. 为简便起见, 我们仅取广州市房屋套数分布热力图局部情况为例进行前后两两环比,

限于篇幅, 仅以图 8 和图 9 例示. 目测可见, 随着预测年份推延, 全市实有累计房屋套数的地理分布热力值增加趋势具有如下特征: 开始变化明显, 后来逐渐减弱. 这里的一个原因是随着预测年份的增长, 模型的预测值准确度下降, 预测结果趋同; 另一个原因是每当热力图图斑开始呈现高亮度色时, 后来的热力值增加对于图斑的增强作用会显著减弱. 图示目测效果大体与我们对广州市近年城市建设区域发展的直观预期相符.

(a) 2017 年 (b) 2018 年

图 8 2017 与 2018 年房屋套数热力图环比



(a) 2022 年 (b) 2023 年

图 9 2022 与 2023 年房屋套数热力图环比

6 结束语

(1) 应用集成方面, 本文基于深度神经网络机器学习的广东省域房产大数据热力图人工智能预测系统, 与广东省域“(房产) 行业数据云同步枢纽平台系统”[20]、“HBDP 省房屋大数据计算集群及作业调度系统”和“热力图播放系统”配合使用, 全面实现了可从过往已知的至后来未知的广东省域房产登记和交易大数据指标的热力图年时序播放展示, 有助于从时空维度俯瞰城市建设演化过程, 为城市资源和区划经济规划等相关的宏观管理活动提供历史和前瞻性的大数据参考, 也可作为满足人们对城市具体属性未来演变的预见愿望的行业大数据直观应用的实践范例.

(2) 技术创新方面, 本文提出“网格累计量预测+市域增量预测修正”的总体预测建模计算框架, 相对于常规直接应用深度神经网络机器学习模型, 不但在建模前期引入了网格粒度调选环节, 为简化计算量提供选项, 更重要的是有效地将细化的局部预测与全局宏观预测修正结合, 较好地化解了因各地网

格单元源数据样本质量参差不齐而引起的模型训练预测误差过大问题, 创造条件来调优应用长短时记忆与全连接层网络 AI 深度学习模型. 这样, 即使在网格单元样本时间和空间数据质量不理想情况下, 仍然实现了课题的技术目标: 系统产生的预测数据可直接应用于百度热力图呈现, 预测模型的可评测应用误差总体囿于合理范畴, 相应数据结果可视化符合人们目测预期.

参考文献

1 广东省统计局, 国家统计局广东调查总队. 2019 年广东省国民经济和社会发展统计公报. http://stats.gd.gov.cn/attachment/0/388/388463/2923609.pdf. [2021-03-03].

2 Yang HT, Lv JM, Xu F, *et al*. Regression approach for optimal purchase of hosts cluster in fixed fund for hadoop big data platform. International Journal of Computer and Information Engineering, 2017, 11(5): 634–641.

3 杨海涛, 程思瀚, 阮镇江, 等. 一种地址关联数据处理方法、用户终端和服务器: 中国 ZL, 201710850892.8. 2021-09-21.

4 百度公司. 百度地图开放平台→开发文档→服务接口→Web 服务 API→正/逆地理编码-地理编码服务. http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-geocoding. (2019-06-18)[2021-03-03].

5 Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. Anderson JA, Rosenfeld E. Neurocomputing: Foundations of Research. London: MIT Press, 1988. 318–362.

6 Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature, 1986, 323(6088): 533–536. [doi: 10.1038/323533a0]

7 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]

8 Goodfellow I, Bengio Y, Courville A. 深度学习. 赵申剑, 黎彧君, 符天凡, 等译. 北京: 人民邮电出版社, 2017. 119–122, 248–250.

9 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

10 Malay FC, Eme M, Zcanhan MH. Comparison of big data time series analysis methods. V. International Scientific and Vocational Studies Congress-Engineering (BILMES EN 2020). 2020: 209–214.

11 周志华. 机器学习. 北京: 清华大学出版社, 2016. 97–115.

12 Google INC. TensorFlow. https://tensorflow.google.cn/. [2021-03-03].

13 Google INC. Keras. https://keras.io/. [2021-03-03].

14 NVIDIA Corporation. CUDA. https://developer.nvidia.com/cuda-downloads. [2021-03-03].

15 NVIDIA Corporation. cuDNN. https://developer.nvidia.com/cudnn. [2021-03-03].

16 Numpy Steering Council. NumPy. https://numpy.org/. [2021-03-03].

17 Anaconda. Anaconda individual edition. https://www.anaconda.com/products/individual. [2021-03-03].

18 Kingma DP, Ba JL. Adam: A method for stochastic optimization. arXiv: 1412.6980v9, 2017.

19 百度地图开放平台. JavaScript API GL. http://lbsyun.baidu.com/index.php?title=jspopularGL.(2020-08-07)[2021-03-03].

20 杨海涛, 张传斌, 阮镇江, 等. 大规模云同步归集数据系统的异步并行优化. 计算机工程与应用, 2017, 53(2): 88–97. [doi: 10.3778/j.issn.1002-8331.1605-0117]