高技术通讯 2023 年 第33 卷 第12 期:1323-1332

doi:10.3772/j.issn.1002-0470.2023.12.010

面向共享单车需求预测的多模型可视比较分析①

张奇奇②* 饶 宁** 朱素佳* 查 梦* 孙国道③*

(*浙江工业大学计算机科学与技术学院 杭州 310023)

(**杭州科技职业技术学院物联网技术学院 杭州 311402)

摘 要 针对不同的区域如何比较不同的模型并选择最佳模型进行预测以及是否存在普遍适用的最佳预测模型等问题,本文设计了一套面向共享单车需求预测的多模型可视比较分析系统。首先,使用隐含狄利克雷分布(LDA)模型,选择具有代表性的预测区域,并在预测区域之间设计相应的字形,通过字形比较多个模型的预测性能。其次,提出一种计算城市空间区域相似度的新方法 area2vec,用于比较相似区域模型预测性能的差异;设计使用网格布局算法,有效地缓解了地图中字形遮挡的问题。最后,通过案例分析和用户调查验证了该系统的有效性和实用性。

关键词 需求预测;隐含狄利克雷分布(LDA)模型;词嵌入;模型比较;可视分析

0 引 言

近年来,共享单车的广泛使用,打通了人们出行的“最后一公里”,让人们的出行更加便捷。但是由于不同区域单车投放量的差异,出现了单车需求量失衡的情况,此时对不同区域单车使用量的精准预测变得至关重要。考虑到共享单车数据的时空性,本文使用时空预测模型对共享单车的需求进行预测。

1 相关研究

时空数据预测的核心步骤是建立预测模型,即通过预测算法训练样本数据生成模型。当前用于时空预测的方法可大致分为 2 类:基于参数模型的传统预测方法,基于数据驱动和机器学习的预测方法。基于参数模型的传统预测方法是通过观察历史数据预测未来的值,包括滑动平均模型(moving average,MA)[1]、自回归滑动平均模型(auto-regressive moving average,ARMA)[2]和差分自回归滑动平均模型(auto-regressive integrated moving average,ARIMA)[3]等。基于数据驱动和机器学习的预测方法是通过对已知数据的学习,实现对新样本类标的预测,包括支持向量机(support vector machine,SVM)[4]、随机森林(random forest,RF)[5]法等。现在研究人员又提出了许多高级的算法,如循环神经网络(recurrent neural network,RNN)[6]和长短期记忆神经网络(long short term memory,LSTM)[7]等,以帮助人们深入理解不同领域的时空数据。

虽然有各种不同的模型可以对时空数据进行预测,但是大多数模型结构复杂、不易理解,且不同模型适用的场景不同,可能存在局限性。研究者对模型选择进行了大量的研究,如 TPOT[8]、Hyperopt[9]、H2O AutoML[10]等最常用的模型选择方法。尽管这些方法在许多应用中取得了显著成功,但是也隐藏着一些挑战。首先,在数据处理方面,现有的数据处

① 浙江省自然科学基金杰出青年科学基金项目(LR23F020003),国家自然科学基金面上项目(61972356)和浙江省属高校基本科研业务费青年科学家专项(RF-A2020001)资助项目。

② 男,1996 年生,硕士生;研究方向:信息可视化与可视分析;E-mail:2966414774@qq.com。

③ 通信作者,E-mail:guodao@zjut.edu.cn。

(收稿日期:2023-03-22)

理方法大多针对传统的关系数据，对于时空数据的处理尚不完善。而时空数据在时间维度存在大量缺失值、噪声以及在空间维度存在尺度不一致等问题。其次，没有对模型的预测性能进行适当的解释，只是根据数据给出一些推荐模型，用户可能想知道是什么因素影响模型的预测结果以及想了解模型的优势及使用范围。

为了解决上述问题，本文基于杭州市共享单车真实数据设计了一套交互式可视分析系统。该系统将多个预测模型与可视分析技术相结合，允许用户在视觉上比较不同区域不同模型之间的预测性能，帮助用户分析预测结果，从而在自己的工作中选择合适的模型。

本文的研究贡献主要有以下 3 个方面：

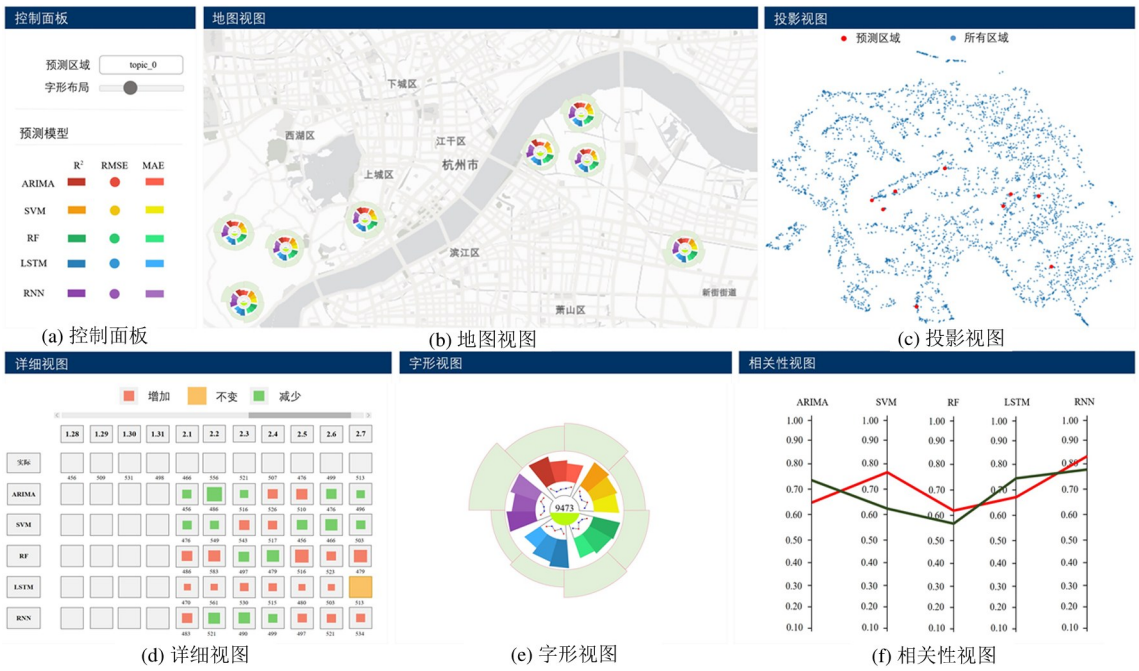
（1）提出了用于区域向量化的“area to vector”（area2vec）方法，进而实现区域相关性的计算。

（2）设计了包含多维属性的字形和网格布局算法，帮助用户比较区域模型的预测性能并为用户提供无遮挡的可视分析图。

（3）设计并实现了一套可视分析系统，通过 2 个案例的分析和用户调查证实了该系统的有效性和实用性。

2　系统概述

本文基于杭州市共享单车真实数据设计了一套可视分析系统，系统整体界面如图 1 所示。该系统包括 6 个视图，分别为允许用户选择预测区域和调节字形布局的控制面板（图 1（a））、展示区域在地理空间中分布的地图视图（图 1（b））、展示区域在向量空间中分布的投影视图（图 1（c））、展示实际数据与各模型预测数据的详细视图（图 1（d））、展示区域字形的字形视图（图 1（e））和展示实际数据与各模型预测数据之间相关性的相关性视图（图 1（f））。



（a）控制面板　（b）地图视图　（c）投影视图

（d）详细视图　（e）字形视图　（f）相关性视图

图 1　面向共享单车需求预测的多模型可视比较分析系统

本文系统的整体流程如图 2 所示，可分为 3 个模块：数据处理模块、数据预测模块和数据可视化模块。

（1）数据处理模块。该模块主要对原始共享单车的起点终点（orginal destination，OD）数据进行处理，包括数据清洗、数据类型转化、提取主要字段存储等。数据处理后，采用隐含狄利克雷分布（latent

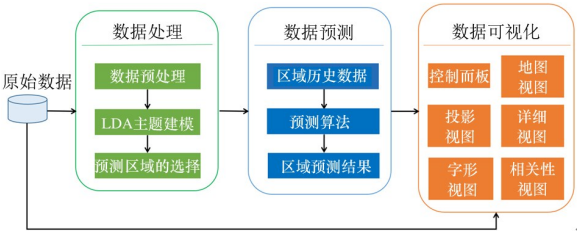


图 2　系统工作流程图

Dirichlet allocation,LDA)主题建模技术选取预测区域,同时完成区域数据特征分析,揭示不同区域数据特征间的相关性以及数据变化规律。

(2)数据预测模块。该模块主要基于各区域历史数据完成数据的预测,包括 3 个环节的工作。首先,基于准备的数据集提取特征,包括工作日、节假日、天气、温度、相对湿度等。其次,选择模型进行训练。将整理好的训练集用于模型训练,然后对数据预测,对比单车真实流量与预测流量。最后,对模型进行评估。本文选择 3 种常用的模型评估指标,分别为:决定系数(coefficient of determination,R^2)、均方根误差(root mean squared error,RMSE)和平均绝对误差(mean absolute error ,MAE)。

(3)数据可视化模块。该模块主要通过可视化技术进行视图设计和系统搭建,并通过多个视图对多模型进行可视比较分析。

3 数据和方法

3.1 数据形式

本文使用浙江省杭州市的共享单车数据、气象数据和兴趣点(point of interest,POI)数据。具体数据介绍如下。

(1)共享单车数据。本文收集了浙江省杭州市 2018 年 1 月至 2 月的共享单车起点终点(origin-destination,OD)数据。每条数据包括:起点经度、起点纬度、起点时间戳、终点经度、终点纬度、终点时间戳等。例如,其中 1 条记录形式为(120.1314,30.2600,2018-01-0114:30:29,120.1541,30.2598,2018-01-0114:48:07)。获取数据后需要对数据进行预处理,包括剔除越界数据、缺失数据和异常数据等并在数据库中存储主要字段。在数据库每一行额外增加 2 列,即是否工作日(1 表示是,0 表示否)和节假日(1 表示是,0 表示否)。

(2)气象数据。为了从可解释性角度出发探究时间序列特征对模型预测的影响,本文引入了外部特征因素,包括天气、温度、相对湿度。天气数据用数字 1、2、3、4 划分为 4 种类型。1 表示:晴朗,少云,晴间多云,部分多云;2 表示:薄雾 + 多云,薄雾 + 裂云,薄雾 + 少云,薄雾;3 表示:小雪,小雨 + 雷暴 + 散云,小雨 + 散云;4 表示:大雨 + 冰托 + 雷暴 + 雾,雪 + 雾。温度和相对湿度用当天采集到的平均温度和平均相对湿度进行表示。

(3)兴趣点(POI)数据。POI 是描述城市中具有位置信息的地点。每个 POI 至少包含 3 项基本信息:名称、类别和地理位置。其中 POI 类别包括交通设施、酒店、美食等。本文根据每个 POI 相对于区域的位置对区域内的 POI 进行统计,并根据 POI 比例将区域划分为不同类型。

3.2 方法描述

基于数据的地理位置,首先对地理空间进行网格划分。考虑到共享单车的停放点间隔一般在 500 m内,结合所研究的杭州市范围,将地理空间均匀划分为 100 × 100 的网格,每个网格大概为 500 m × 500 m,并对这些网格进行编号,从 0 开始一直到 9999 为止。本文将 2018 年 1 月处理好的每天的共享单车 OD 数据汇集起来,作为训练的原始数据。然后将 OD 数据上的全球定位系统(global positioning system,GPS)记录点都映射到该点所属的网格,则每条 OD 记录都可以转换为用数字表示的文本形式。

3.2.1 基于 LDA 模型的主题区域选择方法

将 LDA 引入到文本分析中。首先,将数据与模型中的概念一一对应。把转换后的每条 OD 记录看作模型中的一篇文档;记录中的每个编号看作组成文档的单词;所有 OD 记录即为整个文档集。其次,对 LDA 进行计算,得到文档 - 主题分布 θ 以及主题 - 单词分布 φ, 即每条记录与各个主题的相关度和每个主题与各个空间单位的相关度。最后,对 LDA 进行评估,得到最佳主题个数,每个主题选用相关度最高的前 10 个关键词进行表示,从而得到需要预测的区域。通过文献调研,选择 2 种常用的主题模型评价指标——困惑度(perplexity)[11] 和一致性(pointwise mutual information,PMI)[12]。困惑度表示模型的泛化能力。一致性表示模型是否具有产生语义一致主题词的能力。对于 LDA 模型中的另外 2 个超参数 α 和 δ, 根据相关经验[13],将其分别设置为 $\alpha = 50/k, \delta = 0.1$。

$$Perplexity = -\frac{1}{N}\sum_{d=1}^{N}\frac{1}{N_d}\lg P(d) \tag{1}$$

$$PMI = \frac{1}{K}\sum_{k}\frac{2}{T(T-1)}\sum_{1\leqslant i\leqslant j\leqslant T}\lg\frac{-p(\omega_i,\omega_j)}{p(\omega_i)p(\omega_j)} \tag{2}$$

在困惑度的计算过程中，困惑度越低，通常表示模型的泛化能力越好。其中 N 为文档的语料库，d 为文档，N_d 为文档中单词的数量。而在一致性的计算过程中，*PMI*的值越大说明模型产生的主题语义一致性越高。其中 k 为主题数，T 为每个主题最相关的词汇数，$p(\omega_i)$ 为单词 $\omega_i(i=1,2,\cdots,T)$ 在文档中出现的概率，$p(\omega_i,\omega_j)$ 为词对 (ω_i,ω_j) 共同出现的概率。如图 3 所示，当 $k=29$ 时，模型的困惑度最小，一致性在 25 ~ 30 区间内最大。因此，选取 $k=29$作为最佳主题数量。

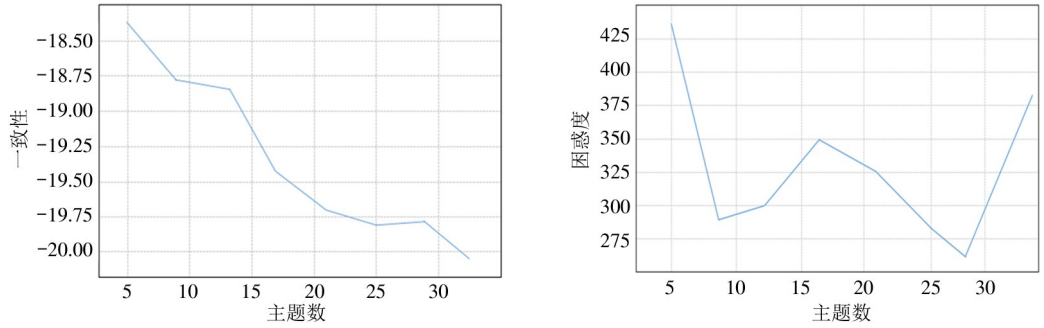


图 3　主题数与模型困惑度和一致性的关系

3.2.2　基于 area2vec 的区域空间表示方法

本文针对区域的地理空间位置提出了一种新的表达方法 area2vec。该方法将上文 OD 数据转换得到的文本类比为文档，将文本中的网格编号类比为单词，通过词嵌入技术将网格编号转换到词向量空间表示。该方法实现步骤如图 4 所示。

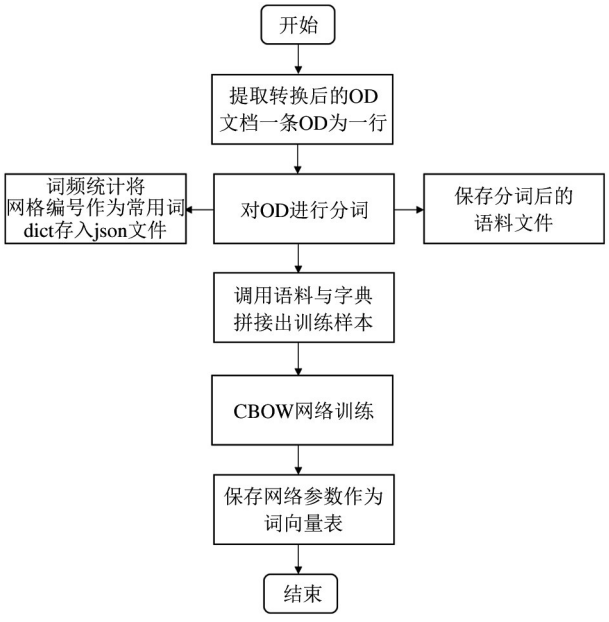


图 4　area2vec 方法运行流程图

在分析单词之间关系的过程中，单词之间的相似性是研究的重点，即如果在同一个文本中经常出现 2 个单词，则这 2 个单词在向量空间中的余弦距离会相对接近。类似地，在地理空间位置中 2 个相近的区域大概率会出现在同一条 OD 中，因此这 2 个区域的词向量表示也是相似的。此时区域的词向量很好地表示了区域之间的空间关系。本文中对转换后的文档使用 word2vec 模型进行训练，每个网格都会在向量空间中生成一个高维向量。以 $grid_i(i=1,2,3,\cdots)$ 表示网格，则网格间的相似性可以使用余弦距离表示为

$$S_{grid_1,\,grid_2} = \frac{v_{grid_1}\cdot v_{grid_2}}{\|v_{grid_1}\|_1\cdot\|v_{grid_2}\|_2} \tag{3}$$

为了观察区域在向量空间中的分布，本文使用 t-SNE[14] 处理向量数据。t-SNE 是用于探索高维数据的非线性降维算法，可以帮助用户将区域高维向量投影成二维平面上的点集。

4　可视分析任务与可视化系统设计

4.1　分析任务

本文的目标是通过可视化技术帮助用户比较分

析不同区域不同模型之间的预测性能,以及相似区域模型预测性能的差异。

针对上述目标,本文系统的主要任务如下。

(1)选择具有代表性的区域进行预测。面对城市不同类型的区域,如何选择既具有代表性又具有普遍性的区域进行研究是本文的分析任务之一。

(2)探究不同区域内兴趣点分布情况。通过对区域内兴趣点的计算可以从语义上丰富区域信息,并帮助用户理解不同区域的城市功能。

(3)比较各模型的预测性能。系统应该提供所有模型的预测性能视图,而不仅仅是选择性能最好的模型来展示,并从可解释性角度出发探究影响模型预测的时间序列特征。

(4)分析字形如何与地理空间位置相结合。根据不同区域的数据集,系统将设计好的字形直接放置在地图相应的区域上,可能出现字形重叠的问题,影响用户对区域字形的探索。因此需要设计使用网格布局算法,对字形重新放置。

(5)分析各预测区域模型输出的相关性。通过区域相似度的计算,探究空间相似的区域在模型预测性能上是否一致。

4.2 可视化视图设计

4.2.1 基于多模型比较的视图设计

为了更好地比较不同区域不同模型的预测性能,本文设计了地图视图、字形视图和详细视图。

(1)地图视图。地图可以为预测区域的地理位置提供一个全局的视角,用户可以通过该视图观察预测区域在地理空间中的分布。

(2)字形视图。由于本文预测区域包含多维混合属性的信息,因此需要设计一套能够集成这些多维信息的图形,通过图形将这些信息同时展示给用户,方便用户对不同预测区域进行探索。本文设计了一套基于径向布局的字形图来展示数据,该字形一共分为4层,每一层的设计都采用人们日常生活中所熟悉的图形,极大增强用户的理解能力,减少感知负担,如图5所示。

字形第1层,在内层圆中放置区域编号。通过LDA计算得到主题后,每个主题选取的关键词即区域编号,将编号均匀分布在内层圆中可以帮助用户更好地理解当前区域在地理空间中的分布。内层圆中的填充比重表示该关键词在该主题下的概率分布占比。

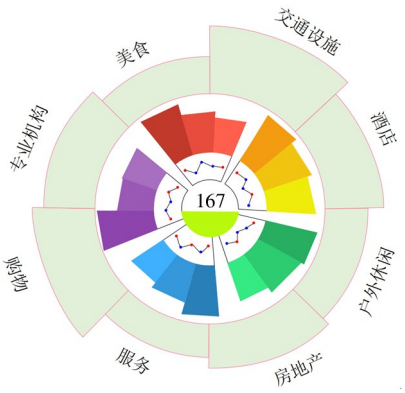


图5 预测区域的字形设计

字形第2层,通过折线图展示各时间序列特征对各模型预测的影响。本文选定了5种时间序列特征,分别为工作日、节假日、天气、温度和相对湿度,并按顺时针方向在字形中依次放置。为了探究各时间序列特征对各模型预测具有促进还是抑制作用,本文选用2种不同的颜色进行编码,并通过点分布的高低表示促进或抑制的大小。

字形第3层,通过环柱形图展示各模型的评估指标。通过调研和阅读文献,本文选定了R^2、RMSE和MAE 3个比较有代表性的评估指标。通过柱形图的高低编码各指标的大小,即柱子越高表示指标数值越大。对每个模型的3个评估指标,本文用顺时针方向的3个柱子分别编码R^2、RMSE和MAE。为了区分各模型,本文使用5种不同的颜色编码不同的模型,并且每种模型的各个评估指标使用渐变色进行表示。用户可以通过比较各模型的评估指标从而比较各模型的预测性能。

字形第4层,通过圆环图展示各预测区域的兴趣点数量分布情况,帮助用户理解不同区域的城市功能。圆环面积越大表示对应的POI数值越高,并按顺时针方向依次表示POI的8种主要类型分别为交通设施、酒店、户外休闲、房地产、服务、购物、专业机构和美食。

(3)详细视图。该视图展示了预测区域过去1个月每天单车的实际数据和未来1周5种模型的

预测数据与实际数据。方块颜色表示模型当日预测数据与实际数据的变化，用 3 种颜色编码数据的变化情况分别表示增加、不变、减少，面积表示变化的多少。方块下方的数字为实际数据和各模型预测数据。用户可以通过多模型预测数据和实际数据的比较得到区域最佳预测模型。

4.2.2　基于区域相关性分析的视图设计

为了探究相似区域模型预测性能的差异，即空空间相似的区域模型预测性能是否一致，本文设计了投影视图和相关性视图。

(1)投影视图。该视图以散点图的形式分析城市区域之间的空间关系，投影视图中的 1 个点代表城市中的 1 个区域。当用户选择预测主题后，预测区域所对应的点会在该视图中高亮显示，帮助用户直观观测各预测区域在向量空间中的分布，并通过点距离的远近判断 2 个区域间的相似性。

(2)相关性视图。为了探究模型预测数据与实际数据之间的关系，本文使用皮尔逊相关系数计算各模型预测数据与实际数据之间的相关性得分，并通过平行坐标图进行展示。可以帮助用户直观观测各预测区域模型预测数据与实际数据的相关性得分及排名。本文使用 10 种不同的颜色编码同一主题下 10 个不同的预测区域。

4.3　区域字形的空间布局设计

如图 6(a)所示，当主题区域距离较近时，地图上的字形会发生重叠，影响用户对区域字形的探索。因此本文设计使用网格布局算法对地图上的字形重新布局，该算法实现步骤如图 7 所示。

图 6　区域模型预测结果可视分析图(a)和网格划分示意图(b)

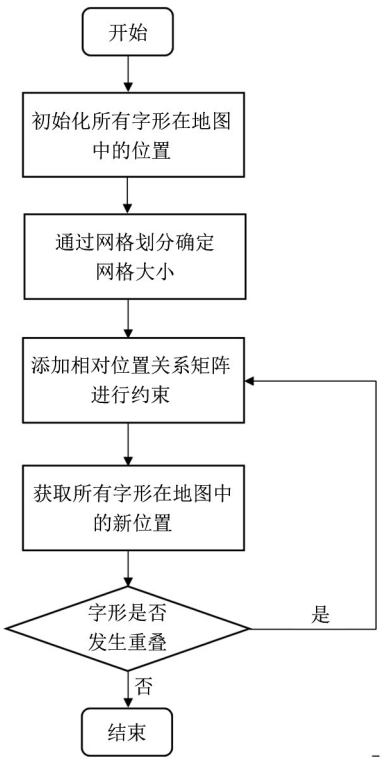


图 7　网格布局算法运行流程图

4.3.1　网格划分

首先，计算所有预测区域在地图中的位置，并将字形放置在预测区域的中心点。其次，采用类似网格索引的方式，遍历所有字形并将字形中最外层圆的最大直径作为一个网格尺寸，建立一个屏幕网格索引。然后，对所有地图要素坐标进行转换，从实际坐标(lng, lat)转换为屏幕坐标(x, y)，转换函数设为$f(x)$。最后，计算所有字形所要占据的网格，并将这些网格标记为不可放置。

网格划分完成后，如图 6(b)所示。每个网格的位置可以用一个二维数组表示，即 $Grid = [iRowLoc][jColLoc]$，其中 $iRowLoc$ 为屏幕网格的最大行数，$jColLoc$ 为屏幕网格的最大列数。用数字 1、0 表示网格是否被占用，如当行 1 列 1 被占用时则 $Grid[1][1]=1$。

$$iRowLoc = [mapScreenHeight/gridHeight]$$

$$jColLoc = [mapScreenWidth/gridWidth]$$

4.3.2　字形避让

初始化网格划分中，一个网格可能包含多个字形或多个网格包含一个字形。因此首先需要计算所有字形中心与字形所占网格中心的距离，并在网格

中放置距离最短的字形或将字形放置在距离最短的网格中,此时网格用 1 表示被占用。同时考虑到字形之间的相对位置关系,本文定义了一个相对位置关系矩阵 $\boldsymbol{P}$ 添加约束,使得其余字形按照相对位置在网格中进行放置。用 $g_1,g_2,\cdots,g_i$ 表示字形,对于单一字形,与自身的相对位置关系定义为 $null$。则 $\boldsymbol{P}$ 可表示为

$$\boldsymbol{P}_{(g_1,g_2,\cdots g_i)}=\begin{bmatrix} null & g_1\rightarrow g_2 & \cdots & g_1\rightarrow g_i \\ g_2\rightarrow g_1 & null & \cdots & g_2\rightarrow g_i \\ \vdots & \vdots & \ddots & \vdots \\ g_i\rightarrow g_1 & g_i\rightarrow g_2 & \cdots & null \end{bmatrix} \tag{4}$$

考虑到人眼视觉的识别能力,我们将一个字形与另一个字形的相对位置关系分为 8 种:上、右上、右、右下、下、左下、左和左上,并按顺时针方向分别定义为:0,1,2,3,4,5,6,7,如图 8 所示。

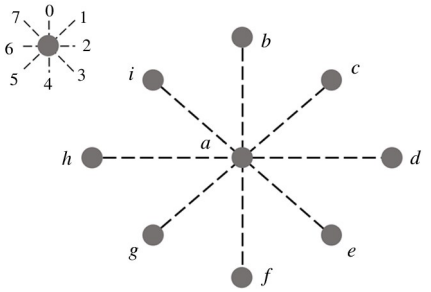


图 8　字形相对位置分析图

当字形位于图 8 中的 a 位置时,则与周边字形所在位置 b、c、d、e、f、g、h、i 组成的相对位置关系向量为 [$null$,0,1,2,3,4,5,6,7]。同理得出其他字形的相对位置关系向量,并组成相对位置关系矩阵。

布局是一个连续迭代的过程,每次运算过程中,每个字形都会在网格中得到一个临时的新位置,直到字形之间没有重叠,迭代才会停止。每次迭代过程中,通过计算临时相对位置关系矩阵 $\boldsymbol{P}$ 添加约束。如果一个字形的相对位置关系矩阵不同于原来的相对位置关系矩阵,也就是说该字形的相对位置发生了变化。根据该字形的初始相对位置关系矩阵和其他字形的相对位置关系矩阵来计算该字形的新位置。最后,得到布局后所有字形的新位置,如图 9 所示。

图 9　字形布局之后的图形

5　实验与案例分析

通过 LDA 主题建模技术和主题困惑度、一致性的计算得到 29 个主题,每个主题选用相关度最高的前 10 个关键词进行表示,即每次选取 10 个预测区域,计算并保存每个预测区域每天的单车流量数据。考虑到预测的实际意义,本文预测了 2018 年 2 月 1 日至 2 月 7 日各预测区域每天的单车流量数据。根据过去一个月的数据预测未来一周的数据。通过大量调研,本文选择 5 种具有代表性的时间序列预测模型,分别为 ARIMA、SVM、RF、LSTM 和 RNN。最后,将所有模型的输出和评估指标等作为输入数据进行可视化展示。

5.1　最佳模型的选择

探索区域模型预测性能,从而选择最合适的模型预测单车流量数据。用户可以通过控制面板中的“预测区域”下拉菜单框选择感兴趣的主题区域进行研究。当用户选择 topic _ 9 主题时,发现地图中的预测区域临近造成了字形重叠。此时用户可以通过调节控制面板中的“字形布局”调节滑条使重叠的字形分离开,如图 10 所示,为用户提供无遮挡的可视分析图。

为了得到区域最佳预测模型,需要对区域各模型的预测性能进行比较。通过观察字形第 3 层并将光标悬停在柱形图上查看具体数据,用户发现对于区域 167,RNN 模型的各个评估指标都优于其他模型,因此初步判定 RNN 为区域 167 的最佳预测模型。为了验证这一想法,用户点击区域编号在详细

图 10　局部区域模型预测结果的可视分析图

视图中观察 5 种模型的预测数据和实际数据，如图 11所示，通过数值的比较验证了这一猜想。

图 11　区域 167 的单车流量图

为了探究各时间序列特征对 RNN 模型预测的影响，观察区域 167 字形的第 2 层，如图 5 所示。用户发现工作日、温度和相对湿度这 3 个特征对模型预测具有促进作用，而节假日和天气对模型预测具有抑制作用。结合区域兴趣点的分布可以将该区域定义为交通小区。

RNN 模型在所有区域预测效果是否都是最佳，结合字形和详细视图中数值的比较，用户发现区域 231 的最佳预测模型为 LSTM。此外，观察发现不存在一个在不同区域都预测很准确的模型，即没有一个模型在任何场景下总是适用的。

5.2　区域相关性分析

为了进一步探究模型预测性能是否与地理空间位置有关，即空间相似的区域在模型预测性能上是否一致，本文使用各预测区域模型预测性能的排名进行表示。由于本文所选的预测模型在原理上是不同的，没有直接相关性，因此无法直接进行比较，所以从模型预测结果上来探索模型的预测性能，即使用模型预测数据与实际数据的相关性得分表示模型的预测性能，并利用各预测区域模型预测性能的排名表示模型预测性能是否与地理空间位置有关。

当用户选择主题后，预测区域会在投影视图中高亮显示。如图 12 所示，通过在投影视图中观察 topic _ 11 主题下区域的空间分布，用户发现区域 8299 和区域 8327 的距离较近且与区域 4894 的距离较远，因此选取这 3 个区域进行研究，点击这 3 个投影点观察这 3 个区域的相关性视图。

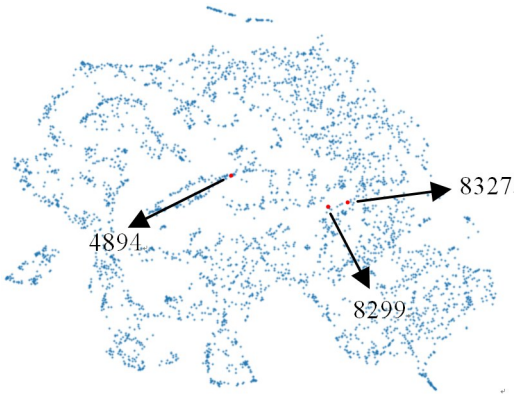


图 12　区域 4894、8299、8327 在向量空间中的位置

如图 13 所示，用户发现区域 4894 的模型预测性能排名依次为 RNN、ARIMA、LSTM、RF、SVM；区域 8299 的模型预测性能排名依次为 LSTM、RNN、ARIMA、SVM、RF；区域8327的模型预测性能排名

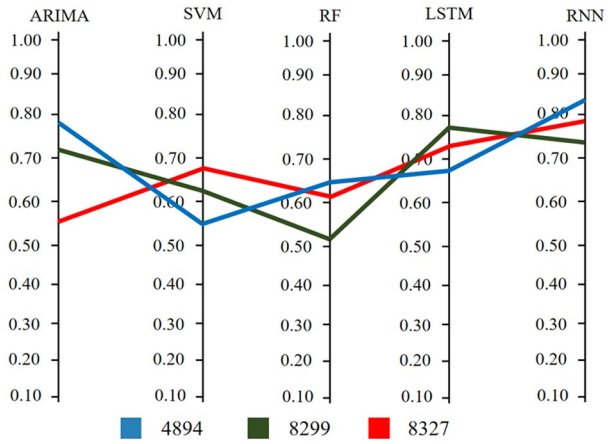


图 13　区域 4894、8299、8327 相关性视图

依次为 RNN、LSTM、SVM、RF、ARIMA。通过这 3 个区域模型预测性能的比较,用户发现模型预测性能的排名没有相关性,得出模型预测性能与所处的地理空间位置无关,即空间相似的区域在模型预测性能上也存在差异。

6 用户调查

为了评估本文可视化系统的有效性和实用性,针对各个视图共设计了 6 个问题并邀请了 20 位计算机相关专业的志愿者和 10 位非计算机相关专业的志愿者进行问卷调查。

在问卷调查开始之前,首先对每一位参与者演示本文系统的具体细节,然后针对不同的问题记录每一位参与者做出的回答,最后对所有参与者的肯定回答进行统计,如表 1 所示。

表 1 问卷调查及结果统计

视图	问题	结果
地图视图	能否快速定位区域所在的位置	30
字形视图	设计是否新颖	29
	能否直观比较区域多模型预测性能	29
详细视图	能否快速发现区域最佳预测模型	28
投影视图	能否快速定位相似区域	30
相关性视图	能否直观比较区域模型相关性排名	30

问卷调查结果显示,参与者普遍认为地图视图能够快速定位区域所在的位置,字形视图设计新颖且能够直观比较区域多模型预测性能,详细视图能够快速发现区域最佳预测模型,投影视图能够快速定位相似区域,相关性视图能够直观比较区域模型性相关性排名。因此本文可视化系统基本符合本文的任务要求。

7 结 论

本文针对共享单车的 OD 数据,设计了一套可视分析系统。该系统能够将多模型预测信息集成到字形中并在地理空间中展示,使用户可以通过交互式界面直观了解模型的预测性能和预测结果,从而指导用户在工作中选择合适的模型。其中 area2vec 方法能够帮助用户分析预测区域之间的空间关系,引入的网格布局算法可以有效地缓解地图中字形遮挡的问题。

虽然本系统基本达到了预期效果,但在系统可扩展性上仍存在一些局限性。首先,本文仅选取 5 种具有代表性的时间序列预测模型,但是由于模型种类众多,无法确定哪种模型预测的准确率更高。其次,在区域相似度的计算过程中,由于 OD 数据只包含起点 - 终点的定位且骑行距离有限,因此只能计算城市小范围内的区域相似度。最后,在字形布局上,由于本文每次仅选取 10 个预测区域进行研究,当包含更多预测区域时,地图上有限的空间无法避免所有预测区域的字形不重叠。未来的工作中,将在可视化系统中加入更多的预测模型。另外,可以考虑融合多种数据源提取城市区域的特征,计算城市部分区域相似度并扩展到整个城市。最后,在考虑更多地理区域的字形布局时,可以通过窗口的自适应功能和地图的缩放功能,实现更多区域预测模型的可视比较分析。

参考文献

[1] ZHAO J, YUE P. Spatiotemporal data cube modeling for integrated analysis of multi-source sensing data[C]// IEEE International Geoscience and Remote Sensing Symposium. Yokohama: IEEE, 2019:4791-4794.

[2] DETTI A H, STEINER N Y, BOUILLAUT L, et al. Fuel cell performance prediction using an auto-regressive moving-average model[C]//IEEE Vehicle Power and Propulsion Conference. Hanoi: IEEE, 2019:1-5.

[3] KHASHEI M, BIJARI M, ARDALI G A R. Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (ANNs)[J]. Neurocomputing, 2009,72(4-6):956-967.

[4] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999,9(3):293-300.

[5] QIAO M, LIU X, YUAN X. Snapshot spatial-temporal compressive imaging[J]. Optics Letters, 2020,45(7):1659-1662.

[6] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B,

et al. Conditional random fields as recurrent neural networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1529-1537.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[8] 杨瑞君,何立君,程燕, 等. 一种改进 TPOT 的工控入侵检测方法[J]. 小型微型计算机系统, 2022, 43(4): 767-772.

[9] 宋瑞蓉,路树华,王斌君,等. 基于 ABiLSTM 与 XGBoost 组合模型的交通时间预测[J]. 软件导刊, 2021, 20(8): 20-28.

[10] LEDELL E, POIRIER S. H2O autoML: Scalable automatic machine learning[C]// Proceedings of the 7th ICML Workshop on Automated Machine Learning. Virtual: ICML, 2020: 1-8.

[11] BeLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2012, 3: 993-1022.

[12] NEWMAN D, LAU J H, GRIESER K, et al. Automatic evaluation of topic coherence[C]// Proceedings of the North American Chapter of the Association for Computational Linguistics. Los Angeles: Association for Computational Linguistics, 2010: 100-108.

[13] WEI X, CROFT W B. LDA-based document models for ad-hoc retrieval[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Massachusetts: Association for Computing Machinery, 2006: 178-185.

[14] 边荣正, 张鉴, 周亮, 等. 面向复杂多流形高维数据的 t-SNE 降维方法[J]. 计算机辅助设计与图形学学报, 2021, 33(11): 116-119.

Multi-model visual comparative analysis for demand forecasting of shared bicycles

ZHANG Qiqi*, RAO Ning**, ZHU Sujia*, ZHA Meng*, SUN Guodao*

(*College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023)

(**School of Internet of Things, Hangzhou Polytechnic, Hangzhou 311402)

Abstract

Aiming at how to compare different models in different regions and choose the best model for prediction and whether there is a universally applicable best prediction model, a set of multi-model visual comparative analysis system for demand forecasting of shared bicycles is designed. First, using the latent Dirichlet allocation(LDA) model, representative prediction regions are selected, and corresponding glyphs are designed between the prediction regions, and the prediction performance of multiple models is compared by the glyphs. Second, a new method, area2vec, for calculating the similarity of urban spatial regions is proposed to compare the differences in prediction performance of models in similar regions. Then, a grid layout algorithm is designed to effectively alleviate the problem of glyph occlusion in the map. Finally, the effectiveness and usefulness of the system is confirmed by case studies and user surveys.

Key words: demand forecasting, latent Dirichlet allocation(LDA) model, word embedding, model comparison, visual analysis