# 智能交通背景下出租车载客收益预测

**作者**：程梁华，万德胜，周恩亦，黄旺辉，王彬，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：数据预处理、时序数据预测、注意力机制、时间注意力、空间注意力、长短期记忆网络、Seq2Seq模型

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 随着信息技术、物联网及汽车工业等不断发展，智能化逐渐成为交通运输的一个重要发展趋势，越来越多的基于智联网汽车的应用场景被提出。出租车（网约车等）作为智联网汽车中的重要部分，具有覆盖面积广、易管理的特点。因此，基于出租车运行中收集到的各种数据，利用算法进行载客收益预测，不仅优化该类车辆执行感知任务的效率，亦可以合理调度出租车分布，从而提升司机收益，提升交通效率等。本案例基于重庆市出租车一个月的轨迹数据，使用深度学习算法预测城市未来一段时间内的载客收益信息。本案例的核心内容有：（1）对重庆市出租车轨迹数据集进行预处理，以适应载客收益预测任务的需要；（2）使用一种基于双注意力机制的循环神经网络进行载客收益的预测，并与长短期记忆网络、Seq2Seq模型、以及时间注意力网络方法进行对比分析。

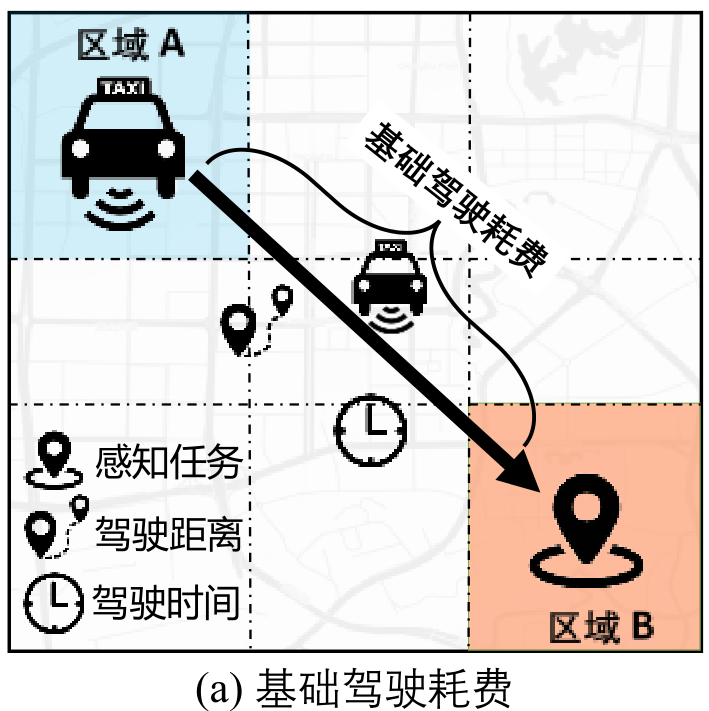
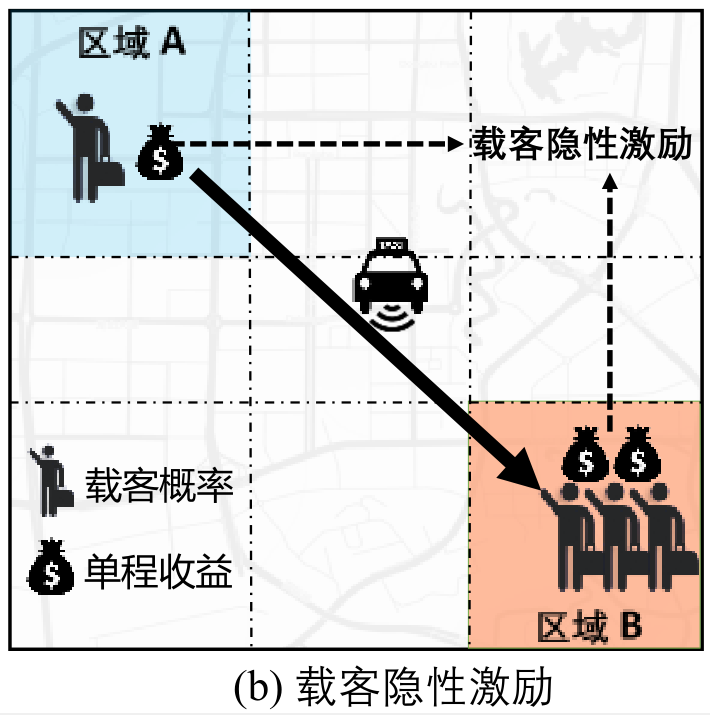
**关键词**：智能交通，载客收益预测，数据挖掘，深度学习

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为智能交通中的智能化交通调度，以出租车的载客收益预测作为具体问题，需引导学生进行的主要内容有：（1）对原始出租车轨迹数据集进行数据预处理，以适应载客收益预测任务的需要；（2）选取适当的深度学习方法，进行出租车载客收益预测任务，本案例以基于双注意力机制的循环神经网络、长短期记忆网络、Seq2Seq模型、以及时间注意力网络作为引导案例。

## 2 背景介绍

随着智能汽车、辅助/无人驾驶汽车、电动汽车的飞速发展和爆炸式普及，当前汽车集成越来越多的传感器，拥有越来越强大的感知、计算和通信能力。近年来，涌现很多基于智联网汽车感知的实际应用。作为智联网汽车中重要一类，出租车、网约车等被大规模被广泛使用，为智联网汽车感知带来两个独特优势：（1）覆盖面积广；（2）大量的空闲时间。因此，智联网汽车司机拥有充足的空闲时间执行感知任务，汽车根据任务分配完成感知任务，并得到平台给予的感知报酬，所以如何基于历史数据来预测得到未来的载客收益是关键，也就是说，利用算法实现载客收益预测，然后基于这个预测结果，实现感知任务的实时最优分配。

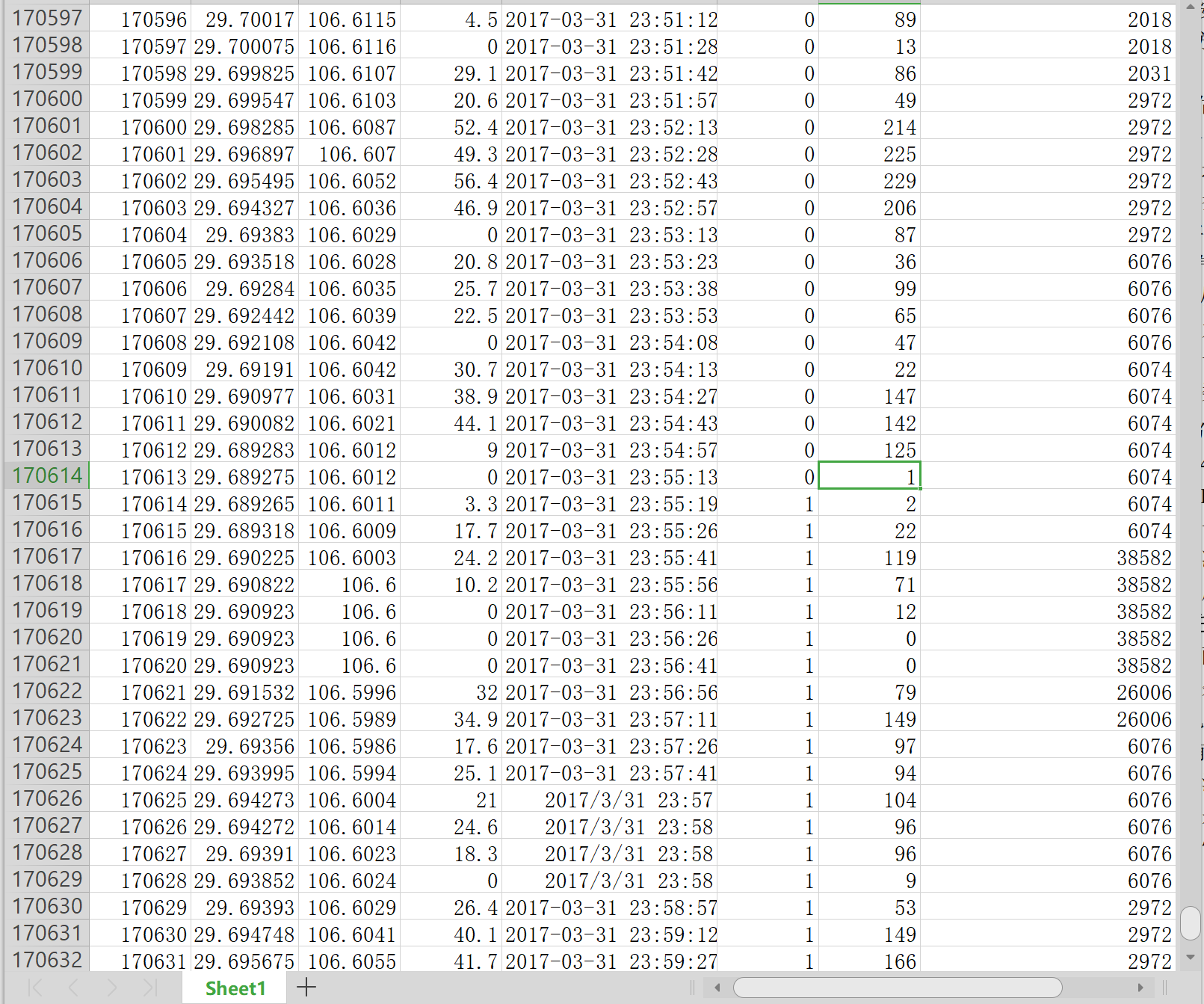


除优化感知任务分配外，基于出租车、网约车等车辆运行中收集到的各种数据，利用算法进行载客收益预测，亦可以提升司机收益，优化交通调度，提升交通效率。

## 3 内容

**3.1 数据集及其预处理**

本案例所采用的数据集为重庆市出租车一个月（2017/03/01/-2017/03/31）的轨迹数据，包括12493辆出租车，覆盖了重庆市约 25472 km2的城市面积。经初步清洗过的数据如下图所示：



在该数据集中，每条记录主要包含汽车ID、时间、GPS位置、速度以及载客/空载状态。数据采样频率为15秒，总存储大小约为 92GB，数据各字段及含义如下表所示：

|  |  |
| --- | --- |
| **字段** | **含义** |
| *TaxiID* | 汽车ID |
| *TimeStamp* | 记录时间 |
| *Lat* | 纬度 |
| *Lon* | 经度 |
| *Speed* | 速度 |
| *Customer* | 是否载客 |

本案例旨在基于实际大规模智联网汽车数据集，利用深度学习算法预测城市各个区域未来多个时间段的载客收益热力图，因此需要计算历史载客收益热力图。

首先按照一定粒度，将城市均匀划分成Z个互不相交的子区域，用zi表示第i个子区域。与之类似，将整个时间区间划分成T个长度相同的时段，每个时段用t表示。某一时段某一区域的载客收益由司机所在区域接客概率和单程平均收益决定。用和分别表示第t时段在区域 zi的接客概率和单程平均收益。因此，智联网汽车的载客收益可表示为：



在得到每一个时段在每一个区域的载客收益后，即可得到历史载客收益热力图。由于在同一时段不同区域的载客收益具有差异性，我们用载客收益热力图Ht表示第t时段所有区域的载客收益，即

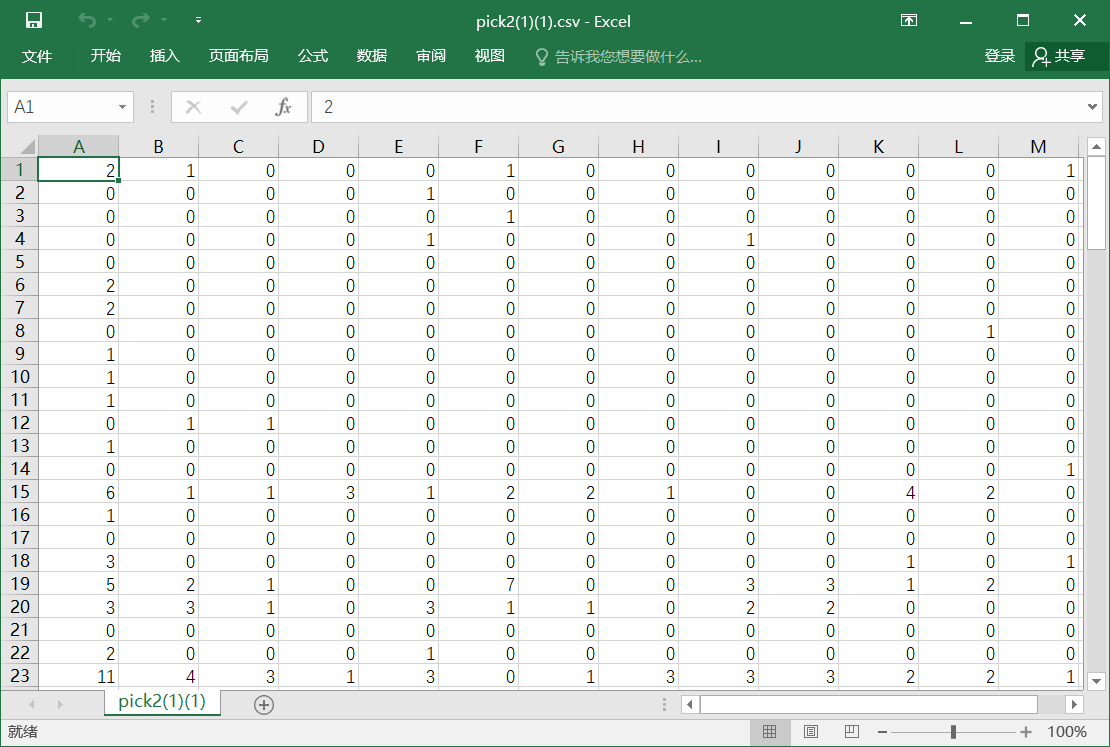
。

因此，智联网汽车在h个历史时段的载客收益可表示为 h 帧热力图，即 。

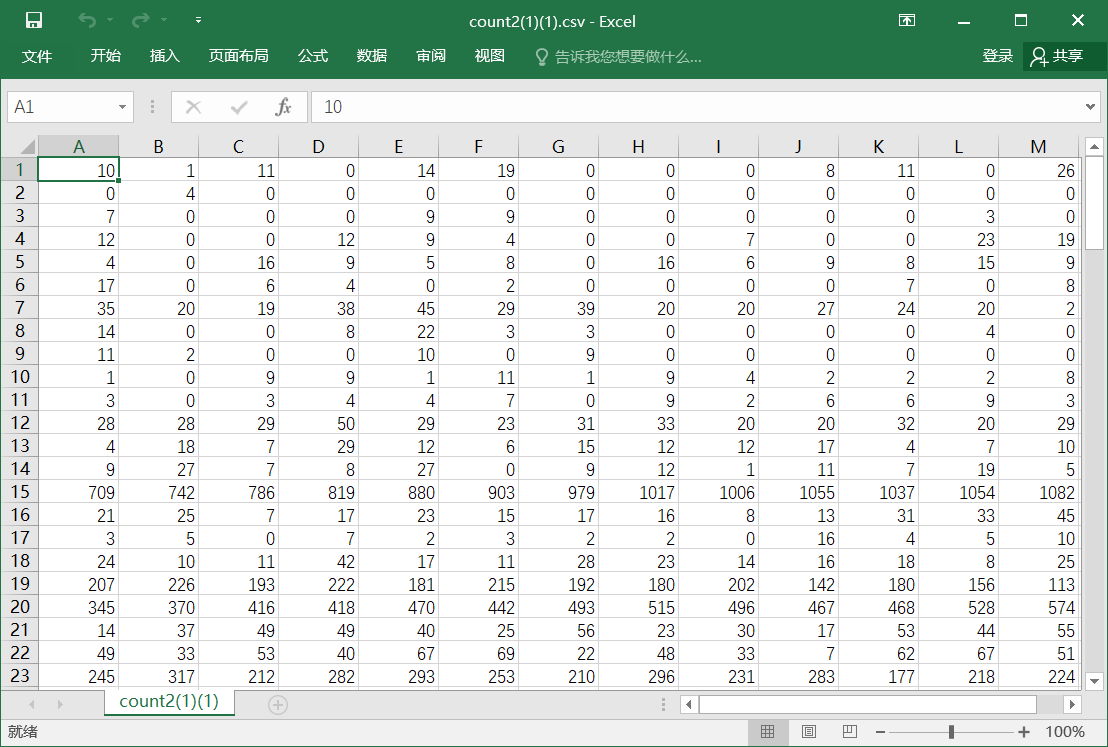
在本案例中，数据集预处理的具体处理步骤如下：

（1）选择重庆市江北区作为实验区域，将它均匀划分成125个子区域(每个子区域大小约为2km2 )。以5min为粒度，将一个月划分为8928个时段。

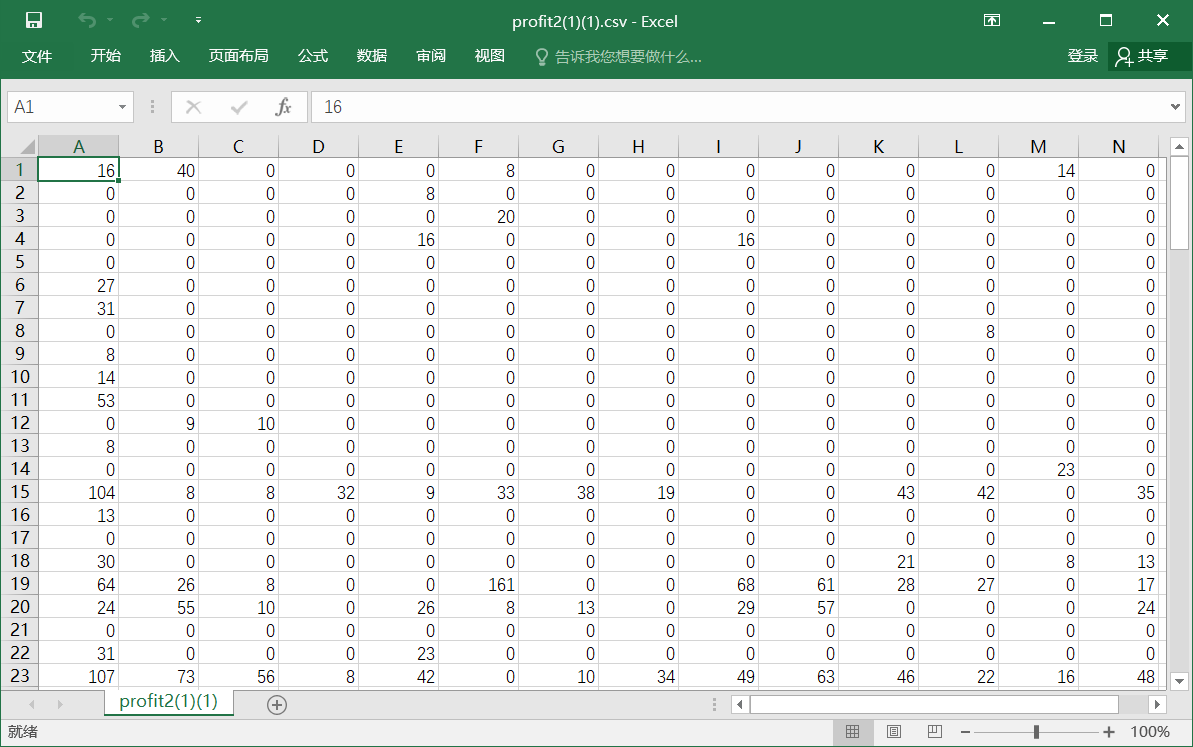
（2）计算每一个子区域中每一个时段的总接客汽车数，将第t时段在区域 zi的总接客汽车数表示为。Customer字段由0变为1表示出租车成功接客，t时段内在区域zi中，载客状态由空载变为载客的汽车总数即为所求，部分结果如下图所示：



（3）计算每一个子区域中每一个时段的总空闲汽车数，第*t*时段在区域 *zi*的总空闲汽车数表示为。*t*时段内若某汽车在区域*zi*中出现，且处于空闲状态（第t时段内该车只要出现过空闲状态，即算作空闲汽车），则将其累加至总空闲汽车数中，部分计算结果如下图所示：



（4）计算每一个子区域中每一个时段的总单程收益，第*t*时段在区域*zi*的总单程收益表示为。对于成功载客的车辆，基于行程轨迹，通过计算轨迹中采样点之间的球面距离，得到行程距离，并基于重庆市出租车计价规则得到载客收益，部分计算结果如下：



（5）计算载客概率和单程平均收益。对于每一个子区域中每一个时段，用该时段子区域内总接客汽车数即总接客成功载客车辆数与总空闲车辆数之比作为载客概率。则第*t*时段在区域 *zi*的接客概率表示为：



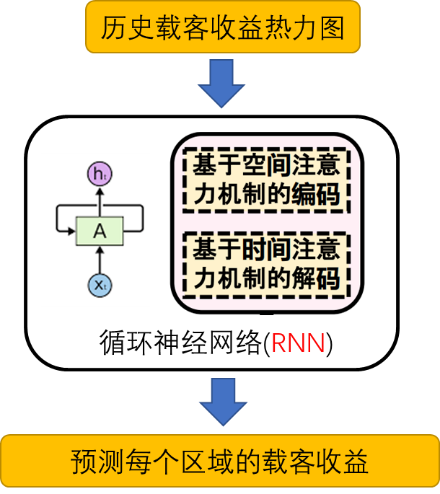
显然，单程平均收益即为总单程收益与总接客汽车数之比，则第*t*时段在区域 *zi*的单程平均收益表示为：



由载客概率和单程平均收益得到每一个子区域中每一时段的载客收益，进而可以得到历史载客收益热力图，基于该热力图和各子区域的拓扑邻接图，即能预测未来每个子区域的载客收益。

**3.2 使用双注意力机制RNN算法进行载客收益预测**

我们以一个子区域为例，详细介绍如何基于双注意力机制的循环神经网络预测载客收益。如下图所示，该方法利用基于长短期记忆的编码器－解码器架构（LSTM based Encoder-Decoder）和时空注意力机制（Temporal-Spatial Attentions），主要包括基于时间注意力机制的编码和基于空间注意力机制的解码两部分。



在上图中，基于空间注意力机制的编码利用空间注意力机制将邻近区域的载客收益热力图编码为特征向量，称为隐藏状态；基于时间注意力机制的解码利用基于LSTM的循环神经网络解码每个时段编码生成的隐藏状态，从而预测载客收益。同时，在解码过程中，利用时间注意力机制，从所有时段中挖掘学习得到最相关的隐藏状态。

对于数据集，本案例选择重庆市江北区作为实验区域，将其均匀划分成125个2km2的子区域（每个子区域大小约为。对于每一个子区域中每一个时段，用该时段子区域内成功载客车辆数与总空闲车辆数之比作为载客概率。

首先，对于成功载客的车辆，基于行程轨迹，通过计算轨迹中采样点之间的球面距离，得到行程距离，并基于重庆市出租车计价规则得到载客收益。然后，计算该时段子区域内所有成功载客车辆的平均载客收益，作为单程收益期望。最后，我们使用预测平均误差（Mean Absolute Error，MAE），即预测值与真实值之间的平均绝对误差，用于评估基于深度学习算法的载客收益预测的准确性。

同时，我们也使用下述三种算法进行对比：

（1）LSTM：即长短期记忆网络深度学习算法，是递归神经网络的变体，其特点在于可以同时捕获序列中的长期与短期变化特征。

（2）Seq2Seq：一种基于递归神经网络的序列转换模型，利用编码器－解码器结构，主要包括编码器、上下文向量和解码器三个主要模块。

（3）TAtt：基于时间注意力机制的循环神经网络。

在该案例中，我们使用Keras工具包实现上述深度学习算法，源代码见附录。每种算法得到的MAE指标结果如下表所示，其中双注意力机制算法取得最佳结果。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 算法名称 | | 双注意力机制算法 | TAtt | Seq2Seq | LSTM |
| MAE | 重庆市数据集 | 153.80 | 184.59 | 216.16 | 178.65 |

同时，在本案例中可以设计一系列参数实验，以确定多种因素对载客收益预测任务的影响，并选取最佳的参数设定。参数实验的内容与结果如下图所示：

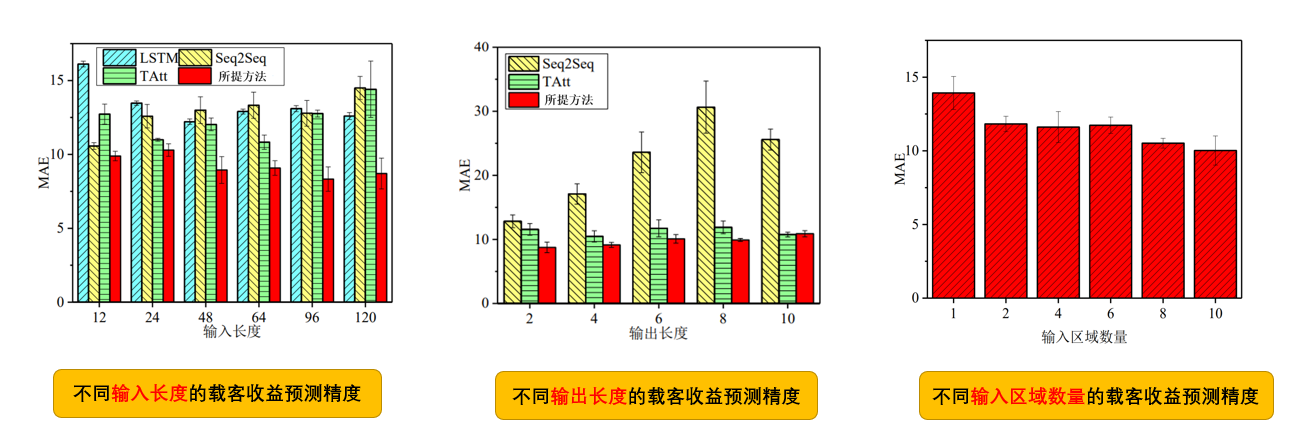


图1 图2 图3

在上图中，输入和输出长度分别表示深度学习算法输入数据的长度和输出数据的时间长度；输入区域数量表示深度学习算法邻近区域的个数。（这里输入长度单位值在设置为5分钟，输入长度12即为5\*12min=60min，输入区域为预测区域的邻近区域个数，例如输入区域数量2为预测子区域的邻近区域个数输入值为2，这与基于空间注意力机制是对应的）

从图1中可以看出，随着输入长度的增加，双注意力机制算法预测精度逐渐提高。当输入长度为96时，预测精度最高，即MAE=8.33。同时，相比于LSTM、Seq2Seq和TAtt，双注意力机制算法分别平均提高预测精度约31.2%、27.1%和25.1%。

图2结果表明，随着输出长度的增加，双注意力机制算法与TAtt性能都比较稳定。同时，双注意力机制算法的预测精度比 Seq2Seq 和 TAtt 分别平均提高 55.5%和 13.5%。

图3是因为由于双注意力机制算法采用时空双注意力机制，因此评估邻近区域输入数量对算法的影响。我们可以看出，随着输入区域数量的增加，双注意力机制算法的预测精度逐步提高。当输入的邻近区域数量为10时，预测误差可以达到MAE=10.02。并且，不同邻近区域的加入对预测结果的影响不相同，主要原因如下：对于目标区域的预测，加入部分相关性较强的邻近区域可提高它的预测性能；反之，加入相关性弱的区域则会给模型学习带来更多的噪声，从而降低预测精度。

**3.3 案例拓展**

（1）系统实现： 本案例仅仅是停留在数据与算法层面，进一步可以基于该预测算法对预测城市实现原型系统，根据城市出租车的实时GPS数据与载客数据，将预测的载客收益进行实时显示，以便于出租车公司进行调控以提高收益。

（2）对任务的进一步拓展：考虑对算法进一步拓展，基于算法的载客收益预测结果，实现出租车资源的实时最优分配，以提高交通效率，并可进行高效的智联网感知任务。当然，最优分配问题往往是一种NP-hard 问题，具有指数级时间复杂度，因此需要设计具备可接受时间复杂度的算法寻找较优的分配方案。

## 4 小结

该案例的选取智能交通中的出租车的载客收益预测作为主要问题，使用重庆市出租车轨迹数据，结合数据挖掘、深度学习的相关方法，进行出租车的载客收益预测。主要内容包括原始数据集的预处理，历史载客收益的计算，以及使用多种深度学习算法进行预测。该案例主题新颖，结合了智能交通的现实需求与大数据分析与挖掘的多种理论与技术，可以充分增强学生的实践能力与理论基础。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、机器学习或深度学习方法完成任务，使用其他公开数据集进行实验研究，并考虑针对实际应用的进一步拓展。

## 附录

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），Keras与TensorFlow（深度学习算法库），matplotlib（可视化绘图库）。

3. 由于数据版权原因，本案例仅提供重庆出租车轨迹数据集的部分示例数据。出于教学与实践需要，本案例给出以下公开出租车轨迹数据集，以供学生选择使用：

（1）T-Drive，包含北京市2008年02月02日至02月08日共一周时间内10357辆出租车的轨迹数据，由微软提供，下载链接：https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/

（2）The epfl/mobility dataset，包含美国旧金山湾区2008年05月17日至2008年06月10日大约500辆出租车的GPS轨迹数据，下载链接：http://crawdad.org/epfl/mobility/20090224/

（3）Smart City Research Group，包括上海市出租车、公交车数据等，来源于香港科技大学，下载链接：https://cse.hkust.edu.hk/scrg/

（4）中华人民共和国交通运输部交通指数，中华人民共和国交通运输部下属交通数据分享平台，具有全国多个城市的出租车GPS轨迹数据，平台网址：https://www.mot.gov.cn/sjkf/

4. 本案例参考文献如下：

[1] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016.

参考内容：深度学习模型。