计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(11):19−28 [doi: 10.15888/j.cnki.csa.007148]

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

有限感知条件下的停车数据批量式修复研究①

张华成, 邹　万, 刘建明, 钟晓雄, 杨　兵

(桂林电子科技大学 计算机与信息完全学院, 桂林 541004)
通讯作者: 杨　兵, E-mail: 170058612@qq.com

摘　要: 随着城市中私家车保有量和使用频次显著增加, “停车难”问题逐渐成为制约城市发展的“瓶颈”. 为了合理利用城市中有限的停车资源, 最好的方式是建立城市级的停车诱导系统, 而现阶段尚且没有有效的方案出现, 究其原因是获取停车数据成本过于高昂导致的. 因此, 如何在不影响停车数据准确性的前提下降低其获取成本成为解决“停车难”问题的关键. 本文首先基于停车数据的时空敏感性, 将数据差异明显的停车场分为不同簇; 再验证同一簇中停车数据符合二八定律后, 筛选出影响力最大的前 20% 的停车场作为样本停车场, 对其安装传感器获取实时停车数据并作为样本数据; 考虑到现有算法得到的修补数据效果不理想, 本文将一维停车数据升至二维, 使用改进后的深度卷积对抗生成网络 (Deep Convolution Generative Adversarial Networks, DCGAN) 生成与样本数据近似同分布的新数据集. 新数据集的任一条可作为同簇中任一缺失的停车数据. 实现结果表明, 本文提出的方案不仅可在有限感知的条件下批量式的获得大量高仿分的“伪数据”, 大幅降低停车数据的获取成本, 而且修复效果较当前研究有明显提高.

关键词: 城市级停车诱导系统; 有限感知; 数据升维; 批量式

引用格式: 张华成,邹万,刘建明,钟晓雄,杨兵.有限感知条件下的停车数据批量式修复研究.计算机系统应用,2019,28(11):19–28. http://www.c-s-a.org.cn/1003-3254/7148.html

Research on Batch Repair of Parking Data under Limited Sensing Conditions

ZHANG Hua-Cheng, ZOU Wan, LIU Jian-Ming, ZHONG Xiao-Xiong, YANG Bing

(School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: With the significant increase in the number and frequency of private cars in cities, the problem of “parking difficulties” has gradually become a “bottleneck” that restricts urban development. In order to make reasonable use of the city’s limited parking resources, the best way is to establish a city-level parking guidance system. At this stage, there is no effective solution. The reason is that the cost of obtaining parking data is too high. Therefore, how to reduce the procurement cost without affecting the accuracy of the parking data becomes the key to solving the problem of “parking difficulties”. First, based on the spatiotemporal sensitivity of parking data, parking lots with significant data differences are divided into different clusters. After verifying that parking data in the same cluster complies with Pareto’s principle, the top 20% of the most influential parking lots are selected as sample parking. At the site, sensors are installed to obtain real-time parking data and used as sample data. Considering that the patch data obtained by the existing algorithm is not satisfactory, this study upgrades the one-dimensional parking data to two-dimensional, and uses the improved Deep Convolution Generative Adversarial Networks (DCGAN) to generate new data settings and sample data. Roughly the

① 基金项目: 国家自然科学基金 (61262074, 61802221, 61802220, 61602125); 广西自然科学基金 (2016GXNSFBA380010, 2016GXNSFBA380153, 2017GXNSFAA198192, 2018GXNSFAA294123)
Foundation item: National Natural Science Foundation of China (61262074, 61802221, 61802220, 61602125); Natural Science Foundation of Guangxi (2016GXNSFBA380010, 2016GXNSFBA380153, 2017GXNSFAA198192, 2018GXNSFAA294123)
收稿时间: 2019-04-11; 修改时间: 2019-05-08; 采用时间: 2019-05-16; csa 在线出版时间: 2019-11-06

same, any new data set can be used as any missing parking data in the same cluster. The implementation results show that the proposed scheme in this study can not only obtain a large number of high-quality “pseudo-data” in batches under the condition of limited perception, greatly reduce the acquisition cost of parking data greatly, but also significantly improve the repair effect compared with the current research.

Key words: city-level parking guidance system; limited sense; dimension-raising; batch

1 引言

伴随着汽车保有量的迅速增加, 停车难问题已成为无法忽视的城市通病, 因停车问题引发的纠纷屡见不鲜. 无论是超大型城市, 还是特大城市, 甚至只有几十万、十几万人口的中小型城市, 车辆迫切的停车需求与可用停车位不充足的矛盾都日益突出. 因此, 如何对城市中有停车需求的车辆进行宏观的停车诱导, 成为地方政府面临的一大难题. 为解决这个问题, 降低停车时间成和经济成本, 智能停车系统[1]是最有效的办法之一. 而城市中所有停车场当前可用停车位都已知是智能停车系统能正常运行的前提. 目前停车数据主要通过传感器采集和与停车场直接合作的方式获得, 但因为停车场类型、产权的多样性、经济成本及安装施工等原因的限制导致大范围的停车场数据处于缺失状态. 其次, 不同停车场以及不同辖区的系统以往都是独立运行, 必然存在数据兼容问题, 而且由于缺乏统一标准, 也不能实现数据共享, 多种原因导致难以将所有停车数据加入到大系统中形成规模. 因此, 城市中相同数据形式的停车数据往往是非充分的[2].

考虑到直接获得充分的停车场数据有足够的挑战性, 本文希望在经济时间成本可控范围内对缺失数据的停车场进行数据修复. 停车数据受时空两个维度的综合影响, 不同停车场间的数据可能差异极大, 如果直接将数据特征明显不同的停车数据当成一个样本集进行学习训练, 会得到无法解释的生成数据. 针对这个问题, 本文使用 K-means 聚类的方法按空间特征相似性将停车场划分为多个簇, 对每个簇单独进行数据修补. 通过可获得的停车场公开数据量化各个停车场的影响力, 对影响力高的停车场安装传感器以获得真实数据, 在此基础上修复其余停车场的停车数据. 因为不同地理位置、规模、收费标准等多种状况对停车场的影响, 不能简单的用一般插值法的方式修复. 因此本文采用数据增强技术, 也就是通俗意义上的数据生成/修补技术. 通过这种方法, 能在有限感知条件下下获取大量高仿真的停车数据.

2 相关工作

本文的停车数据为时序数据, 目前时序数据修补领域的研究主要是一些插值法[3], 主要为 3 类, 基于拉格朗日插值法的数据修补方法、基于牛顿插值法的数据修补方法、基于分段线性插值法的数据修补方法[4], 其中基于分段线性插值法的数据修补方法效果较好. 上述方法的特点是需要一定的先验知识, 常常被用于有一定历史数据的修补领域中, 然而由于经济成本、安装施工、经济产权等原因, 多数停车场历史数据难以获得, 因此是用插值法在对停车场数据修补时会有较大局限性.

在机器学习中, 处理数据缺失问题一般采用数据增强技术, 也就是传统意义上的数据生成或数据修复技术. 数据生成目前已成为机器学习领域的研究热点, 其中优秀的生成模型有生成对抗网络 (Generative Adversarial Nets, GAN)[5], 该模型由一个生成网络和判别网络组成, 在每一次迭代中, 判别器的目的是区分真实数据和生成数据, 而生成器则期望生成以假乱真的数据, 在零和博弈的思想下, 最终达到一个两者都可接收的结果. 由于 GAN 在时序数据中训练起来非常不稳定, 因此直接使用 GAN 可能会生成无意义的数据. 目前一种高效的的生成方式是将在时序数据表现良好的 LSTM 网络与 GAN 结合得到的循环生成式对抗网络 (Recurrent Generative Adversarial Networks, RGAN)[6], 该方法虽然能快速生成时序数据, 但缺点是生成结果伴有明显的抖动和相位差, 针对生成数据震动较大这个不足, 本文的解决思路是将时间序列升维, 使用在二维数据有强大特征提取能力的深度卷积对抗生成网络 (Deep Convolution Generative Adversarial Networks, DCGAN)[7]提高生成数据的稳定性, 实验表明该方法使生成结果更加稳定, 抖动减少.

3 系统模型

3.1 模型搭建

本文的停车数据指的是关于停车场的空车位的时间序列, 考虑到不同停车场同一时刻可用停车位数量差异较大引起的数据难以训练问题, 本文将空车位的时间序列除以停车场的规模, 转换成空车率的时间序列. 对于区域 O 中的 n 个停车场, 表示为$P=\{p_1,p_2,\cdots,p_i,\cdots|i=1,2,\cdots,n\}$则停车场 p_i 的空车率数据可表示为$U=\{r_1,r_2,\cdots,r_j,\cdots|j=1,2,\cdots,s\}$, s 为时间序列的长度. 实验将样本点的停车数据作为训练集, 使用改进后的 DCGAN 模型生成其余停车场的停车数据. 改进后的生成器如图 1 所示.

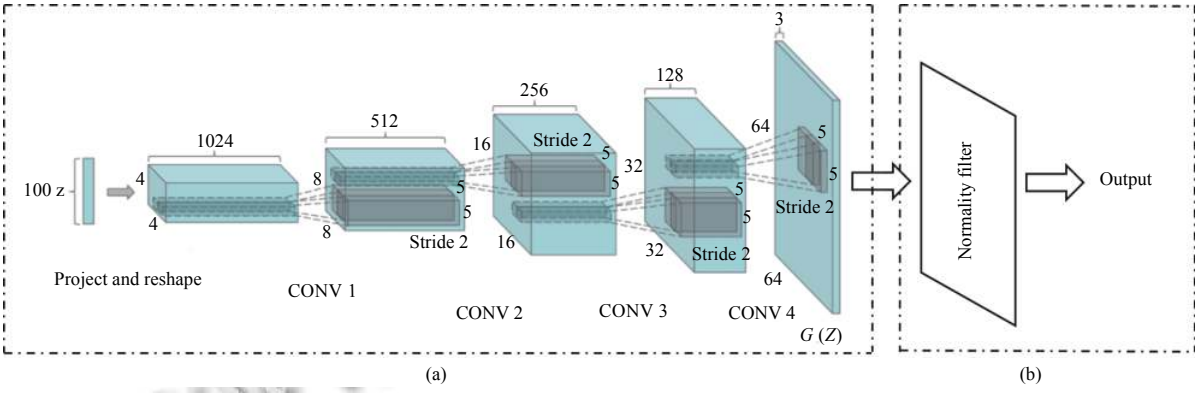

图 1 改进后的 DCGAN 生成模型

图 1(a) 描述的是, 将 100 维均匀分布的 z 投影到一个具有多个特征图的小空间范围的卷积中表示, 用一系列步长为 4 的卷积将这种高维的表示方式转换为 64×64 像素的图像, 注意不适用全连接层和池化层. 另外, 在生成器中引入正态性检验的过滤器, 用来保证生成的曲线效果可以接受, 如图 1(b) 所示. 其中, 正态性过滤器使用的算法为 D'Agostino-Pearson 检验, 该方法是一种对分布的偏度和峰度进行综合评定的方法[8]. D 值的公式如下:

$$D=\frac{\sum\left[i-\frac{n+1}{2}\right]\cdot U_i}{\sqrt{n^3-\left[\sqrt{U^2}-\left(\sum^{U}\right)^2/n\right]}} \quad (1)$$

在式 (1) 中, n 为样本总数, r_i 为停车场 p_i 的的停车数据. 得到 D 值后, 通过 D 界值表确定 P'的值, 按照 P'值判断这组样本是否符合正太性分布. 如果 P'小于设定的阈值∂, 则这组生成样本不符合正态性检验, 反之符合正态性检验, 并保留这组生成样本.

3.2 停车场的高维聚类

不同停车场的停车数据差异较大, 这是因为停车场不可能单独存在于地理空间中, 一定会受周围空间信息的影响, 地理空间信息其实也就是地理兴趣点 (Point Of Interest, POI), 包括住宅、商场、学校、公交站等, 不同 POI 对停车场有不同的影响. 比如某景区附近的停车场, 其停车位在节假日使用率明显高过工作日; 住宅区附近的停车场车位占用率在下班时段明显高于上班时段; 商场周围停车场的车位使用率在周末显著上升. 换句话说, 附近空间信息相似的两停车场其数据一定具有相似性, 如图 2 所示.

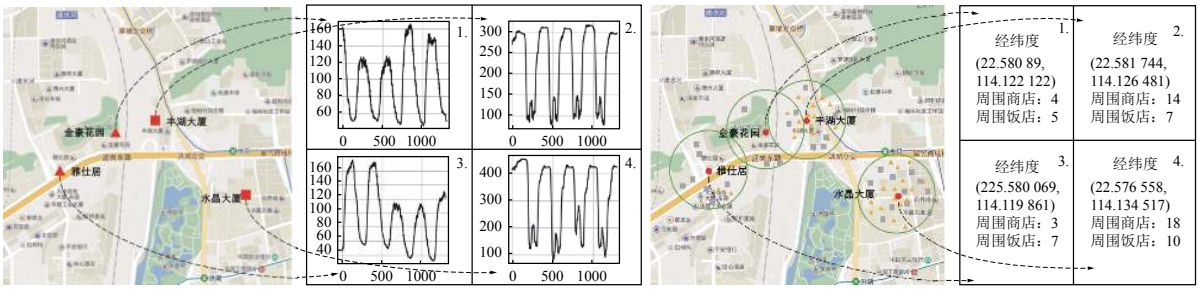

图 2 停车场间的拓扑关系

可以看到两对停车场停车数据差异明显; 而对于两对中的任何一对, 对内停车数据却很接近. 因此本文根据停车场的各 POI 数量, 对每个停车场转成高维向量的形式, 通过对停车场高维向量的聚类实现将停车场按数据差异分类的目的.

以停车场为圆心、R_t 为容忍度半径的圆用 O_p 表示. 圆内的 POI 会对停车场产生影响, 圆外的 POI 对停车场的影响不考虑. 假设在区域 Ω 内有 n 个停车场, 如果城市中主要的 POI 有 h 种, 那么, 对于任意一个停车场 p_i 统计 R_t 内 h 种主要 POI 的数量可构建一个 h 维的向量作其特征向量, 记为v_i, 表示为$\left\{v_i^1, v_i^2, \cdots, v_i^h\right\}$, 用来表示停车场受地理空间的影响. 考虑到停车场的经纬度也会对停车数据产生影响, 将停车场 p_i 的经纬度信息用一个 2 维向量, 记为 μ_i. 对于停车场 p_i 的地理空间信息用 (2+h) 维向量$es_i = (u_i, v_i)$唯一标定, 则 n 个停车场的高维向量记为$ES = \{es_1, es_2, \cdots, es_i, \cdots | i = 1, 2, \cdots, n\}$. 基于 K-means 的聚类算法更适合对高维向量进行聚类, 在本文中, 将对停车场高维聚类的公式为:

$$\begin{cases} m_j = \left(\mu_{j'}, v_{j'}\right) \\ \mu_{j'} = \dfrac{1}{|C_j|}\displaystyle\sum_{es_i \in C_j} \mu_i \\ v_{j'} = \dfrac{1}{|C_j|}\displaystyle\sum_{es_i \in C_j} v_i \\ E = \displaystyle\sum_{1}^{k}\sum_{es_i \in C_j}\left(a_0\left\|\mu_i - \mu_{j'}\right\|_2^2 + (1-a_0)\left\|v_i - v_{j'}\right\|_2^2\right) \end{cases} \tag{2}$$

其中, $C = \{C_1, C_2, \cdots C_j, \cdots | j = 1, 2, \cdots, k\}$为聚类产生的 k 个簇, m_j 为簇 C_j 的质心, E 为成簇 C_j 内样本与簇均值向量 m_j 的靠近程度, a_0 是经纬度 2 维向量和 POI 高维向量的权重.

3.3 一种停车场影响力评价算法

区域内的停车场间相互影响, 具体表现为影响力越强的停车场吸引车辆停车的能力越强, 当一个影响力强的停车场因为无空闲停车位而无法继续停车时, 车辆会向周围影响力较弱的停车场进行疏散. 换句话说, 在一个区域内某个停车场的影响力越强的, 则这个停车场越能代表这个区域的停车场. 因此, 如果对停车场按影响力进行排序, 那么只要知道影响力较高的停车场的停车数据, 就可以通过某种方式对其余停车场的数据进行修复. 本节的目的是筛选出影响力较强的停车场.

基于一般的认知标准, 与周围其它停车场连通度越高的停车场往往表现出更高的影响力, 因为人们更倾向于前往停车场较密集的区域, 这样会增加停车的成功率, 当一个停车场由于某种原因无法停车时, 可以轻松的向与之连通的停车场疏散. 此外, 相邻停车场也会互相影响, 具体表现在一个区域内影响力最强的停车场附近停车场评分会稍低, 但明显高于更远处的停车场 (类似于地理上的等高线). 因此, 实验需要将停车场的拓扑关系用数学方式描述. 假设有 6 个停车场, 它们的拓扑关系可用无向图表示, 如图 3 所示.

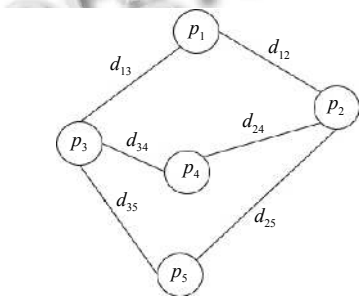

图 3 停车场间的拓扑关系

对于图 3 中任意两个停车场 p_i 和 p_j, 如果存在连通关系, 那么它们间的距离设为 d_{ij}. 上图中的拓扑关系也可用矩阵的形式表示:

$$M = \begin{bmatrix} 0 & d_{12} & d_{13} & \infty & \infty \\ d_{12} & 0 & \infty & d_{24} & d_{25} \\ \mathrm{d}_{13} & \infty & 0 & d_{34} & d_{35} \\ \infty & d_{24} & d_{34} & 0 & \infty \\ \infty & d_{25} & d_{35} & \infty & 0 \end{bmatrix} \tag{3}$$

考虑到距离的数值差异较大难以计算, 将 p_i 和 p_j 距离 d_{ij} 换成 p_i 和 p_j 的连通度, 并做归一化处理, 用 s_{ij} 表示 p_i 和 p_j 的转移概率, 则新的矩阵也就是概率转移矩阵如下:

$$M' = \begin{bmatrix} 0 & s_{12} & s_{13} & \infty & \infty \\ s_{12} & 0 & \infty & s_{24} & s_{25} \\ s_{13} & \infty & 0 & s_{34} & s_{35} \\ \infty & s_{24} & s_{34} & 0 & \infty \\ \infty & s_{25} & s_{35} & \infty & 0 \end{bmatrix} \tag{4}$$

一个连通度很高的停车场会存在多种无法停车的情况, 比如私有停车场、收费过高的停车场. 因此, 除了考虑停车场的连通度之外, 还需量化停车场的静态信息.

根据我国相关的法律法规, 任何一个合法经营的停车场都必须以公示牌的形式公开的展示该停车场的

类型、收费标准、规模等信息, 这些信息一定程度上代表了停车场相对于其它停车场的影响力. 不难发现, 不同的信息有着不同的深层含义, 具体如下:

(1) 停车场类型: 主要分为 4 种, 住宅、办公、政府、商场. 其中, 住宅类型和政府类型的停车场开放程度最低, 外来车辆很难进入, 商场类型停车场开放程度最高. 用 $1\geq x\geq 0$ 来表示不同类型停车场的开放程度, 当 $x=0$ 时不对外开放.

(2) 收费标准: 不同收入阶层能接受的收费区间不同, 低收费的停车场能被大多数收入阶层的人接受, 而收费较高的停车场只吸引高收入阶层的人. 因此, 收费标准可以表示停车场的受欢迎程度, 用 $y\geq 0$ 表示停车场的收费标准, 当 $y=0$ 时, 停车场最受欢迎.

(3) 停车场规模: 停车场的规模可以表示停车场的服务能力, 显然, 规模越大的停车场无疑影响力越强, 用 $z>0$ 表示停车场的服务能力

通过式 (5) 来量化静态信息对停车场影响力的影响, 也就是对其评分[2]:

$$SV_i = x_i \cdot \frac{z_i/\|z\|}{1+y_i/\|y\|} \tag{5}$$

其中, SV 表示停车场的评分值. i 为停车场 p_i 的编号, $z_i/\|z\|$和 $y_i/\|y\|$的目的是对收费标准和停车场规模归一化处理. 用 $SV=\{SV_1,SV_2,\cdots,SV_i,\cdots|i=1,2,\cdots,n\}$表示区域 O 中的 n 个停车场的评分向量.

考虑到停车场的静态信息和拓扑关系都对停车场的影响力有显著影响, 因此如何合理的统筹这两部分, 成为必须要解决的问题. 本文的方法是在已知概率转移矩阵 M'的条件下, 求解平稳状态下向量 SV 的值, 可表示为 $SV=M'.SV$, 显然, SV 是循环定义的, 所以引入著名的幂迭代法求解平稳状态下向量 SV. 将式 (1) 得到的评分向量作为初始评分向量 SV^0, 与转移概率矩阵 M'作为幂迭代法的输入, 多次迭代不断修正评分值 SV, 直到 SV^i 和 SV^{i+1} 的差值小于一个阈值 θ 时, 迭代终止, 并取 SV^i 作为最终的评分向量. 迭代公式如下:

$$SV^{i+1} = M'.SV^i \tag{6}$$

实验选取影响力较大的部分停车场为样本点, 来修复其余停车场的停车数据. 由于停车数据受到人类社会的活动的影响, 一定程度上停车数据是满足正态分布的, 可用式 (1) 中的 D 值检验证真实停车数据是否存在正态性, 当数据存在明显正态性时, 则可根据二八定律[9], 也就是影响力最大的前 20% 的停车场基本包括该簇停车场全部的特征, 停车数据因受多种复杂因素的影响, 在服从同一分布的前提下必然含有一定的多样性. 一般数学方法生成的同分布数据极为相似, 不满足停车数据的特点, 而这正是 GAN 的优势所在, 因此本文基于 GAN 的思想学习样本数据分布并生成新的停车数据. 与传统大量部署传感器获得数据的方式相比, 显著降低了时间和经济成本.

3.4 修复缺失停车数据

考虑到直接使用一维时间序列生成同簇数据时, 其结果难免伴随有明显抖动[6]. 为了使生成数据的效果更平滑, 需要对一维时间序列升维. 本文解决方式是将其转为二维曲线, 并以图像方式保存, 如图 4 所示.

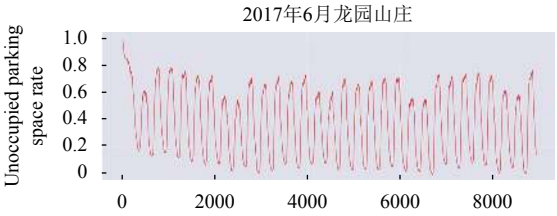

图 4 一条真实的空车率曲线

将筛选出的二维曲线集做为学习样本, 采用基于图 1 的 DCGAN 模型中训练. 一条生成曲线如图 5 所示.

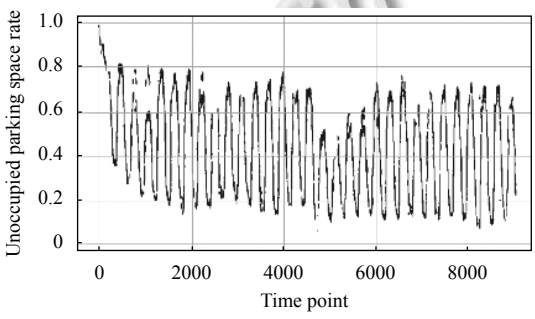

图 5 一条生成的空车率曲线

从图 5 可以看到, 生成图像伴随有明显的噪声, 因此需要对生成数据进行降噪处理. 本文试验中对生成图像的处理包括如下 3 步.

第一步, 需要把产生的图片灰度化. 即将灰度化之前的 RGB 值分别设为 R_1、G_1 和 B_1, 相应的, 灰度化后的值设为 R_2、G_2 和 B_2. 用公式表示为:

$$R_2 = G_2 = B_2 = a_1.R_1 + a_2.G_1 + a_3.B_1 \tag{7}$$

第二步, 将灰度化的做二值化处理. 二值化处理方

法为设定一个阈值 γ, 遍历矩阵中每一个数值, 如果该数值大于 γ 则设为 255, 若像素点值小于该阈值则设为 0.

第三步, 将异常值处理. 下一节将提到.

3.5 异常值处理

图 5 中异常值分为两类, 在曲线峰值处像素点过于密集, 称为毛刺点; 在曲线外零星的像素点, 称为离群点.

对于毛刺点, 实验采用均值滤波的方法[10], 降低毛刺点处像素点的密度. 均值滤波的公式如下所示.

$$f'(t,r)=\frac{\sum f(t,r)}{W} \tag{8}$$

其中, t 代表时间轴和 r 为空车率; W 表示滤波窗口, 大小取默认的 3×3; $\sum f(t,r)$表示遍历原图像所有像素点; 最后$f'(t,r)$表示滤波之后的新图像.

对于离群点, 实验采用局部异常因子 LOF 算法 (Local Outlier Factor)[11]来寻找. 思想是通过比较每个点 q 和其邻域点的密度来判断该点是否为离群点. 设 $N_q(k)$ 表示以 q 为圆心, $d_k(q)$ 为半径的圆, 其中 $d_k(q)$ 为点 q 到第 k 远点的距离. 实验中选 k 为 3. 寻找离群点用到的公式如下:

$$dist_k(q,o)=\max(d_k(q),d(q,o)) \tag{9}$$

式 (9) 中, $dist_k(q,o)$ 表示可达距离, $d(q,o)$ 表示点 q 到点 o 的距离. 当点 o 在 $N_q(k)$ 圆内, 则 $dist_k(q,o)$ 等于 $d_k(q)$, 当点 o 在圆 $N_q(k)$ 外, 则 $dist_k(q,o)$ 等于 $d(q,o)$.

$$lrd_k(q)=1.0/\left(\frac{\sum_{o\in N_k(q)}dist_k(q,o)}{|N_k(q)|}\right) \tag{10}$$

式 (10) 定义了局部可达密度 $lrd_k(q)$, 可以理解为一个密度, 密度越高, 则认为越可能属于同一簇, 反之, 越可能是离群点. 其中$|N_k(q)|$描述的是 q 为圆心, 邻域为 $d_k(o)$ 点的个数.

$$LOF_k(q)=\frac{\sum_{o\in N_k(q)}\frac{lrd_k(o)}{lrd_k(q)}}{|N_q(k)|} \tag{11}$$

因为密度的阈值难以选定, 实验引入局部离群因子来判定每个点 q 是否为离群点, 如式 (11) 所示. 其中, $LOF_k(q)$ 描述的是点 q 在圆 $N_q(k)$ 的局部可达密度 $lrd_k(q)$ 与点 q 的局部可达密度之比的平均数. 如果这个比值越接近 1, 说明点 q 的邻域点密度越接近, q 和邻域同属一簇; 如果这个比值越小于 1, 说明 q 的密度高于其邻域点密度, q 为正常点; 如果这个比值越大于 1, 说明 q 的密度小于其邻域点密度, q 越可能是离群点.

4 仿真实验

本实验的目的是在有限感知的前提下, 获得足够多的停车数据, 为基于机器学习的停车诱导系统提供充足的数据支撑. 目标区域停车场的静态信息, 通过百度地图拓展包 BMap 得到. 本文的思路是筛选出样本点, 通过对样本点安装传感器可以得到实时停车数据, 基于这些样本点来修复剩余的实时停车数据, 达到实验目的. 而现实情况是没有条件安装这些传感器, 因此选择已知 2017 年 6 月停车数据的深圳市罗湖区的 392 个停车场来进行仿真实验 (实验收据为采购获得), 最后将修复的数据与测试数据对比来筛选出合理的生成数据.

4.1 样本停车场的筛选

目标区域主要 POI 分布如图 6 所示.

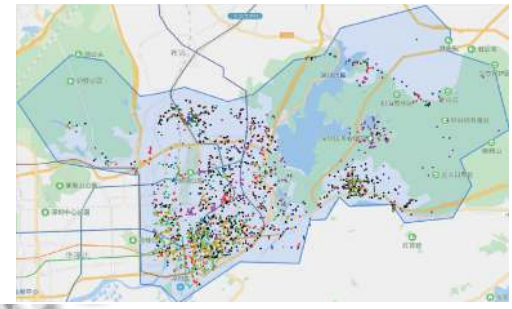

图 6 深圳市罗湖区主要 POI 分布

考虑到空间信息差异大的停车场间停车数据同样差异过大, 在筛选样本停车场前需要将停车场进行聚类. 实验中, 为方便计算, 取容忍度 R_t 为 310 米, 在地图中恰好约等于 1′, 对每个停车场统计其半径 310 米方位内 POI 的 7 维向量. 结合其位置得到 9 维向量. 部分停车场的 9 维向量如表 1 所示.

对 392 个停车场进行高维聚类, 结果如图 7 所示.

表 1 部分停车场的 9 维向量

停车场	经纬度	POI 向量
鸿翔花园	(22.561,114.112)	(2,2,3,1,4,4,3)
龙园山庄	(22.593,114,116)	(1,11,1,1,3,0,2)
鹤围村	(22.592,114.119)	(0,9,1,0,8,0,2)
农行大厦	(22.548,114.113)	(1,0,7,5,0,19,4)

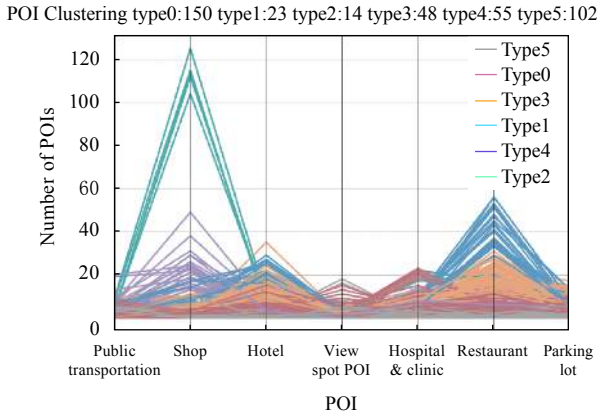

图 7 对停车场的聚类结果

可以看到停车场数量最多的簇为‘type0’, 因此仿真实验选取簇‘type0’的 150 个停车场进行后续实验.

‘type0’的 150 个的停车曲线和正态性检验结果如图 8 和表 2 所示. 从图 8 可以看到, 有 3 条数据明显异常的噪声数据, 做剔除处理. 对其余 147 条数据在 8928 个时刻检验使用式 (1) 其正态性, 设阈值∂为 0.05. 结果如表 2 所示. 不满足正态性的组数不足 25%, 可以认为整体是符合正态性的, 因此停车数据适用于二八定律.

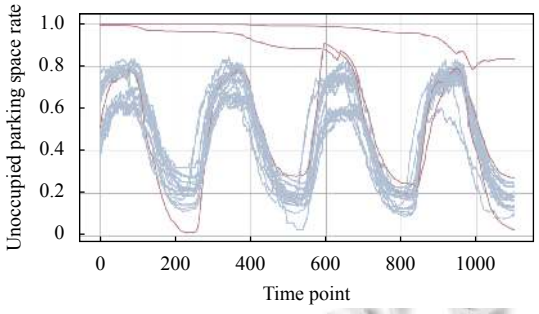

图 8 簇“type0”停车场的空车率数据

表 2 样本数据的正态性检验

检验组数	不符合正态性组数	所占比例 (%)
8928	2131	23.87

第二步对 150 个停车场进行编号, 根据式 (3)–式 (6) 及停车场间的空间拓扑关系, 计算所有停车场的评分, 并排序. 如表 3 所示.

取 θ 为 0.01, 当迭代次数达到 105 时, 评分值趋于稳定, 并全部保存. 可以看到序号为 74 和 73 的停车场存在 C_{74}^{0} 约等于 C_{73}^{0}, 而最终的 C_{74}^{105} 却远远大于 C_{73}^{105} 的情况. 导致这种情况的原因是两停车场的拓扑关系也就是连通度差异较大. 具体来说, 序号为 74 的停车场与周围停车场的连通度远远大于序号为 73 的停车场, 当车辆在 74 号或 73 号停车场无法停车时, 处于 74 号停车场的车辆更容易疏散到附近停车场. 序号为 39 和 74 的停车场, 存在 C_{39}^{0} 和 C_{74}^{0} 差异较大, 而 C_{39}^{105} 和 C_{74}^{105} 却比较接近, 这还是由连通度差异较大导致的. 74 号停车场连通度大于 39 号停车场, 一定程度上弥补了 74 号停车场先天条件的不足. 评分结果符合公众认知, 可被接受.

表 3 停车场评分的迭代过程

序号迭代	1	2	…	104	105
39	0.7369	0.6603	…	0.34093	0.3409
74	0.481	0.4372	…	0.341	0.3407
139	0.7106	0.6403	…	0.3343	0.3341
…	…	…	…	…	…
86	0.1471	0.1288	…	0.0856	0.0853
73	0.4314	0.1288	…	0.0853	0.0849
84	0.1491	0.1322	…	0.0847	0.0839

4.2 数据生成和处理

在深圳市罗湖区, 对应簇‘type0’中 147 个有数据的停车场, 筛选出 30 个样本点, 在此基础上修复其余停车场的停车数据. 实验以 2017 年 6 月整月为时间跨度, 每 5 分钟为时间间隔, 可划分出 8928 个时间节点, 并绘制空车率线图像. 一条真实的停车曲线如图 4 所示.

使用 DCGAN 为生成模型, 设置隐层神经元为 600 个, 批处理大小为 1, 学习率为 0.004. 在生成过程中, 每一次迭代生成器都会学习样本点在 2017 年 6 月的空车率数据, 并尽可能生成与样本点相似的数据. 由于在 DCGAN 生成模型中加了正态性检验过滤器, 所以生成的数据一定是符合正态性的. DCGAN 的生成过程如图 9 所示.

从图 9 中可以看到随着迭代次数的增加, 生成图像从模糊逐渐变得清晰, 实验取第 800 次迭代的结果. 图 5 为最终得到的一条生成结果, 从中可以看到生成的数据存在较多的噪声, 需要进行降噪处理. 降噪的第一步是要进行灰度化处理, 灰度化公式中的系数如表 4 所示.

具体二值化的做法是从图 5 左上角遍历每一个数值点, 设定阈值 γ 为 140, 当像素点的像素值大于该阈值将该值重新设为 255, 当像素值小于该阈值时则设为 0. 最后对二值化后的图像删除毛刺点和离群点. 图 10 为图 5 经过去噪的效果.

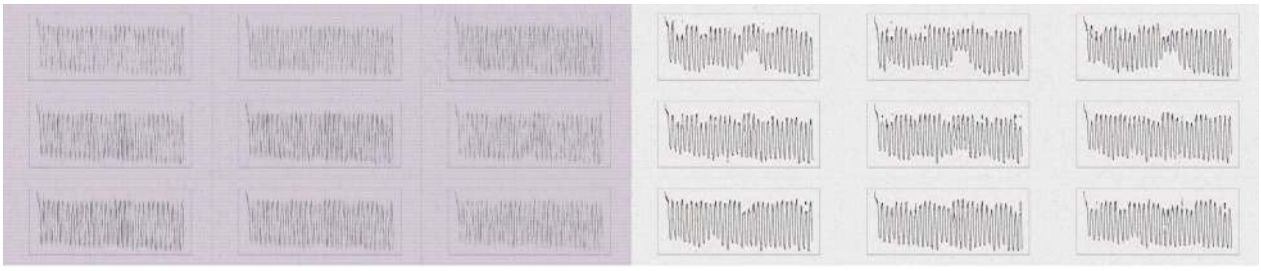

图 9 DCGAN 生成数据的过程.

表 4 式 (7) 的系数设置.

参数	a_1	a_2	a_3
取值	0.31	0.60	0.19

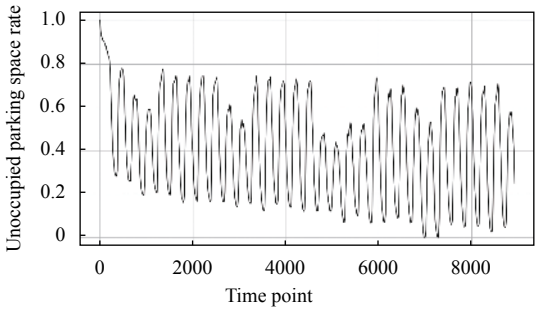

图 10 图 5 经过去噪处理的效果

从图 10 中可以看到, 生成数据一定程度的保留了原始空车率数据的概率分布信息, 即可表示出时间和空车率的关系. 另一方面, 由于 DCGAN 本身的特性, 生成结果不仅和真实数据有相似的概率分布, 且有一定的多样性. 因此, 只要生成结果集足够大, 就会包含该簇所有停车场的空车率数据.

为了对比生成数据的效果, 本文还复现了 RGAN 生成停车数据的实验. 设置隐层神经元 50, 批处理大小为 1, 学习率为 0.03. 使用与上文实验相同的训练集进行学习, 生成过程如图 11 所示.

在图 11 中, Iteration 表示迭代次数. 在迭代 100 次时曲线无规律, 随着学习的进行, 曲线渐渐变得平滑, 当迭代次数达到 4000 时, 曲线趋于平稳, 但仍有数据跳变的情况. 本文实验选取第 4000 次迭代结果.

考虑到GAN网络本身的缺陷,无论是本文基于DCGAN的生成模型还是已有的基于 RGAN 的生成模型, 都难免会生成十分异常的输出, 因此需要对生成的数据进行评估, 及时剔除明显错误的生成数据. 具体做法是在每一个时间节点计算生成数据与 117 条真实数据误差, 当一条生成数据在 85% 的时间节点上与真实数据集的误差都小于 0.05, 则保留这条数据, 反之则丢弃. 两种方法耗时对比和生成数据结果对比如表 5 和图 12 所示.

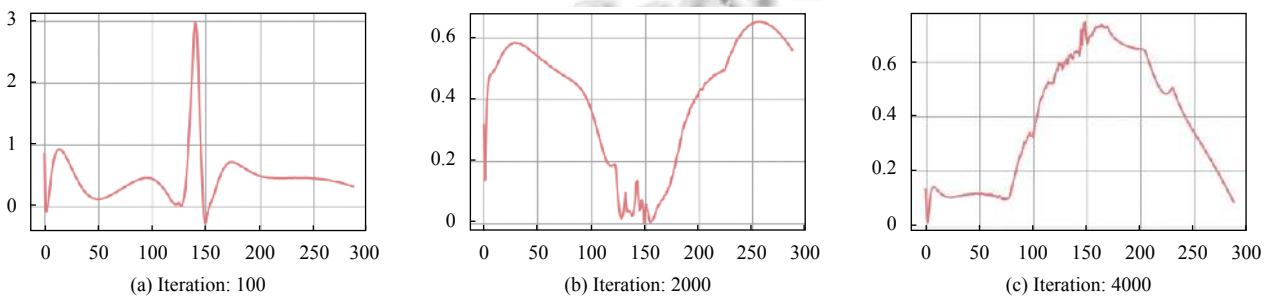

图 11 RGAN 生成数据的过程.

表 5 两种数据修补方法耗时对比.

修补方法	基于 RGAN 的数据修补方法	基于 DCGAN 的数据修补方法
修复一个月停车数据所需时间 (秒)	28 800	7600

从表 5 中, 可以看到基于 DCGAN 的停车场数据修补方法相较基于 RGAN 的停车场数据修补方法耗时显著降低. 从图 12 可以看出, 一方面无论 RGAN 生成模型还是 DCGAN 生成模型, 其生成数据和真实数据视觉上大致相似, 因此两种方法均存在一定合理性.

另一方面 RGAN 生成的数据出现了异常偏移和明显抖动, 而 DCGAN 生成数据较 RGAN 生成数据更为平滑. 这可能是 RGAN 网络对样本集学习过于充分而导致的泛化性能不强, 且 DCGAN 面向的二维数据比 RGAN 处理的一维数据有更多的特征. 考虑到停车数据受人类社会活动的影响, 一般情况下其数据变化是一个循序渐进的过程 (如图 12 中的真实数据), 特殊情况下有出现短期大幅跳变的可能, 比如体育场周边的停车场, 在有球赛的时其空车率会急促降低, 但如果多数生成曲线均存在急剧抖动现象, 会导致其与真实数据间的方差变大, 因此需要曲线平滑. 两种生成数据与真实数据方差如表 6 所示.

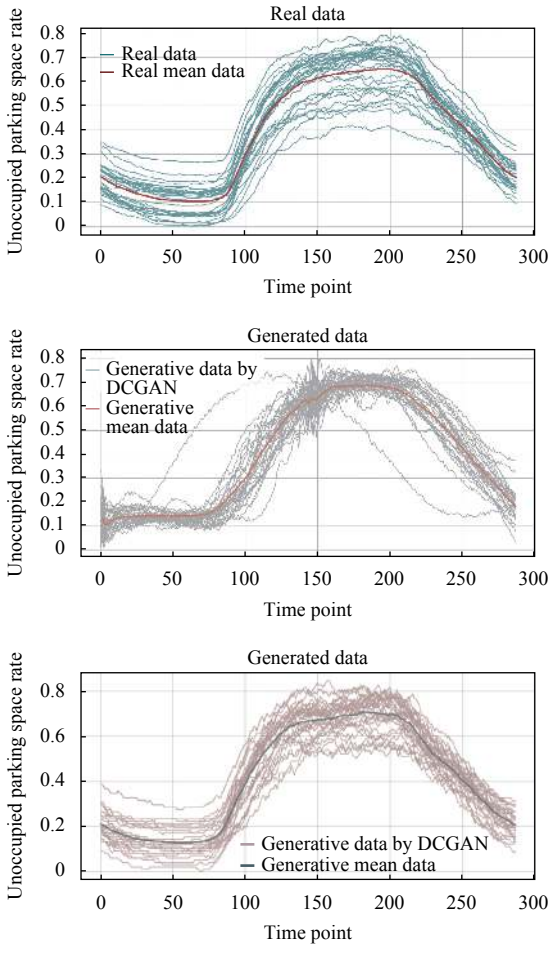

图 12 真实数据和生成数据对比效果

表 6 两种修补方法生成数据的离散程度

修补方法	基于 RGAN 的数据修补方法	基于 DCGAN 的数据修补方法
与真实数据的离散程度 (方差)	0.1484	0.1428

因此就修补速度和生成数据直观效果, 基于 DCGAN 模型的修补方法均明显优于基于 RGAN 模型的修补方法, 更符合公众认知.

为了进一步比较两种生成数据的质量, 还需衡量生成数据与非样本点的真实数据之间的误差. 本文引入均方根值 (RMS)、均方根分误差 (RMSE)、平均绝对误差 (MAE) 来描述这种误差, 并用卡方检验 (Chi-square test) 计算两种生成数据和真实数据同分布的比例, 设置卡方检验显著性水平为 0.05, 在 8928 个时间点上判断生成数据和真实数据是否属于同一分布, 如果卡方检验的 P 值大于 0.05, 则此时刻的生成数据与真实数据属于同一分布. 两种方法修复数据对比如表 7 所示.

表 7 两种数据修补方法效果对比

评价指标	检验方法	
	基于 RGAN 的数据修补方法	基于 DCGAN 的数据修补方法
均方根值 (RMS)	0.4546	0.4732
均方根误差 (RMSE)	0.0335	0.0483
平均绝对误差 (MAE)	0.0274	0.0377
和真实数据同分布的比例 (卡方检验, %)	92.94%	91.24%

从表 7 可以看出, 基于 RGAN 的数据修补方法的误差分析和正确率均稍好于基于 DCGAN 的数据修补方法, 这是因为 RGAN 中的 LSTM 对文本数据解释性较好. 因此, 总结两种方法的优缺点如表 8 所示.

表 8 两种生成方法的优缺点

优缺点	修补方法	
	基于 RGAN 的数据修补方法	基于 DCGAN 的数据修补方法
优点	1) LSTM 处理一维数据有明显优势. 2) 生成过程直接, 可从文本数据直接生成文本数据. 3) 生成数据准确性较高.	1) 生成数据速度更快. 2) 数据升维后, 生成结果抖动小、更平滑. 3) 生成数据直接可视化, 更直观.
缺点	生成数据伴随明显抖动.	生成过程不够直接, 并会出现噪声点, 增加了工作量.

5 结论

为了在降低经济时间成本的前提下, 获得城市中的所有停车场的停车数据, 本文提出了一种基于 DCGAN 生成模型来修复缺失数据的全新技术, 可通过对样本

停车数据的学习训练生成与之同分布的新数据, 由于GAN 网络生成数据多样的特征, 理论上只要新数据数量足够大, 就一定会包含该簇所有停车场的停车数据. 其中要解决的细节问题主要由两点组成. 首先, 不同地理信息的停车场数据差异巨大, 这样会导致生成数据可解释性差. 本文的方法是统计停车场周围 POI 的类型和数量将停车场映射为高维向量, 使用 K-means 算法将数据特征相似的停车场归为一个簇, 针对各个簇分别进行数据修复实验; 其次, 为了降低成本, 本文希望仅通过少量数据就能学习到足够特征来生成同分布的新数据. 对于任意一个簇, 本文做法是利用 PageRank 算法的思想通过对停车场的公开信息和停车场间的连通度的迭代计算, 算出各个停车场的影响力评分值, 在验证停车数据遵循二八定律的后, 将影响力最大的 20% 停车场作为样本停车场, 通过安装传感器等方式获取样本停车场数据, 以此为样本修复该簇其余的停车场停车数据.

本文的方法目前还不能针对具体停车场进行点对点的修复. 下一步主要研究方向是对特定停车场的数据进行修复.

参考文献

1 Jaurkar HV, Mulay GN, Gohokar V. Parking guidance system using internet of things. Proceedings of 2016 International Conference on Inventive Computation Technologies. Coimbatore, India. 2016. 1–6.

2 陈名松, 董适, 周信玲, 等. 一种基于公开信息的停车场推荐算法. 桂林电子科技大学学报, 2018, 38(3): 173–177. [doi: 10.3969/j.issn.1673-808X.2018.03.001]

3 韩卫国, 王劲峰, 胡建军. 交通流量数据缺失值的插补方法. 交通与计算机, 2005, 23(1): 39–42.

4 岳勇, 田考聪. 数据缺失及其填补方法综述. 预防医学情报杂志, 2005, 21(6): 683–685. [doi: 10.3969/j.issn.1006-4028.2005.06.013]

5 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 2672–2680.

6 Sun YQ, Peng L, Li HY, *et al*. Exploration on spatiotemporal data repairing of parking lots based on recurrent GANs. Proceedings of the 21st International Conference on Intelligent Transportation Systems. Maui, HI, USA. 2018. 467–472.

7 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv: 1511.06434, 2015.

8 D'Agostino RB, Belanger A, D'Agostino RB Jr. A suggestion for using powerful and informative tests of normality. The American Statistician, 1990, 44(4): 316–321.

9 刘良忠. 二八定律在零售企业中的运用. 商业时代·理论, 2004, (33): 31–32.

10 江巨浪, 章瀚, 朱柱, 等. 高密度椒盐噪声的多方向加权均值滤波. 计算机工程与应用, 2016, 52(6): 204–208. [doi: 10.3778/j.issn.1002-8331.1501-0332]

11 Ma MX, Ngan HYT, Liu W. Density-based outlier detection by local outlier factor on largescale traffic data. Electronic Imaging, 2016: 1–4. [doi: 10.2352/ISSN.2470-1173.2016.14.IPMVA-385]