(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109670277 B
(45)授权公告日 2022.09.09

(21)申请号 201910123626.4

(22)申请日 2019.02.19

(65)同一申请的已公布的文献号
申请公布号 CN 109670277 A

(43)申请公布日 2019.04.23

(73)专利权人 南京邮电大学
地址 210023 江苏省南京市亚东新城区文苑路9号

(72)发明人 邹志强 杨浩宇 吴家皋 蔡韬 王兴源

(74)专利代理机构 南京瑞弘专利商标事务所(普通合伙) 32249
专利代理师 张耀文

(51)Int.Cl.
G06F 30/27(2020.01)
G06Q 10/04(2012.01)
G06N 3/08(2006.01)

(56)对比文件
CN 106981198 A,2017.07.25

审查员 朱琳玲

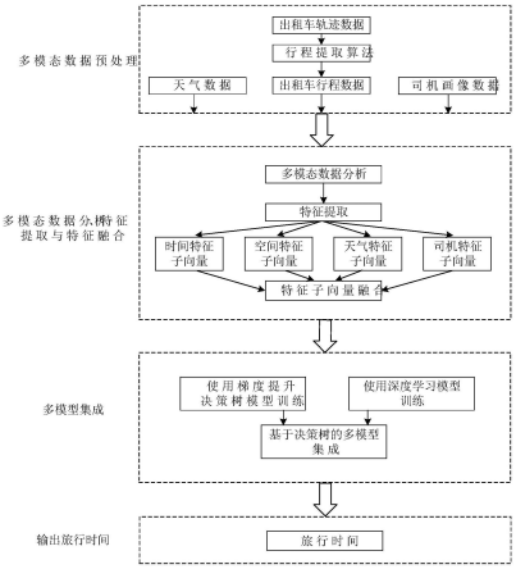
权利要求书3页 说明书8页 附图3页

(54)发明名称

一种基于多模态数据融合与多模型集成的旅行时间预测方法

(57)摘要

本发明公开了一种基于多模态数据融合与多模型集成的旅行时间预测方法,包括:多模态数据预处理模块:从出租车GPS轨迹数据中根据载客状态提取出租车行程数据;多模态数据分析、特征提取与特征融合模块:从出租车轨迹数据、天气数据、司机画像数据等领域分别提取相应的特征子向量,并完成特征拼接;多模型集成模块:分别建立梯度提升决策树模型和深度神经网络模型,并使用决策树模型对以上模型的预测结果进行集成。本发明的旅行时间预测方法融合了出租车轨迹数据、天气数据、司机画像数据等多模态数据,充分提取与挖掘对旅行时间有影响的因素,建立了基于决策树的集成模型,使得本发明以较小的计算代价获得了较高的行程旅行时间预测准确率。



CN 109670277 B

1.一种基于多模态数据融合与多模型集成的旅行时间预测方法,其特征在于,包括以下步骤:

a.多模态数据预处理

采用行程提取算法对出租车轨迹数据进行预处理,提取出租车行程数据;

b.多模态数据分析、特征提取与特征融合

通过分析包括出租车行程数据、天气数据、司机画像数据的多模态数据,分别提取相应的特征子向量,将输入特征的离散特征使用独热编码,与其他连续型特征拼接构成特征子向量,完成特征拼接;

c.多模型集成

所述模型建立过程包括子模型建立和模型集成两部分:将描述行程的特征进行特定的处理,分别输入到梯度提升决策树模型和深度神经网络模型这两个子模型中,然后基于子模型的预测结果建立基于决策树的集成模型,最终使用该模型预测得到行程的旅行时间;

所述c.多模型集成包括:

c1.建立梯度提升决策树模型:

输入:样本集$\{trip_j|trip=((trip_s,trip_t,trip_w,trip_d),trip_{traveltime}),j\in(1,all)\}$,其中all为样本的总数目

输出:梯度提升决策树模型

1.1)初始化第一棵决策树$T_0(trip)=0$;

1.2)设置参与梯度提升决策树模型训练的决策树的总数目为M,分别对m=1,2,...,M,在生成第m棵决策树时,计算每个样本的残差r_{mj},其中$r_{mj}=trip_{traveltimej}-T_{m-1}(trip_j)$,j=1,2,...,all,通过使用回归决策树模型拟合每个样本的残差r_{mj},学习得到第m棵决策树$T_m(trip;\Theta_m)$,其中Θ_m为第m棵决策树中的参数;

1.3)更新$f_M(trip)=\sum_{m=1}^{M}T_m(trip;\Theta_m)$;

1.4)得到梯度提升决策树模型$f_M(trip)$,其输出就是使用梯度提升决策树模型预测出的旅行时间;

c2.建立深度神经网络模型:

2.1)基于空间网格划分进一步提取空间特征

将城市细粒度划分为200*200的网格,遍历行程数据集中的行程,将行程的起始地与目的地归属到对应的网格中,获得出发地编号和目的地编号;

2.2)建立深度神经网络模型

对出发地编号、目的地编号以及司机编号通过实体嵌入层将正整数编号转换为具有固定大小的向量,获得的实体嵌入向量分别经过多层隐藏层处理后和特征向量提取模块提取出的其余特征进行拼接,通过多层全连接网络进行训练;

2.3)损失函数及优化方法

在上述模型构建后,训练该模型,其中设置训练样本的批大小为512,选用均方误差损失函数,然后使用激活函数为修正线性单元,由激活函数完成非线性变换,通过Adam优化算法进行参数寻优,其中学习率为0.001,衰减项1e-08,动量0.9,迭代次数设置为100,使用5折交叉验证的获得最佳模型g(trip);

c3.建立基于决策树的集成模型：

对于训练集中的所有样本{$trip_j$|trip＝(($trip_s$,$trip_t$,$trip_w$,$trip_d$),$trip_{traveltime}$),j∈(1,all)}，将$trip_j$分别输入训练后的梯度提升决策树模型f_M(trip)得到预测结果y_{GBDT}和深度神经网络模型g(trip)得到预测结果y_{DNN}，构建出集成模型的新的训练样本集((y_{GBDT},y_{DNN}),$trip_{traveltime}$)，在新的样本集所在的输入空间中，使用最小二乘法对切分变量y_{GBDT}与y_{DNN}遍历并计算每个切分点的损失值，选择具有最小损失值的切分变量及切分点，递归地将每个区域划分为两个子区域并决定每个子区域上的输出值，构建回归决策树模型，该模型的输出就是预测的旅行时间。

2.根据权利要求1所述的一种基于多模态数据融合与多模型集成的旅行时间预测方法，其特征在于，所述行程提取算法具体包括：

输入：一辆出租车轨迹序列T＝{P_1,P_2,P_3,...,P_n}

输出：行程数据集

a1.对轨迹序列进行遍历，设置循环变量i从1到n-1，n表示轨迹点的总数，初始时i＝1，行程状态位为0；

a2.当P_i的载客状态为1，跳转至a3，否则跳转至a4；

a3.当行程状态位为1时，跳转至a6，否则将P_i记录为行程的起始点，行程状态位置1，跳转至a6；

a4.当行程状态位为1时，跳转至a5，否则跳转至a6；

a5.当P_{i+1}载客状态位为1时当前行程记录完毕，将该行程计入行程数据集中，行程状态位置0，否则跳转到a6；

a6.执行i＝i+1；

a7.当i<n时，跳转至a2，否则完成行程数据集的提取。

3.根据权利要求2所述的一种基于多模态数据融合与多模型集成的旅行时间预测方法，其特征在于，所述b.多模态数据分析、特征提取与特征融合包括：

b1.分析出租车行程数据，提取行程空间特征：根据半正矢公式计算在地球上两点之间的大圆距离，进而提取曼哈顿距离和方位角，使用k-means方法进行空间聚类，将聚类后得到的类簇编号使用独热方式编码，构成空间特征子向量$trip_s$；

b2.分析出租车行程数据，提取行程时间特征：分别提取行程出发时间的离散型周期性信息以及状态信息，构成时间特征子向量$trip_t$；

b3.分析天气数据，提取天气特征：根据天气状况将天气划分为不同的等级，构成天气特征子向量$trip_w$；

b4.分析司机画像数据，提取司机特征：获取驾驶出租车的司机画像信息，构成司机特征子向量$trip_d$；

b5.将行程空间特征子向量、行程时间特征子向量、天气特征子向量与司机特征子向量进行拼接，并与该行程对应的旅行时间$trip_{traveltime}$组成描述行程信息的完整特征向量：trip＝(($trip_s$,$trip_t$,$trip_w$,$trip_d$),$trip_{traveltime}$)。

4.根据权利要求3所述的一种基于多模态数据融合与多模型集成的旅行时间预测方法，其特征在于，所述深度神经网络模型具体包括：

第一部分：输入数据为出发地编号和目的地编号，包括实体嵌入层和三层隐藏层，嵌入

层输入维度为200,输出维度为16的数据,隐藏层每层有256个神经元节点;

第二部分:输入数据为司机编号1*1,包括实体嵌入层和三层隐藏层,嵌入层输入维度为NumDriver,其中NumDriver为司机数目,输出维度为32的数据;

第三部分:将第一部分输出的特征向量、第二部分输出的特征向量与其余特征组成的特征向量进行拼接;

第四部分:是一个包含512个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为512的数据;

第五部分:是一个包含256个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为256的数据;

第六部分:是一个包含128个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为128的数据;

第七部分:是一个包含64个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为64的数据;

第八部分:是一个包含1个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为1的数据,该数据就是使用深度神经网络模型预测出的旅行时间。

一种基于多模态数据融合与多模型集成的旅行时间预测方法

技术领域

[0001] 本发明涉及一种基于多模态数据融合与多模型集成的旅行时间预测方法,属于智能交通信息处理技术领域。

背景技术

[0002] 在智能交通系统(Intelligent Traffic Systems,ITS)和基于位置的服务中,旅行时间预测是一个关键、复杂且具有挑战性的问题。交通监管机构可以通过旅行时间间接了解城市流量的变化,实时旅行时间的预测和提示一定程度上还可以缓解交通拥堵,旅行时间估计为ITS中的交通流量控制提供了有效的决策支持。旅行时间预测也是地图导航与出行服务软件的重要模块,如百度地图,滴滴出行等,人们可以通过旅行时间估计来合理安排和规划自己的出行活动。

[0003] 目前旅行时间预测方法可以分为两种类型:包括基于路段的方法和基于路径的方法。传统的旅行时间预测方法最初依赖于环路探测器收集的车辆行驶数据,结合路段的道路交通状况,根据路段实际行驶过程中大量驾驶员的旅行时间消耗来估计特定路段的行驶时间。然而,在大型运输网络中安装和维护环路检测器的成本很高,因此该解决方案无法大规模有效扩展。随着GPS(Global Positioning System,GPS)技术的进步,安装有GPS设备的出租车收集了大量的出租车轨迹数据,这些数据逐渐开始用于估计路段的行驶速度和行程时间。旅行时间预测方法初步研究侧重于基于路段的研究,固定路段的旅行时间预测被认为是时间序列预测问题。经典的时间序列预测模型,包括差分整合移动平均自回归模型、卡尔曼滤波器以及长短时记忆网络都曾被用来解决旅行时间预测的问题。然而在现实生活中,车辆实际行驶的路线由多个路段组成。基于路径的方法分别计算每个路段的行程时间并进行累积,这些方法虽然考虑了路段的连续性,但是没有考虑交叉路口,交通灯等的影响,已有的基于路段和路径的预测方法将不可避免地产生较大误差。

[0004] 城市的交通网络涉及极其复杂的情况,然而上述方法仅基于现实世界的物理模型,它还需要考虑交通系统的空间特征,例如交通信号灯和经过道路的数目、速度限制等。近年来随着交通数据的规模越来越大以及机器学习技术的发展,基于行程起讫点的方法尝试忽略行驶过程中的轨迹信息,重新思考旅行时间预测问题。与基于路段和路线的方法不同,基于行程起讫点的旅行时间预测方法主要基于以下三个立足点:(1)部分城市没有完整的轨迹数据集,数据集中仅包括行程的起始地和目的地;(2)如果考虑路线信息,问题将转化为路线选择和时间计算两部分。例如,百度、高德等地图数据服务提供商首先预测路线,然后使用基于路径的方法预测相应的时间,这将会带来较大的计算代价。由于密集的城市道路网络和复杂多变的交通条件,很难考虑行驶过程中可能遇到的交通状况,特别在交叉路口等。另外,乘客在出行时,更关注行程的旅行时间,而非具体的行驶路线。

发明内容

[0005] 发明目的:旅行时间预测问题存在多种复杂情况,为了克服现有技术中存在的不

足，本发明提供一种基于多模态数据融合与多模型集成的旅行时间预测方法，以解决城市中任意起讫点行程的旅行时间难以获得精准预测结果的问题。

[0006] 要实现上述发明内容，必须要解决几个核心问题：(1)目前存在的旅行时间预测的方法中考虑的数据领域相对单一。不仅应该考虑交通系统的空间特征，还要考虑时间特征，例如早晚峰值的频繁拥堵，以及交通事故造成的意外拥堵。因为交通系统由人和车组成，并受外部因素的影响，旅行时间预测还需要引入个性化特征和外部特征的建模；(2)旅行时间预测的研究大多局限于单一模型。由于交通流变化过程是一个实时的、非线性的、高维的、非平稳的随机过程，旅行时间变化的随机性和不确定性变得越来越强，单个模型容易出现偏差且较难消除。由于单模型固有的缺陷，对于ITS中的各种情况，很难做出良好的预测。

[0007] 技术方案：为实现上述目的，本发明采用的技术方案为：

[0008] 一种基于多模态数据融合与多模型集成的旅行时间预测方法，其特征在于，包括以下步骤：

[0009] a.多模态数据预处理

[0010] 因为本方法忽略了行程过程中的轨迹点，因此首先需要采用行程提取算法对出租车轨迹数据进行预处理，忽略异常轨迹点并根据载客状态提取出租车行程数据；

[0011] b.多模态数据分析、特征提取与特征融合

[0012] 旅行时间受到多个因素的影响，本方法从出租车轨迹数据、天气数据、司机画像数据等领域分别提取相应的特征子向量，将输入特征的离散特征使用独热编码，与其他连续型特征拼接构成特征子向量，完成特征拼接；

[0013] c.多模型集成

[0014] 所述模型建立过程包括子模型建立和模型集成两部分：将描述行程的特征进行特定的处理，分别输入到梯度提升决策树模型和深度神经网络模型中，然后基于子模型的预测结果建立以决策树为基础的集成模型，最终使用该模型预测得到行程的旅行时间。

[0015] 进一步的，所述行程提取算法具体包括：

[0016] 输入：一辆出租车轨迹序列$T=\{P_1,P_2,P_3,...,P_n\}$

[0017] 输出：行程数据集

[0018] a1.对轨迹序列进行遍历，设置循环变量i从1到n-1，n表示轨迹点的总数，初始时i＝1，行程状态位为0；

[0019] a2.当P_i的载客状态为1，跳转至a3，否则跳转至a4；

[0020] a3.当行程状态位为1时，跳转至a6，否则将P_i记录为行程的起始点，行程状态位置1，跳转至a6；

[0021] a4.当行程状态位为1时，跳转至a5，否则跳转至a6；

[0022] a5.当P_{i+1}载客状态位为1时当前行程记录完毕，将该行程计入行程数据集中，行程状态位置0，否则跳转到a6；

[0023] a6.执行i＝i+1；

[0024] a7.当i＜n时，跳转至a2，否则完成行程数据集的提取。

[0025] 进一步的，所述b.多模态数据分析、特征提取与特征融合具体包括：

[0026] b1.分析出租车行程数据，提取行程空间特征：根据半正矢公式计算在地球上两点之间的大圆距离，进而提取曼哈顿距离和方位角，使用k-means方法进行空间聚类等，构成

空间特征子向量$trip_s$；

[0027] b2.分析出租车行程数据，提取行程时间特征：分别提取行程出发时间的月份、日期等离散型周期性信息以及是否为工作日、节假日等状态信息，构成时间特征子向量$trip_t$；

[0028] b3.分析天气数据，提取天气特征：根据天气状况将天气划分为不同的等级，使用独热编码方法得到天气数据的特征，构成天气特征子向量$trip_w$；

[0029] b4.分析司机画像数据，提取司机特征：获取驾驶出租车的司机编号、性别、年龄、驾龄等信息(从司机画像数据中针对司机编号使用实体嵌入式处理方法，得到每一个司机的特征)，构成司机特征子向量$trip_d$；

[0030] b5.将行程空间特征子向量、行程时间特征子向量、天气特征子向量与司机画像特征子向量进行拼接，并与该行程对应的旅行时间$trip_{traveltime}$组成描述行程信息的完整特征向量：$trip=((trip_s,trip_t,trip_w,trip_d),trip_{traveltime})$。

[0031] 进一步的，所述c.多模型集成具体包括：

[0032] c1.建立梯度提升决策树模型：梯度提升决策树模型结合人工挖掘的特征来获得非线性映射和高阶特征，此处设置梯度提升决策树的损失函数为均方误差损失函数；

[0033] c2.建立深度神经网络模型：深度神经网络适用于处理高维稀疏特征；

[0034] c3.建立基于决策树的集成模型：集成模型指训练多个基础模型并将它们组合起来，这样的算法可以比单个模型实现更好的预测结果。

[0035] 其中，c1.梯度提升决策树模型具体包括：

[0036] 输入：样本集$\{trip_j|trip=((trip_s,trip_t,trip_w,trip_d),trip_{traveltime}),j\in(1,all)\}$，其中all为样本的总数目

[0037] 输出：梯度提升决策树模型

[0038] 1.1)初始化第一棵决策树$T_0(trip)=0$；

[0039] 1.2)设置参与梯度提升决策树模型训练的决策树的总数目为M，分别对$m=1,2,...,M$，在生成第m棵决策树时，计算每个样本的残差r_{mj}，其中 $r_{mj}=trip_{traveltime_j}-T_{m-1}(trip_j)$，$j=1,2,...,all$，通过使用回归决策树模型拟合每个样本的残差$r_{mj}$，学习得到第m棵决策树$T_m(trip;\Theta_m)$，其中$\Theta_m$为第m棵决策树中的参数；

[0040] 1.3)更新$f_M(trip)=\sum_{m=1}^{M}T_m(trip;\Theta_m)$；

[0041] 1.4)得到梯度提升决策树模型$f_M(trip)$，其输出就是使用梯度提升决策树模型预测出的旅行时间；

[0042] c2.深度神经网络模型具体包括：

[0043] 2.1)基于空间网格划分进一步提取空间特征

[0044] 为了细化不同区域的特性，将城市细粒度划分为200*200的网格。遍历数据集中的行程，将行程的起始地与目的地归属到对应的网格中，获得出发地编号和目的地编号；

[0045] 2.2)建立深度神经网络模型

[0046] 对起始地编号、目的地编号以及司机编号通过实体嵌入层将正整数编号转换为具有固定大小的向量，获得到的实体嵌入向量分别经过多层隐藏层处理后和特征向量提取模块提取出的其余特征进行拼接，通过多层全连接网络进行训练；

[0047] 2.3)损失函数及优化方法

[0048] 在上述模型构建后,需要训练该模型,其中设置训练样本的批大小为512,选用均方误差损失函数,然后使用激活函数为修正线性单元(Rectified Linear Unit,ReLU),由激活函数完成非线性变换,增强本模型对特征的学习能力,通过Adam优化算法进行参数寻优以最小化损失函数,其中学习率为0.001,衰减项1e-08,动量0.9。迭代次数设置为100,使用5折交叉验证的获得最佳模型g(trip)。

[0049] c3.基于决策树的集成模型具体包括:

[0050] 对于训练集中的所有样本$\{trip_j | trip=((trip_s,trip_t,trip_w,trip_d),trip_{traveltime}),j\in(1,all)\}$,将$trip_j$分别输入训练后的梯度提升决策树模型$f_M(trip)$得到预测结果$y_{GBDT}$和深度神经网络模型g(trip)得到预测结果$y_{DNN}$,构建出集成模型的新的训练样本集$((y_{GBDr},y_{DNN}),trip_{traveltime})$,在新的样本集所在的输入空间中,使用最小二乘法对切分变量y_{GBDT}与y_{DNN}遍历并计算每个切分点的损失值,选择具有最小损失值的切分变量及切分点,递归地将每个区域划分为两个子区域并决定每个子区域上的输出值,构建回归决策树模型,该模型的输出就是预测的旅行时间。

[0051] 进一步的,所述深度神经网络模型具体包括:

[0052] 第一部分:输入数据为起始地编号和目的地编号,包括实体嵌入层和三层隐藏层,嵌入层输入维度为200(空间特征提取中将城市细粒度划分为的200*200的网格),输出维度为16的数据,隐藏层每层有256个神经元节点;

[0053] 第二部分:输入数据为司机编号1*1,包括实体嵌入层和三层隐藏层,嵌入层输入维度为NumDriver(NumDriver为司机数目),输出维度为32的数据;

[0054] 第三部分:将第一部分输出的特征向量、第二部分输出的特征向量与其余特征组成的特征向量进行拼接;

[0055] 第四部分:是一个包含512个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为512的数据;

[0056] 第五部分:是一个包含256个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为256的数据;

[0057] 第六部分:是一个包含128个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为128的数据;

[0058] 第七部分:是一个包含64个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为64的数据;

[0059] 第八部分:是一个包含1个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为1的数据,该数据就是使用深度神经网络模型预测出的旅行时间。

[0060] 有益效果:本发明提供的一种基于多模态数据融合与多模型集成的旅行时间预测方法,相对于现有技术,具有以下优点:

[0061] (1)由于当前解决旅行时间预测问题,往往需要来自现实生活的大规模车辆轨迹数据,如果所有车辆行驶过程中的轨迹数据都参与建模过程,训练任务将会有很大的计算代价,本发明从大量的轨迹数据中提取少量的行程数据参与模型计算,将大大减小运算代价,提高了运算速度;

[0062] (2)由于旅行时间问题受到多个复杂因素的影响,不仅需要从时空角度分析问题,

还需要结合天气数据和司机画像等其他数据领域的特征。由于梯度提升决策树在预测短途行程方面具有优势，深度神经网络擅长预测长途或者包含复杂交通状况的行程，本发明提出了一种基于决策树的融合方法，综合了梯度提升决策树和深度神经网络方法两个子模型的优点，有效地提高了模型的预测精度。

附图说明

[0063] 图1为本发明中一种基于多模态数据融合与多模型集成的旅行时间预测方法的流程图；

[0064] 图2为本发明中行程提取算法的流程图；

[0065] 图3为本发明中深度神经网络模型的结构图。

具体实施方式

[0066] 下面结合附图对本发明作更进一步的说明。

[0067] 如图1所示为一种基于多模态数据融合与多模型集成的旅行时间预测方法，主要包括以下步骤：

[0068] a.多模态数据预处理

[0069] 因为本方法忽略了行程过程中的轨迹点，因此首先需要采用行程提取算法对出租车GPS轨迹数据进行预处理，主要包括异常轨迹点的纠正与根据载客状态位提取有效行程两部分；

[0070] b.特征子向量提取与特征融合

[0071] 旅行时间受到多个因素的影响，本方法通过特征向量提取模块从出租车轨迹数据、天气数据、司机画像数据等领域分别提取相应的特征子向量并进行特征拼接；

[0072] c.多模型集成

[0073] 所述模型建立过程包括子模型建立和模型融合两部分：首先根据梯度提升决策树和深度神经网处理回归问题的不同特性，对描述行程的特征进行特定的处理与拼接输入到子模型中，然后使用子模型分别预测旅行时间；然后基于子模型的预测结果建立以决策树为基础的集成模型，最终使用模型预测得到行程的旅行时间。

[0074] 如图2所示，所述行程提取算法具体包括：

[0075] 输入：一辆出租车轨迹序列$T=\{P_1,P_2,P_3,...,P_n\}$

[0076] 输出：行程数据集

[0077] a1.对轨迹序列进行遍历，设置循环变量i从1到n-1，n表示轨迹点的总数，初始时i＝1，行程状态位为0；

[0078] a2.当P_i的载客状态为1，跳转至a3，否则跳转至a4；

[0079] a3.当行程状态位为1时，跳转至a6，否则将P_i记录为行程的起始点，行程状态位置1，跳转至a6；

[0080] a4.当行程状态位为1时，跳转至a5，否则跳转至a6；

[0081] a5.当P_{i+1}载客状态位为1时当前行程记录完毕，将该行程计入行程数据集中，行程状态位置0，否则跳转到a6；

[0082] a6.执行i＝i+1；

[0083] a7.当i＜n时,跳转至a2,否则完成行程数据集的提取。

[0084] 进一步的,所述b.多模态数据分析、特征提取与特征融合具体包括:

[0085] b1.提取空间特征,设lat_1,lng_1表示A点的经纬度,lat_2,lng_2表示B点的经纬度;a＝lat_1-lat_2为两点的纬度之差,b＝lng_1-lng_2为两点的经度之差;r为地球的半径,约为6371km

[0086] 根据半正矢公式计算在地球上两点之间的大圆距离,半正矢公式如下:

[0087] $$d_{Haversine}=2r\arcsin\sqrt{\sin^2\frac{a}{2}+\cos(lat_1)\times\cos(lat_2)\times\sin^2\frac{b}{2}} \quad (1)$$

[0088] 计算曼哈顿距离,曼哈顿距离又称马氏距离(Manhattan distance),计算公式为:

[0089] $d_{Manhattan}$＝|a|+|b| (2)

[0090] 提取方位角,从两地的经纬度,我们不仅可以计算两地的距离信息,还可以得到两地的方向信息,计算公式为

[0091] α＝180°-90°sgn(a)-arctan(b/a) (3)

[0092] 将输入的起始地坐标与目的地坐标使用k-means方法聚类,将聚类后得到的类簇编号使用独热方式编码,最终构成空间特征子向量$trip_s$。

[0093] b2.提取行程时间特征,分别提取行程出发时间的月份、日期等离散型周期性信息以及是否为工作日、节假日等状态信息,最终构成时间特征子向量$trip_t$;

[0094] b3.提取天气特征,根据天气状况将天气划分为不同的等级,如下表所示

[0095]

天气状况	量化等级
晴天、阴天	1
雷阵雨、小雨、雾	2
小到中雨、中雨	3
大雨、大雾、小雪	4
中雪、大雪	5

[0096] 最终构成天气特征子向量$trip_w$

[0097] b4.提取司机特征,获取驾驶出租车的司机编号、性别、年龄、驾龄等信息,最终构成司机特征子向量$trip_d$;

[0098] b5.将行程空间特征子向量、行程时间特征子向量、天气特征子向量与司机画像特征子向量进行拼接,并与该行程对应的旅行时间$trip_{traveltime}$组成描述行程信息的完整特征向量:trip＝(($trip_s$,$trip_t$,$trip_w$,$trip_d$),$trip_{traveltime}$)。

[0099] 进一步的,所述c.多模型集成具体包括:

[0100] c1.建立梯度提升决策树模型

[0101] 梯度提升决策树擅长结合人工挖掘的特征来获得高阶属性或非线性映射,此处设置梯度提升决策树的损失函数为均方误差损失函数。

[0102] 输入:样本集{$trip_j$|trip＝(($trip_s$,$trip_t$,$trip_w$,$trip_d$),$trip_{traveltime}$),j∈(1,all)},其中all为样本的总数目

[0103] 输出:梯度提升决策树模型

[0104] 1.1)初始化第一棵决策树T_0(trip)＝0;

[0105] 1.2)设置参与梯度提升决策树模型训练的决策树的总数目为M,分别对m＝1,2,...,M,在生成第m棵决策树时,计算每个样本的残差r_{mj},其中$r_{mj}=trip_{traveltime_j}-T_{m-1}(trip_j)$,j＝1,2,...,all,通过使用回归决策树模型拟合每个样本的残差r_{mj},学习得到第m棵决策树$T_m(trip;\Theta_m)$,其中Θ_m为第m棵决策树中的参数;

[0106] 1.3)更新$f_M(trip)=\sum_{m=1}^{M}T_m(trip;\Theta_m)$;

[0107] 1.4)得到梯度提升决策树模型$f_M(trip)$,其输出就是使用梯度提升决策树模型预测出的旅行时间;

[0108] c2.建立深度神经网络模型。

[0109] 由于梯度提升决策树算法的局限性,处理高维稀疏特征并不容易,而深度神经网络适用于具有高维特征的场景。

[0110] 1.基于空间网格划分进一步提取空间特征

[0111] 为了细化不同区域的特性,将城市细粒度划分为200*200的网格。遍历数据集中的行程,将行程的起始地与目的地归属到对应的网格中,获得出发地编号和目的地编号;

[0112] 2.建立深度神经网络模型

[0113] 对起始地编号、目的地编号以及司机编号通过实体嵌入层将正整数编号转换为具有固定大小的向量,获得到的实体嵌入向量分别经过多层隐藏层处理后和特征向量提取模块提取出的其余特征进行拼接,通过多层全连接网络进行训练;

[0114] 3.损失函数及优化方法

[0115] 在上述模型构建后,需要训练该模型,其中设置训练样本的批大小为512,选用均方误差损失函数,然后使用激活函数为修正线性单元(Rectified Linear Unit,ReLU),由激活函数完成非线性变换,增强本模型对特征的学习能力,通过Adam优化算法进行参数寻优以最小化损失函数,其中学习率为0.001,衰减项1e-08,动量0.9。迭代次数设置为100,使用5折交叉验证的获得最佳模型g(trip)。

[0116] 其中,如图3所示,所述深度神经网络模型具体包括:

[0117] 第一部分:输入数据为起始地编号和目的地编号,包括实体嵌入层和三层隐藏层,嵌入层输入维度为200(空间特征提取中将城市细粒度划分为的200*200的网格),输出维度为16的数据,隐藏层每层有256个神经元节点;

[0118] 第二部分:输入数据为司机编号1*1,包括实体嵌入层和三层隐藏层,嵌入层输入维度为NumDriver(NumDriver为司机数目),输出维度为32的数据;

[0119] 第三部分:将第一部分输出的特征向量、第二部分输出的特征向量与其余特征组成的特征向量进行拼接;

[0120] 第四部分:是一个包含512个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为512的数据;

[0121] 第五部分:是一个包含256个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为256的数据;

[0122] 第六部分:是一个包含128个神经元节点的隐藏层,经过ReLU激活函数处理,得到维度为128的数据;

[0123] 第七部分:是一个包含64个神经元节点的隐藏层,经过ReLU激活函数处理,得到维

度为64的数据；

[0124] 第八部分：是一个包含1个神经元节点的隐藏层，经过ReLU激活函数处理，得到维度为1的数据，该数据就是使用深度神经网络模型预测出的旅行时间。

[0125] c3.建立基于决策树的集成模型

[0126] 对于训练集中的所有样本$\{trip_j | trip=((trip_s,trip_t,trip_w,trip_d),trip_{traveltime}),j\in(1,all)\}$，将$trip_j$分别输入训练后的梯度提升决策树模型$f_M(trip)$得到预测结果$y_{GBDT}$和深度神经网络模型$g(trip)$得到预测结果$y_{DNN}$，构建出集成模型的新的训练样本集$((y_{GBDT},y_{DNN}),trip_{traveltime})$，在新的样本集所在的输入空间中，使用最小二乘法对切分变量y_{GBDT}与y_{DNN}遍历并计算每个切分点的损失值，选择具有最小损失值的切分变量及切分点，递归地将每个区域划分为两个子区域并决定每个子区域上的输出值，构建回归决策树模型，该模型的输出就是预测的旅行时间。

[0127] 相对于现有技术，在本发明中旅行时间预测的方法融合了轨迹数据、天气数据、司机画像数据等，充分提取与挖掘对旅行时间有影响的因素，融合来自不同模态的信息，使用决策树模型集成了梯度提升决策树模型和深度神经网络模型的预测结果，在计算代价减小的情况下也能获得较高的预测准确率。

[0128] 以上所述仅是本发明的优选实施方式，应当指出：对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

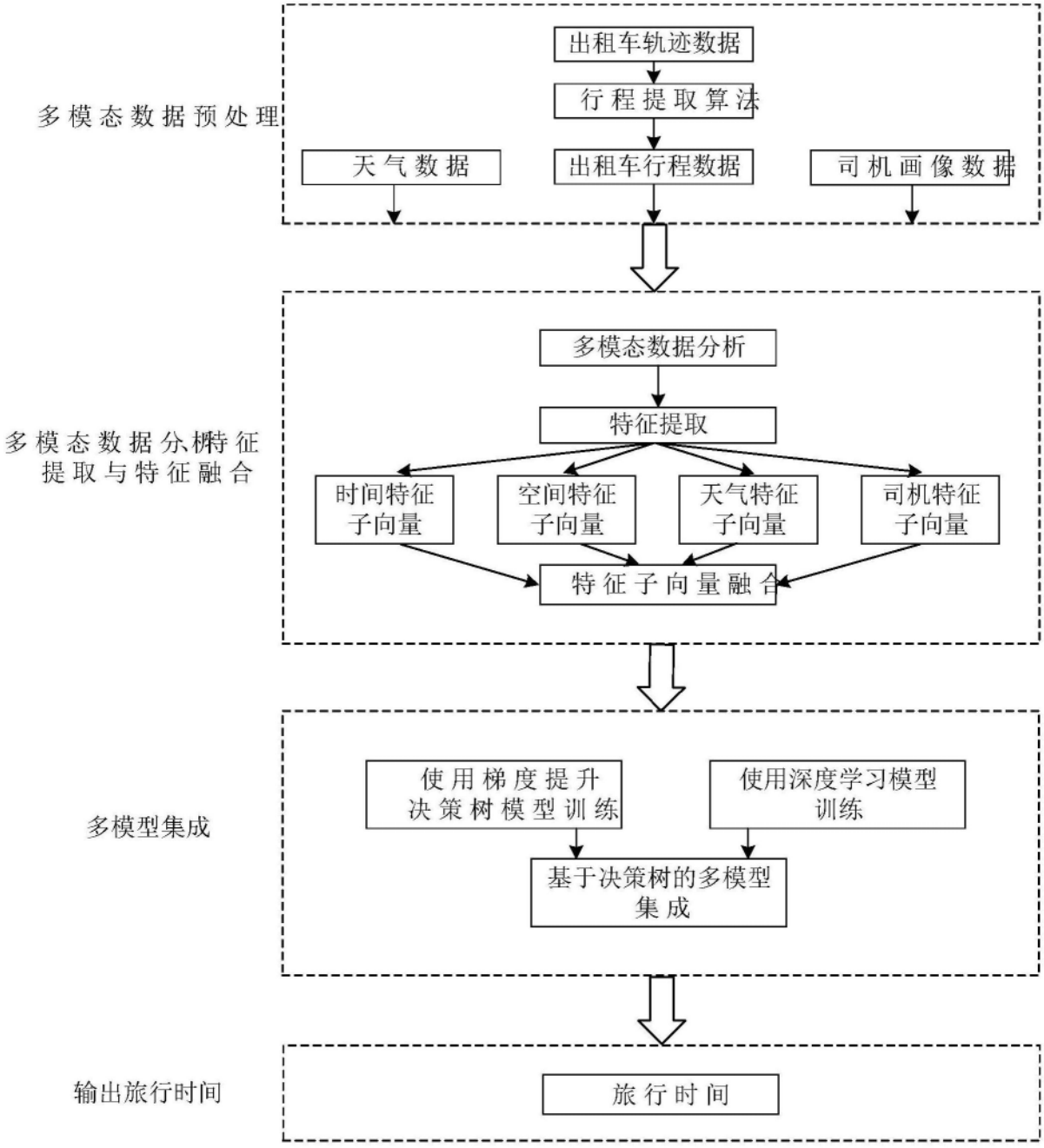

多模态数据预处理
出租车轨迹数据
行程提取算法
天气数据
出租车行程数据
司机画像数据
多模态数据分析特征提取与特征融合
多模态数据分析
特征提取
时间特征子向量
空间特征子向量
天气特征子向量
司机特征子向量
特征子向量融合
多模型集成
使用梯度提升决策树模型训练
使用深度学习模型训练
基于决策树的多模型集成
输出旅行时间
旅行时间

图1

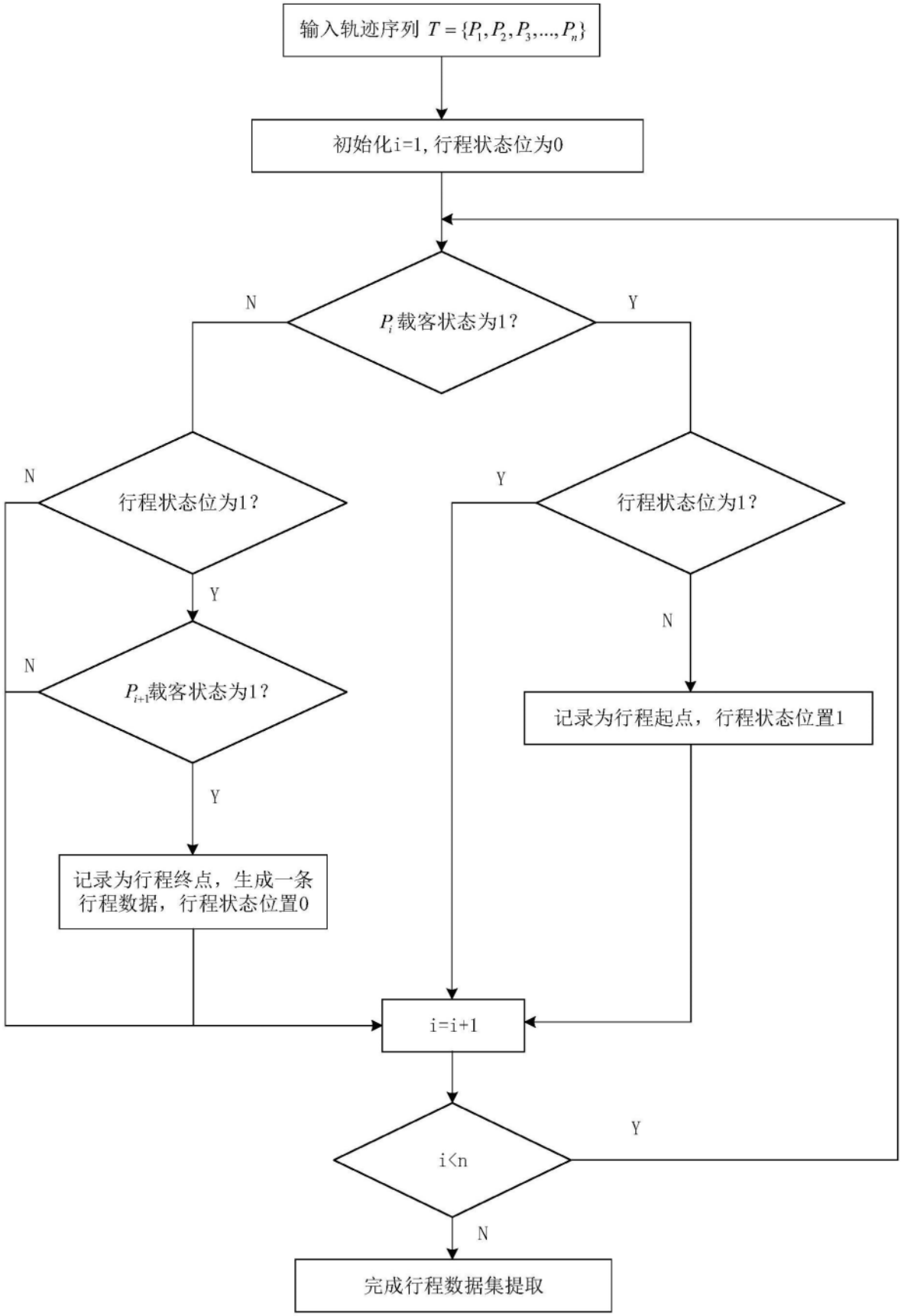

输入轨迹序列 $T=\{P_1,P_2,P_3,...,P_n\}$
初始化i=1,行程状态位为0
P_i载客状态为1?
N
Y
行程状态位为1?
N
Y
P_{i+1}载客状态为1?
N
Y
记录为行程终点，生成一条行程数据，行程状态位置0
行程状态位为1?
Y
N
记录为行程起点，行程状态位置1
i=i+1
i<n
Y
N
完成行程数据集提取

图2

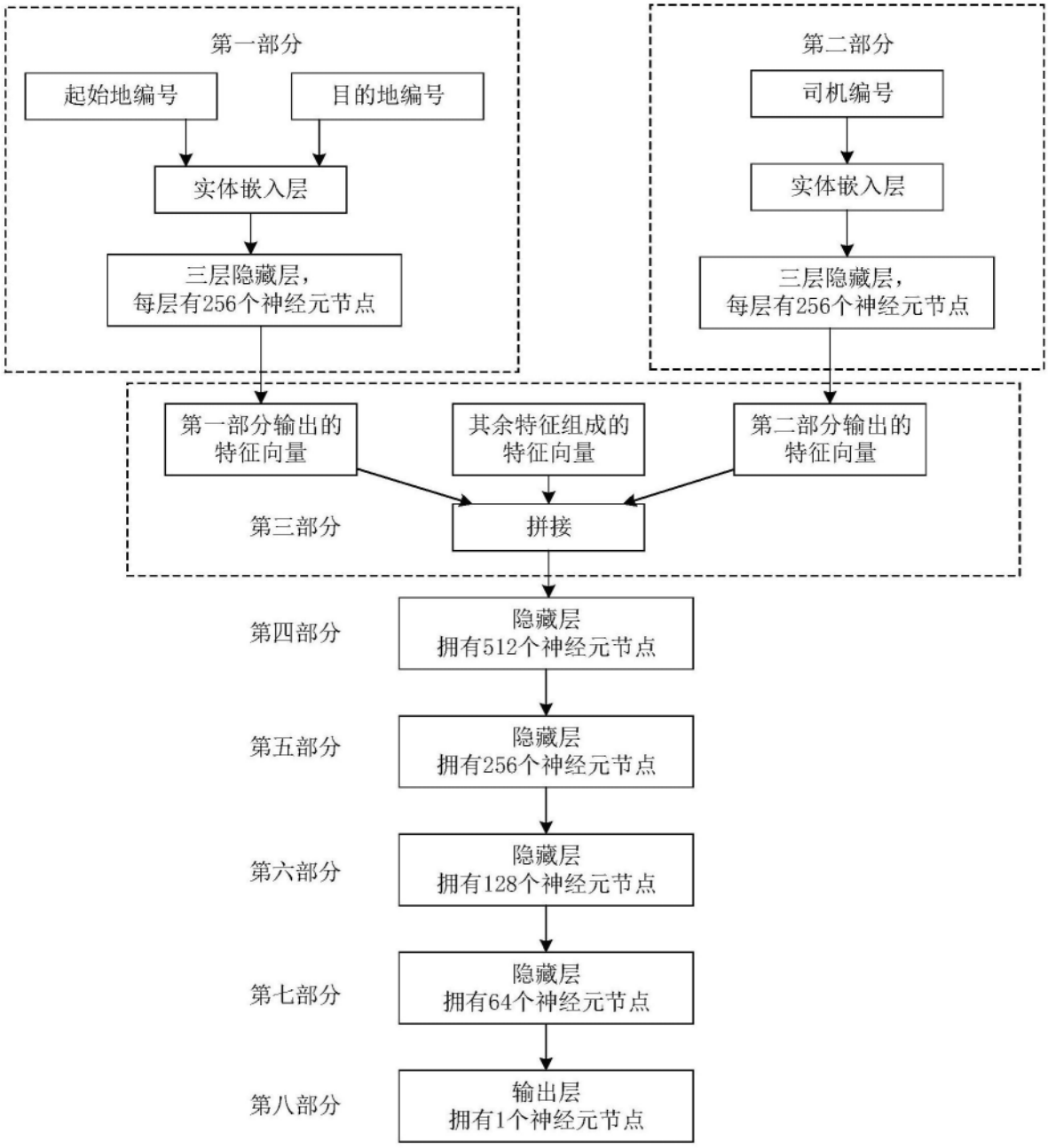

第一部分
起始地编号
目的地编号
实体嵌入层
三层隐藏层，
每层有256个神经元节点
第二部分
司机编号
实体嵌入层
三层隐藏层，
每层有256个神经元节点
第一部分输出的
特征向量
其余特征组成的
特征向量
第二部分输出的
特征向量
第三部分
拼接
第四部分
隐藏层
拥有512个神经元节点
第五部分
隐藏层
拥有256个神经元节点
第六部分
隐藏层
拥有128个神经元节点
第七部分
隐藏层
拥有64个神经元节点
第八部分
输出层
拥有1个神经元节点

图3