一种基于深度学习的城市公交线路查询方法

1. **背景技术**

随着我国汽车产业的逐步发展，城市道路拥堵问题日益严重。通过有效分析目前城市公交线路和站点的相关数据，并在此基础上建立合理的数学分析模型对大规模公交线路数据进行充分挖掘，找出不同线路之间彼此存在的内在关系，从而为乘客提供合理的公交出行路线。传统的公交线路查询算法有迪杰斯特拉(DijKstra)算法及其相应的改进算法、Floyd算法、Moore算法、深度优先搜索算法以及遗传算法等。基于目前的研究现状，诸如此类算法对于大规模的公交线路数据进行有效成分分析及彼此内联关系的分析及时间优化上还存在一定的不足，基于此种现状，本发明提供一种基于机器学习的深度学习算法对大规模公交数据进行学习，找出不同数据链之间彼此存在的关联性。

本发明的深度学习算法提供乘客准确、及时、优化的公交信息服务，生成出行辅助计划，主动而合理的引导乘客出行，不仅方便了乘客，还让缺少秩序的交通出行变得有秩序，使客流分布更加合理。

1. **城市公交相关数据采集**

针对现有城市公交的相关数据，根据目前传统的算法进行线路查询，由于换成次数、出行时间、起始距离以及行程费用等因素的相互制约，因此传统方法难以得到一个较为合理的反馈结果，本发明提供了一种基于深度学习的城市公交线路查询方法。

采集的各类公交数据为后续的深度学习提供丰富的历史数据和当前数据。主要包含以下方面：

1. 采集各城市公交线路数据。如1路、2路公交等；
2. 采集公交各线路的站点数据。如1路公交从甲地始往乙地途经的各个公交站点；
3. 采集各线路的起止出发时间，停运时间，各相邻站点之间行驶需要的时间等数据，形成线路时间数据链条；
4. 采集各线路相邻站点间的距离数据，形成线路距离数据链条；
5. 采集各线路的行程费用数据。

对于公交数据要按照一定规律对其进行分组，形成相应的数据集。以城市为单位进行汇总，最后将形成全国的公交线路数据。将城市数据当成父节点，对每个城市的所有公交路线进行统计，将上述五种数据分别放在不同路线的子数据集中，最后将得到一个较大规模的全国公交路网数据集合，列举两个城市加以说明，如图1所示。

 图1 全国公交线路数据集合示意图

1. **基于深度学习的公交线路查询方法**

当全国公交路网数据集完成挖掘整合后，就要通过数据运算平台进行深度学习和不间断的训练。本发明采用深度学习中的自动编码器(DAE)模型来进行机器学习训练，提取公交线路数据中的有效信息，从而为乘客规划有效的出行路线。本发明提供的公交线路方案具备以下特点：

(1)有直达路线的优先采用直达路线；

(2)在没有直达的情况下优先考虑一次换乘方案，并将所有的一次换乘方案依次按照出行时间长短、距离远近、费用多少等因素进行排序，供乘客进行选择；

(3)当没有一次换乘方案的条件下优先考虑二次换乘方案，并将所有的二次换乘方案依次按照(2)中给定的相关因素进行排序，供乘客进行选择；

(4)同理，当没有二次换乘方案的条件下优先考虑三次换乘方案，并将所有的三次换乘方案也依次按照(2)中给定的相关因素进行排序，供乘客进行选择；

(5)当需要超过3次以上换乘才能达到目的地的情况下，能告知乘客合理选择其它的出行方式；

DAE 的产生和应用免去了人工提取数据特征的巨大工作量，提高了特征提取的效率，降低了原始输入的维数，得到数据的逆向映射特征，展现了从少数类标样本和大量无类标数据中学习输入数据本质特征的强大能力，并将学习到的特征分层表示，为构建深度结构奠定了基础，成为神经网络研究的一个里程碑。

DAE 是一种利用经过无监督逐层贪心预训练和系统性参数优化的多层非线性网络从无类标数据中提取高维复杂输入数据的分层特征，并得到原始数据的分布式特征表示的深度学习神经网络结构。DAE由编码器、解码器和隐含层组成。编码器是输入x到隐含表示h 的映射，表示为:，其中，是非线性激活函数，一般为逻辑函数，其表达式为：



解码器函数将隐含层数据映射回重构y，表示为:



其中，是解码器的激活函数，一般为线性函数或者函数。训练DAE 的过程是在训练样本集D上寻找参数的最小化重构误差，重构误差的表达式为：



其中，L为重构误差函数，一般可以用平方误差函数或交叉熵损失函数，二者分别表示为：





其中，平方误差用于线性，交叉熵损失函数用于。

DAE 的构建主要有2个步骤，第一步是改进原型自动编码器的结构(原型自动编码器如图2所示)，即增加隐含层和神经元的数量，调整隐含层节点的分布并改变权值的分享方式等，构建DAE 的基本框架。第二步是根据不同任务的需要选取合适的代价函数及其优化策略、隐含层品质因数和系统性参数优化时的性能指数等，确定DAE 的训练方案。



图2 原型自动编码器

1. DAE的预训练

预训练的目的是将所有权值链接和偏置限定在一定的参数空间内，防止随机初始化的发生进而降低每个隐含层的品质因数，便于对整个神经网络进行系统性参数优化，该算法的核心是用无监督的方法将DAE的输入层和隐含层全部初始化，然后再用逐层wake-sleep算法将每个隐含层训练为自动关联器，实现输入数据的重构，其基本步骤可总结如下:

（1）逐层构建单层神经元，这样每次都是构造一个单层网络。

（2）当所有层构造完后，每层采用wake-sleep算法进行调优。每次仅调整一层，逐层调整。wake-sleep算法原理如下：

A、wake阶段：认知过程，通过下层的输入特征（Input）和向上的认知（Encoder）权重产生每一层的抽象表示（Code），在通过当前的生成（Decoder）权重产生一个重建信息（Reconstruction），计算输入特征和重建信息残差，使用梯度下降修改层间的下行生成（Decoder）权重。

B、sleep阶段：生成过程，通过上层概念（Code）和向下的生成（Decoder）权重，生成下层的状态，再利用认知（Encoder）权重产生一个抽象景象。利用初始上层概念和新建抽象景象的残差，利用梯度下降修改层间向上的认知（Encoder）权重。以上步骤的主要目的是进行网络参数的初始化。

2、DAE 的微调

DAE的微调是构建DAE 的必要步骤，通常采用BP 算法来完成这一任务 。微调的核心思想是将自动编码器的输入层、输出层和所有隐含层视为一个整体，用有监督学习算法进一步调整经过预训练的神经网络，经过多次迭代后，所有权值及偏置均能被优化。由于最后一个隐含层只能输出原始数据的重构，因而不具有分类识别功能。为了让DAE 具有分类识别的功能，需要在完成微调的神经网络的输出层之后加入SVM分类器，将整个神经网络训练成能完成分层特征提取和数据分类任务的多重感知器。

1. DAE训练大规模公交线路数据具体实现

由于深度神经网络模型庞大，计算量非常大，模型训练时间经常要很长的时间，单个计算机根本在短期无法完成，需要在分布式系统上通过并行计算来加速。本发明的深度神经网络的训练阶段是在一定规模的CPU集群上实现的，从而使神经网络机器学习阶段的耗时大大缩短。本发明系统总体方案如图3所示。



图3 系统总体方案图

深度神经网络的具体实现过程为：

1. 为深度学习模型准备数据集

根据上述的五类公交数据，对其进行统一归一化处理，获得各个城市的公交线路数据记为其中表示第一条线路的相关数据，表示第二条线路的相关数据，其它的依次类推。每条公交线路数据具体包含文中上述的五种数据，如路线数据，中包含的站点数据，相邻站点行程时间及起止出发时间数据，相邻站点行程距离数据，票价数据等。

(2) 采用自动编码器深度模型进行数据训练

自动编码器就是一种尽可能复现输入信号的神经网络。为了实现这种复现，自动编码器就必须捕捉可以代表输入数据的最重要的因素，就像主成分分析方法PCA那样，找到可以代表原信息的主要成分。在本发明中也就是要找到大量公交线路数据的有用成分及彼此之间存在的联系，具体过程如下：

A、给定无标签数据，用非监督学习学习特征

在之前的神经网络中，如图4所示，输入的样本是有标签的，即（input, target），这样根据当前输出和target（label）之间的差去改变前面各层的参数，直到收敛。但现在只有无标签数据，如图5所示。此时就要考虑误差如何得到。



图4 有标签神经网络模型



图5 无标签神经网络模型

本发明采用的自动编码模型如图6所示，将input公交线路数据流输入一个encoder编码器，就会得到一个code，这个code也就是输入的一个表示，为了确定输出code就是input的一个表示，加一个decoder解码器，这时候decoder就会输出一个信息reconstruction，输出的这个信息和一开始的输入数据input是比较接近的，在理想情况下几乎是一样的，我们就有理由相信这个code是输入input的最好诠释。所以，每一层就是通过调整encoder编码器和decoder解码器的参数，使得重构误差最小，此时就得到了输入input数据流的第一个表示了，即编码code。因为是无标签数据，所以误差的来源就是直接重构后与原输入input相比得到。



图6 自动编码模型

B、通过编码器产生特征然后训练下一层如此逐层训练

通过上面步骤得到第一层公交数据流的code，此时重构误差最小所以这个code就是原输入数据的良好表达。第二层和第一层的训练方式没有差别利用相同的原理，我们将第一层输出的code当成第二层的输入input数据流，同样最小化重构误差，就会得到第二层的参数，并且得到第二层输入公交数据流的code，也就是原输入信息的第二个表达了。其他层利用同样的方法进行训练，训练当前层，前面层的参数都是固定的，这时候他们的decoder已经不需要了。

C、有监督微调

经过上面的方法，就可以得到很多层了。每一层都会得到原始输入不同的表达。此时深度自动编码器还不能用来分类数据，因为它还没有学习如何去连结一个输入和一个类。它只是学会了如何去重构或者复现它的输入而已。即它只是学习获得了一个可以良好代表输入公交数据流的特征，这个特征可以最大程度上代表原输入信号。因此为了实现分类，需要在自动编码器最顶的编码层添加一个支持向量机分类器（SVM）然后通过标准的多层神经网络的监督训练方法（梯度下降法）去训练。其基本步骤如下：

1) 随机选取类标数据样本用BP 算法对神经网络进行训练，计算各层的输出;

2) 求出各层的重构误差，并根据误差修正权值和偏置;

3) 根据性能指数判定误差是否满足要求，如果未能满足要求则重复步骤1)和2)，直到整个网络输出满足期望要求。

4) 得出最后深度自动编码器DAE预测模型如图7所示。



图7

(4) 利用训练好的DAE预测模型进行公交线路查询

得到训练好的深度学习DAE预测模型后，就可以进行公交路线的查询。预测模型的输入数据为乘客出行始发站和终点站的公交站点数据，此时的输出结果即为本发明给出的公交出行路线方案。

1. **总结**

本发明使用深度学习自动编码器DAE模型对大规模公交线路数据进行分析，从而完成起止站点间线路的查询任务。完成预训练和微调的DAE具有生物神经系统的基本特征，在一定程度上反映了人脑的若干功能，是对生物系统的成功模拟，实现了无监督学习和监督学习的有机结合，且不同隐含层能学到不同的特征，低阶隐含层为高阶隐含层提供特征模式，便于高阶隐含层进行模式组合。同时，DAE 的拓扑结构完整，具有强大的非线性拟合能力，能发现数据的分布式特征表示，提取数据的潜在分层特征，获取原始输入的“层次型分组”或“部分—整体分解”结构，因而可以更好地表示输入数据，具有大规模并行、分布式处理、自组织和自学习等优点。