重庆师范大学

专业学位硕士研究生论文开题报告

及实施计划

|  |  |
| --- | --- |
| 所在学院 | 计算机信息与科学学院 |
| 专业学位类型 | 工程硕士 |
| 专业领域 | 计算机技术 |
| 研究生 | 李泽冯 |
| 指导老师 | 杨德刚 |
| 入学年月 | 202109 |

重庆师范大学研究生院

2023年 2月

**说 明**

一、开题报告作为学位论文撰写和学位申请的重要材料，应及时完成，不得随意更改。确需更改开题报告，应由学生提出申请，导师签字同意，院学位评定分委员会审查盖章后交由校学位评定委员会办公室审核同意。

二、本表一式二份，A4纸张双面打印，研究生院、学院各一份。

1. **论文开题报告**

|  |  |
| --- | --- |
| **拟定论文题目: 基于图编辑距离的车辆旅游轨迹聚类算法研究与应用** | |
| **选题分类（√）** | **1. 基础研究 2. 应用研究√ 3. 综合研究 4. 其它：** |
| **选题来源 （√）** | **1. 国家基金项目 2. 教育部人文社科项目**  **3. 省部级项目 4. 其它√** |
| **开题报告内容（选题价值、文献综述、研究或开发设计、可行性分析等）**   1. **选题价值**   地理信息系统与科学的不断发展使得时间地理学从数据采集、存储、分析到可视化等整个过程都获得了极大的便利，其强大的可视化和空间分析功能增强了个体时空轨迹的直观显示及其隐含模式的识别能力［1,2］。伴随着神经网络与高性能计算技术的不断发展，交叉学科研究的推进，越来越多的地理信息已经被计算机技术所处理，学科交叉的结果是给地理学科的发展提供了不同的方向。其中轨迹数据隐含了丰富的城市信息，往往规模巨大且来源广泛。轨迹数据主要包括人类活动轨迹数据、交通轨迹数据、动物活动轨迹数据和自然现象轨迹数据[3]。这些轨迹数据的时空尺度与粒度各有不同，分别适用于不同研究。其中人类活动轨迹和交通轨迹数据是与人类日常生活联系最紧密的两类轨迹数据，因此有大量的研究围绕着他们。  人类活动轨迹数据，包括微博定位签到数据、照片数据、手机基站定位数据、信用卡消费数据、手机定位数据等，按照时间顺序跟踪这些位置记录就形成关联人在一段时间内的移动轨迹。这类轨迹数据精度不高，通常在200ｍ左右［4］，而且大多涉及个人隐私，并没有得到广泛的应用，一般在研究中作为辅助数据用于丰富语义信息。 交通轨迹数据[5]包括两类：一类是通过车载GNSS设备采样得到的交通工具移动轨迹，如公交车轨迹数据、出租车轨迹数据、船只轨迹数据等。另一类是通过固定设备采集得到的数据，如城市道路卡口照片、视频监控数据、地铁刷卡数据，高速公路收费站数据等，城市道路卡口照片数据和视频监控数据太复杂，涉及的信息太多不便于有用信息的提取。高速公路收费站的数据，数据信息简单明了，能够清楚的表明车辆在那些地区进行流动，也便于实验数据的处理。  基于图神经网络研究个体或群体的轨迹，发现其活动中蕴含的空间认识规律及空间行为和交互模式，建立以人为本的地理信息服务，进而支持个体或群体时空行为决策，已成为地理信息科学研究的前沿问题。此类研究也将为联系“计量”和“行为”两个地理学思想流派以及重新审视人“上—地”关系提供一个全新的视角，从而推动理论地理学的进一步发展[6]。  基于图编辑距离的车辆旅游轨迹聚类算法通过改进GED（图编辑距离）来衡量两个车辆轨迹之间的相似度，使用图神经网络SIMGNN模型将求两个轨迹之间的相似度问题转换为一个可学习问题，提高了计算两个轨迹之间相似度的效率，得出每两个车辆轨迹之间的相似度得分，再使用DBSCAN、K-Means聚类算法进行聚类，分析聚类结果从而得到车辆运动轨迹的规律。同时我们可以根据聚类结果，挖掘车辆轨迹特征，可以发现车辆的行为规律，例如：节假日车开往有景区的地方和无景区的地方车辆的比较，那些因素导致车辆往一个地方去，景点地区间的车辆的流通形式等等。高速公路出入站口数据可以反映在特定的时间节点（春节、劳动节、国庆节等）上人口的流向和社会生活动态。把此类数据分析好，不仅可以推动地区展示自己的特色、帮助游客更好的规划自己的行程，还可以为旅游监测与分析领域的研究提供工作基础和方法指导[7,8]。   1. **文献综述**   **2.1、图编辑距离**  Sanfeliu和Fu受字符串编辑距离概念的启发，将其扩展到图领域[9]。文献[10]将图编辑距离定义为所有子图同构的最小代价，由此加快了图编辑距离的研究。图可以用来表示对象之间的复杂关系，因此在很多领域有着大量的应用。分析并管理图数据有很多基本问题需要解决，近年来针对这些问题有着大量的研究，其中如何度量两个图的相似性是很多应用的基础。例如，图的分类、图的聚类以及图的相似性查找等等。  近年来，研究者们提出了诸多的图相似性计算方法，在这些方法中，基于图编辑距离（Graph Edit Distance）的方法受到了很多关注。因为其适用于多种类型的图。图编辑距离是指将一个图转变成另一个图所需要的最小操作数，应用在车辆轨迹也是一样的，可以用来衡量两个图的相似程度。图编辑距离问题的求解是一个NP-Hard问题，时间和空间都有很大的开销。因此，图编辑距离算法需要解决复杂度过高带来的开销过大的问题。现有的图编辑距离算法主要是通过搜索顶点之间的映射来进行，搜索过程可以视为扩展搜索树的过程，搜索树的节点即为顶点的映射。现有的方法可以被分为两类：最佳优先搜索和深度优先搜索。最佳优先搜索基于经典的A\*搜索算法（A\*\_GED）,在搜索时优先扩展编辑代价，估计值最小的映射，搜索到的第一个完整映射对应的编辑代价就是图编辑距离。深度优先搜索（DF\_GED）则优先扩展搜索树中层次更深的映射，直到遍历完所有完整映射，取其中最小的编辑代价作为图编辑距离。A\*\_GED会占用大量空间，而DF\_GED在找到局部最优解时会产生大量回溯，消耗大量时间，效率不如A\*\_GED。  计算精确图编辑距离是NP难问题，由于其庞大的解空间，精确图编辑距离算法的时间及空间复杂度很高，减少搜索空间变成了研究热点，为减少对空间的占用和时间的消耗，往往都是采用剪枝策略。陈锌扬[11]等提出利用Symmetry-Breaking方法，通过降低扩展映射数量，开以同时降低时间和空间开销。Bunke 和 Riesen 将二部图概念引入到图编辑距离中,提出一种基于二部图的近似图编辑距离算法(Bipartite Graph Edit Distance, BP-GED)。通过构造编辑距离的二部图代价矩阵框架(BP代价阵框架)，利用代价矩阵求解 GED，将图编辑距离问题转化为二次指派问题[12]。文献[13]从理论上分析比较Hungarian算法[14]Munkres算法[15]和Volgenant-Jonker[16]算法求解BP代价阵框架的性能，时间复杂度均为O(n3)。基于BP代价矩阵框架的近似图编辑距离算法可以在多项式时间求解,但随着图的规模增大，会出现存储消耗高和精度不足的问题。文献[17]采用六种不同的搜索策略以提高BP-GED算法的精度。Fischer等人针对不同的局部结构之间的不同度量标准,提出了八个不同的中心度度量，其中特征向量和 PageRank对稀疏图的精度提升效果最佳,但是所有这些方法都是以增加运行时间作为代价[18]。[19]中提出一种环结构，利用更多的结构信息构建代价阵，从而获得更高的精度。文献[20]利用指派问题中矩阵的性质，在满足度量公理下，构造规模更小代价矩阵，时间和空间都得到提升，但在图的规模变大时，精度会降低。此外，Dwivedi提出的一种通过利用中心性度量来忽略最小中心顶点的算法框架，并采用不同的中心度标准删除不重要的顶点来达到降低求解规模，以牺牲精度为代价大大减少总体匹配时间，应用在大图上时准确性会严重降低[21,22]。（图编辑距离求解技术研究）  **2.2、轨迹聚类算法**  近年来，聚类成为模式挖掘中的热点之一。聚类根据数据集中数据间的相似度，按定的结构规则进行分组，即将数据集中相似的数据对象聚在一起，形成一个类别。因此，类内的对象相似度较高，类间的对象相似度较低。  现有的一些聚类方法是基于划分的方法。该类方法大部分基于距离，将一个含有n个对象的集合划分为k组，每个组内包含一个以上对象。K-meanst[23]为该类方法中具有代表性的算法，K-means算法简单，但需要手动选取聚类个数以及初始质心。此外，k中心点算法（K-medoids）[24]、CLARANS[25]也较为常用。  基于层次的聚类方法通过数据间相似度，从大到小或从小到大进行排序,逐步连接各个节点。2017年，Zhao 等人[26]提出了一种基于 DBSCAN的自适应分层聚类方法。文中对船舶运动轨迹进行分层分析，在每一层中选择合适的参数，从非均匀密度数据集中发现同类的船舶轨迹，以达到聚类的目的。  基于模型的聚类方法。此类方法即使用神经网络模型或概率模型进行聚类的方法。此类方法划分出来的同一类数据，具有同一种概率类型。算法中簇间没有严格的边界，但数据量少分布数量多时，算法效率不高。2014年，文献[27]提出了一种基于隐马尔可夫模型(Hyper Media Management Schema, HMMs)的方法。此方法将每条轨迹映射到一个HMM （Hidden Markov Model）中，然后定义HMMs之间的距离，最后使用基于距离矩阵的方法对 HMMs进行聚类。2017年，Chamroukhi等人[28]提出使用基于隐马尔可夫模型对时间序列进行回归，再利用极大似然法确定每条轨迹所属的簇,从而实现轨迹聚类。  轨迹聚类算法的核心是将轨迹划分成不同的，具有相似运动规律的轨迹组成的子集。国内外学者对轨道聚类进行了一系列的研究[29]。吐尔逊[30]等采用DBSCAN（Density-Based Spatial Clustering of Applications with Noise）算法对模拟农业机械作为轨迹进行分析，分析农机作业班次的有效作业轨迹、空间转移轨迹和停歇轨迹，得出农机利用率。周培培[31]等提出了基于增强聚类的异常轨迹检测算法，基于速度的最小描述长度准则把轨迹简化成有序线段，再利用DBSCAN算法把线段分成不同的类，从而监测时空异常轨迹。陈锦阳[32]等利用特征点概念将轨迹分成轨迹子段，提出一种改进的轨迹子段距离度量方法，计算轨迹子段之间的相似度，再用CTIHD聚类算法进行轨迹聚类。曹妍妍[33]等针对不完整的运动轨迹，导致分析结果不准确的问题，提出利用改进的Hausdorff距离进行轨迹相似度度量，然后采用谱聚类方法对距离矩阵进行聚类，从而得到符合实际的聚类结果。  根据上方所述，我们可以了解到已经有研究人员在研究不同领域中的轨迹问题，并通过实验研究发现了其中的一些规律指导了实践。  **2.3、图神经网络**  近年来，深度学习方法在许多机器学习任务得到了广泛应用，例如计算机视觉(Computer Vision, CV)领域的图像分类、NLP领域的自然语言理解等，这些任务中的数据通常在欧几里德空间中表示，然而随着新的应用不断涌现，由非欧几里德空间生成的数据也逐步增多，通常表示为具有对象节点和复杂依赖关系的图数据。例如，在电子商务中，基于GNN 的学习系统可以通过构建用户-产品关系图并利用GNN提取信息从而给用户提供更加高效的产品推荐信息;在引文网络中，通过利用GNN学习引文之间的相互链接构建的相互引用的关系信息,可以对论文进行分类和聚类分析。  图是一种为一组对象（节点）及其关系（边）建模的不规则的数据结构，由于图中节点之间通过复杂类型的边相互联系，使得图数据节点之间不再相互独立，因此一些重要的计算方法在图数据中很难应用，图结构数据的复杂性为经典的机器学习算法带来了较大的困难和挑战[34]。  图分析是一种独特的用于处理图数据结构的机器学习方法，关注于图中节点的分类和聚类、边的预测等。GNN是一种基于图数据的深度学习方法，随着其不断发展，近年来在图分析中得到广泛应用，GNN的由来主要有两个动机。一个动机是卷积神经网络CNN [35]，CNN作为深度学习模型的典型代表,随着研究学者对 CNN 不断的深入研究发现了CNN 主要包括三个关键特性:一是利用局部连接提取特征，二是利用权重共享降低模型参数量,三是利用残差机制引入多层架构,进一步捕捉深层次的特征。然而，CNN只能对二维结构的图像和一维序列的文本等欧几里德数据进行操作，由于图是局部连通的并且这些数据结构可以被视为图的实例，因此，一些学者想将CNN推广到图中，但对于图中卷积核的定义较为困难，故并未成功。另一个动机是图嵌入[36],即模型利用学习的低维嵌入向量表示节点或者边。随着单词嵌入的成功发展，Deepwalk[37]遵循表征学习的思想，将SkipGram模型应用于图中随机游动生成的序列上并生成节点的嵌入表示，随后，Node2vec[38]等图嵌入方法不断涌现。然而，这些方法存在以下缺陷，首先,模型没有共享参数，使得计算效率低下，其次，泛化能力较弱，对于一些新的且较为复杂的图难以处理。在此基础上，图神经网络 GNN应运而生。  Sperduti等人[39]在1997年首次将神经网络应用于有向无环图，激发了学者对GNN的早期研究。图神经网络的概念最初在2005年 Gori等人[40]的论文中概述，在2009年Scarselli [41]等人的论文中进一步阐述，这些早期研究基于不动点理论学习目标节点的表示，属于递归图神经网络的范畴，但这一过程的计算成本很高。一些图数据的卷积方法不断发展，比如基于卷积神经网络CNN的卷积图神经网络(ConvGNN)。ConvGNN分为两大主流，基于频谱的方法和基于空间的方法。Bruna等人[42]首先提出了一种基于频域的图卷积神经网络，通过傅里叶变换在频域中实现图中节点的卷积操作计算。然而,该方法具有较高的计算复杂度。其次，ChebyNetst[43]使用切比雪夫多项式去除了计算代价高昂的拉普拉斯特征分解过程，进一步降低了计算复杂度。然后，GCN[44] (GraphConvolutionalNetwork）只考虑了一阶切比雪夫多项式滤波器，通过应用多通道卷积，重新定义了传播矩阵S，进一步简化了图卷积，降低了时间复杂度，此外，GCN可以堆叠多层,以增加图中节点的感受野。基于空间的ConvGNN的研究比基于频谱的ConvGNN早得多,起初，Micheli等人[45]通过复合非递归层首次解决了图的相互依赖性问题。随着进一步发展，过去几年中还诞生了不同类型的GNN，包括图自动编码器（Graph Autoencoders,GAE）[46]和时空图形神经网络(Spatial-temporal Graph Neural Networks ,STGNN)[46]。  GNN可以有效地处理具有丰富关系结构的任务，并能保存图的全局结构信息，在有监督、半监督、无监督等问题领域中得到了广泛应用，例如物理系统、知识图谱、文本、图像等。   1. **项目的研究内容、研究目标，以及拟解决的关键科学问题**   **1、 研究内容**  本研究是基于图编辑距离的聚类算法及其应用，研究内容主要分为如下几个方面：  1.1 数据收集与预处理  本次研究的数据集是在2012、2014、2016三年国庆七天湖北、四川、广东、贵州四地的车在重庆高速的通行情况。由于只统计了车辆在高速路上的出发点和目的地，没有记录车辆偶尔选择其它道路出行的情况，导致有很多车的轨迹并不完整，中间偶尔会出现断连的情况，这也是符合现实的。对这类数据，我们有选择的将其完整完善，使其成为一个完整的轨迹。  模拟数据的制作，生成N个顶点，连接每两个顶点之间形成边，随机选取M（0<M<=N）个边，其中边代表车辆行驶的轨迹，顶点就代表着车辆经过的地方。  使用真实数据和模拟数据来丰富我们的数据库，更好的训练神经网络模型，使得以该模型的结果作为聚类依据的聚类结果和以真实GED的聚类结果相似。  1.2 图编辑距离的改进  图编辑距离方法是一种高效的图匹配方式，其是同过衡量两个图的相似性来衡量匹配的准确性。其核心思想通过图编辑操作将一个图转换通过一系列的图编辑操作为另一个图。图编辑操作包括插入或删除顶点或边，或更改顶点或边标签（称为重新标记）。    其中默认为每一次图编辑操作都是等价的，但是这种方案并不适合我们的所使用的数据集，我们所使用的数据集中的每一个子轨迹(边)和每一个地区（点）都有着自己的权重和比值，并不能简简单单的统计从一个图转到另一个图的图编辑操作数。因此我们对GED进行改进，在每进行一步图编辑操作的时候，给对应的操作赋予权值，最后相加起来的结果称为图编辑权重距离GEWD(Graph Edit Weight Distance)。  由于GEWD是基于GED改进的，但是并没有对其寻找最小的图编辑操作数的解法进行优化，因此GEWD也存在空间占用大和时间消耗多的问题。故我们在基于图编辑权重距离的车辆旅行轨迹聚类算法中添加图神经网络来解决计算GEWD的空间和时间复杂度高的问题（详情见1.3）。不过对GED算法的修改，使得GED的应用领域更广。使得GEWD可以作为衡量车辆旅游轨迹之间相似性，我们发现GEWD值越小的代表他们相互转换的操作权重就越小，也就是越相似，反之亦然。  1.3 图神经网络模型的改进  自图神经网络提出以来，该技术已经被广泛的应用在生物学化合物聚类、程序BUG检测等各个领域，不断有研究者将图神经网络技术的应用领域扩宽。其中Yunsheng Bai等人将图神经网络用来计算两个图数据之间的相似度得分，用来衡量两个图数据之间的相似程度。将图形相似度计算转换为一个学习问题来解决传统计算GED时间复杂度高的问题，并提出了一种基于神经网络的方法SIMGNN来得到两个图之间的相似度得分。设计了一个可学习的嵌入函数，将每个图映射到一个嵌入向量，它提供了一个图的全局摘要。提出了一种新的注意机制来强调重要节点与特定的相似度度量。其次，我们设计了一种成对节点比较方法，用细粒度节点级信息补充图级嵌入。我们的模型在未见图上实现了更好的泛化，并且在最坏的情况下，相对于两个图中的节点数量，在二次时间内运行。该模型不是使用组合搜索直接计算两个图数据的相似度，而是将其转化为一个学习问题，从另一个方向解决了图编辑距离计算时空间占用大和时间消耗大的问题。在训练阶段，通过最小化预测相似度分数和ground truth之间的差值来学习该函数中涉及的参数，其中每个训练数据点是一对图及其真实相似度分数。在测试阶段，通过向学习的函数输入任意一对图，我们可以获得预测的相似度分数。其模型的结构如图所示。    在使用模型SIMGNN计算两个图数据之间的相似度得分的时候，由于只输入了图数据的结构，并没有将其它详细的数据输入到模型中训练，导致模型得到的相似度得分与真实的GED还是有一定的差别。  本文拟提出丰富输入，修改注意力机制以增加模型对图数据结构特征的提取，使得模型计算出来的相似度得分与真实的GED的差距更小。然后可以将该相似度得分作为聚类的依据，在一定程度上提高了聚类的准确率。  1.4 聚类算法的改进  对于车辆旅行轨迹的聚类属于无监督问题，聚类分析能够发现数据集自身隐含的内蕴结构信息，最终目的是将相似程度大的轨迹划分为一类。DBSCAN算法是基于密度算法的经典代表，算法能够发现任意形状的簇，且不易受噪音以及离群点的影响。因此DBSCAN算法是轨迹聚类常用的一种算法。其中（聚类密度）与（聚类半径为DBSCAN算法的两个参数，其主要思想为：   1. 首先将数据集中所有反射样本标记为Unvisited； 2. 任取一点，，若到其距离内得其余样本点不少于个，则称为核心点。将以及距离内的所有样本点标记为Visited，并隶属于子集, *。* 3. 在剩余的Unvisited样本中任取一点，重复步骤2，继续判断该店是否为核心点。若该核心点距离内的样本已在隶属于先前形成子集的情况，则将该核心点以及它的距离内所有样本点包括进，否则，将这些样本隶属于新的子集。 4. 遍历所有的剩余Unvisited样本，获得中所有聚类子集，若不属于任何一个子集，则将其视为噪音点。   轨迹聚类的难点是如何定义轨迹在空间维度上的相似性度量，基于传统的DBSCAN算法中是用样本间的欧式距离来进行相似度量，本研究提出使用图编辑权重距离来代替欧式距离进行相似度量，使得DBSCAN算法的适用范围更广，使用 DBSCAN方法以实现轨迹聚类。  **2、 研究目标**  针对本研究现状和所存在的问题，本文提出基于改进的图编辑距离车辆旅游轨迹聚类算法，使图神经网络与聚类算法联合，并将该算法应用在车辆旅行轨迹中，分析轨迹所存在的相关关系。   * + 1. 本课题在图编辑距离的基础之上，提出图编辑权重距离，使其更适合度量两个车辆旅游轨迹的相似性。     2. 本课题将从数据集的搜集与处理出发，使用基于图编辑权重距离的数据集训练图神经网络模型用于计算两个轨迹之间的相似性得分，使其图神经网络模型得出来的相似性得分更加贴近两轨迹之间的真实距离。     3. 结合聚类算法，将图神经网络模型得出来的结果应用在聚类中，建立一套基于图神经网络计算图形相似性得分的聚类算法。对车辆旅游轨迹进行聚类，使得属于同一个簇的车辆轨迹之间的相似度大，而不同簇间的车辆轨迹相似度小。     4. 本课题在聚类步骤中，需要分别根据真实的GED和模型得到的相似度得分两个准则聚类，并比较两者结果的相似程度，从而分析同一簇中的轨迹规律。   **3、 拟解决的关键问题**   1. 现有的GED并不适合衡量我们的数据集中样本间的相似性，我们需要重新构建一个新的度量方式来衡量两个轨迹之间的相似性。 2. 缩小神经网络模型SIMGNN所计算出来的相似性得分与真实值之间的差异，使得该相似度得分能够被聚类所使用。 3. 聚类算法的选取，选择或改进后的聚类算法应该适用于本研究中的数据集，能够很好的将车辆旅游轨迹聚成不同的簇。 4. **拟采取的研究方案及可行性分析** 5. **技术路线**   本课题研究的技术路线有数据集的制作与预处理、图编辑距离的改进、图神经网络模型的优化和聚类算法的使用四部分组成。此4部分的内容贯穿整个车辆旅游轨迹的聚类过程，以下对此4个部分进行简单的路线介绍。  本课题通过获取高速收费站的出站与入站车辆的信息进行采集，转换成车辆旅游轨迹，使用python语言对数据进行预处理，使得采集的数据可以供模型和聚类算法所使用。并且添加自制数据集，完善数据种类，增加数据量，对模型训练有正向作用。  本课题算法的改进部分主要是针对车辆轨迹之间的相似性度量，在图编辑距离的基础之上，提出图编辑权重距离，对每一步图编辑操作赋予权重，使其更加适用于车辆旅游轨迹之间的相似性度量。  本课题通过模型训练来得到车辆旅游轨迹之间的相似度得分，并根据模型所得到的得分与真实值进行比较，从而反向调整模型参数，使得最终模型所得的相似度得分能够聚类算法使用。  本课题通过需求分析，针对车辆旅行轨迹聚类总体设计， python是主要开发语言，并配合PyCharm Community Edition集成开发环境进行实验代码的编写工作。     1. **研究方法** 2. 文献综述法   大量收集国内外目标检测算法和目标跟踪算法的期刊和会议论文，了解所使用的，技术路线，以及每个技术的优缺点，并以此进行相关理论分析。最终能够从理论上寻找可行性技术路线。   1. 实验法   大量阅读论文后，筛选一些与本研究相关的论文，并对其思想和代码进行了解与实践，尝试着发现其中的不足点，并寻找改进方案。结合各个论文中优秀的思路，与此同时也可以将自己的一些创新性想法加入其中以达到改进的目的。图神经网络模型的加入主要是解决GED算法的时间和空间复杂度高的问题，并且GED算法对于超过16个节点以上的图数据，是不能计算它们之间的一个距离。可以增加算法覆盖不全的问题。   1. 对比分析法   本课题会将GED的相似度得分和图神经网络相似度得分分别作为聚类条件，对比两者的聚类结果，看图神经网络的引入是否回有利于聚类。并且也会与其它轨迹得度量方式和聚类算法作对比，发现各自的优缺点，针对本课题数据集的算法，选其优点，并改正缺点，最终使得该算法适合车辆旅游轨迹的聚类。   1. **实验方案**   根据本课题得研究内容、研究方法和技术路线，对各个研究部分拟定以下研究方案：   * 1. 图神经网络模型评价指标   2. 时间评价指标：主要是用来评价模型的时间开销，用于收集每个模型计算一对图的相似度得分所需的必要时间。   3. MSE：均方误差测量模型计算的相似度和真实相似度之间的平均平方差。其公式为：     其中为权重，为真实相似度，为模型计算的相似度。  3.2 聚类结果的评价指标  根据真实的GED和模型SIMGNN的相似度得分分别作聚类，观察结果，以两种不同方式得到的相似度得分的聚类结果是否一致或者存在某种联系。  3.3 改进的GED评价指标  改进后的GED是否能比较准确的将两个轨迹之间的相似度衡量出来，可以观察根据真实GED聚类过后的结果，是否存在一定的规律或其它隐藏关系。   1. **关键技术**   本课题研究主要使用了以下关键技术：   1. 图编辑权重距离   本实验采用图编辑权重距离来衡量两个轨迹的相似度，是在图编辑距离上的改进，使得该距离能够应用在本课题的数据集中。   1. SIMGNN   传统计算GED的时间和空间复杂度太高，且有一定的局限性。本课题采用图神经网络模型SIMGNN用以提高计算两个轨迹的相似度的效率，也可以计算节点超过16个的轨迹之间的相似度。   1. 聚类算法   提出使用图编辑权重距离来代替传统DBSCAN算法中的欧式距离进行相似性度量，改进DBSCAN方法以实现轨迹聚类。  **5. 可行性分析**  5.1 研究问题的可行性  本课题所研究的车辆轨迹聚类问题一直都在被研究，且国内外对车辆轨迹聚类算法具有一定的基础，我是在站在前人的肩膀上重新审视车辆轨迹聚类问题。本课题基于前人的研究结果对算法进行改进和创新，并对算法的基本理论作以解释，所以本课题所研究的问题具有可行性。  5.2 研究方案的可行性  本课题的研究方案是对研究内容进行推导，一步一步的构成整个研究，在层次和结构上循序渐进，不存在跳跃研究，并且每个研究步骤都可达，不存在不能完成的情况，所以本课题的研究方案具有可行性。  5.3 研究技术的可行性  本课题使用深度学习作为研究框架，使用图神经网络模型SIMGNN算法和DBSCAN聚类作为基本模型对车辆旅行轨迹聚类问题进行研究，研究的主要技术都是整个领域的基础，研究的创新和改进部分也是基于这些基础在思想上进行创新，在技术上进行改进，所以本课题的研究技术具有可行性。  5.4 研究环境的可行性  本课题研究者所在实验室拥有一台GPU计算设备并有充足的时间进行课题的研究，保障了本课题实验研究的进行，所以本课题具有研究环境的可行性。   1. **预期研究成果与研究思路** 2. 提出图编辑权重距离（GEWD），推广GED的使用范围，使之可以衡量车辆旅行轨迹之间的相似性，为后续的聚类和图神经网络训练提供良好的数据支撑。 3. 将改进SIMGNN图神经网络模型，减小模型结果与真实结果之间的差异，提高模型预测的准确率。 4. 将图神经网络模型和聚类算法框架相结合，减少计算两个轨迹之间相似性的时间复杂度，建立比较健壮的基于图编辑距离的车辆旅行轨迹聚类算法。   **参考文献：**  [1] Kraak M J, Koussoulakou A. A Visualization Environment for the Space-Time-Cube[C]. Developments in Spatial Data Handling. Berlin Heideberg:Springer,2005:189-200  [2] Ren F, Kwan M P. Geovisualization of Human Hybrid Activity-Travel Patterns[J]. Transactions in GIS,2007,11(5):721-744  [3] 高强，张凤荔，王瑞锦等轨迹大数据：数据处理关键技术综合研究综述[J]. 软件学报，2017，28（4）：959-992.  [4] PAN Gang, QI Guande, ZHANG Wangsheng, et al. Trace analysis and mining for smart cities: issues, methods, and applications[J]. IEEE Communications Magazine, 2013,51(6): 120-126  [5] 吴华意，黄蕊，游兰等出租车轨迹数据挖掘进展[J], 测绘学报. 2019,48(11)  [6] 刘瑜，康朝贵，王法辉等大数据驱动的人类移动模式和模型研究[J], 武汉大学学报(信息科学版). 2014,39(06)  [7] Ahas R, Aasa A, MarkD, et al. Seasonal Tourism Spaces in Estonia: Case Study with Mobile Positioning Data[J]. Tourism Management, 2007, 28(3): 898-910  [8] Tiru M, Kuusik A, Lamp M L, et al. LBS in Marketing and Tourism Management: Measuring Destination Loyalty with Mobile Positioning Data[J].Journal of Location Based Services, 2010,4(2):120-140  [9] Sanfeliu A, Fu K S. A distance measure between attributed relational graphs for pattern recognition[J]. IEEE Transactions on Systems Man & Cybernetics, 2012, SMC-13(3):353-362.  [10] Messmer B T, Bunke H. A new algorithm for error-tolerant subgraph isomorphism detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(5): 493-504.  [11] 陈锌扬，王璿，周军锋等一种高效的图编辑距离计算方法[J]. 智能计算机与应用，2020,10(12),94-98  [12] Riesen K, Bunke H. Approximate graph edit distance computation by means of bipartite graph matching[J]. Image & Vision Computing, 2009,27(7):950-9597.  [13] Fankhauser S, Riesen K, Bunke H. Speeding Up Graph Edit Distance Computation through Fast Bipartite Matching[C] Graph-based Representations in Pattern Recognition-iapr-tc-15 International Workshop. DBLP, 2011.  [14] Kuhn H W. The Hungarian Method for the assignment problem[J]. Naval Research Logistics, 2005,52(1): pp.7-21.  [15] Munkres J. Algorithms for the Assignment and Transportation Problems[J]. Journal of the  Society for Industrial & Applied Mathematics, 1957,5(1):32-38.  [16] Jonker R, Volgenant A. A shortest augmenting path algorithm for dense and sparse linear  assignment problems[J]. Computing, 1987,38(4): 325-340.  [17] Riesen K, Bunke H. Improving bipartite graph edit distance approximation using various search strategies[J]. Pattern Recognition: The Journal of the Pattern Recognition Society，2015,48(4):1349-1363.  [18] Riesen K, Fischer A, Bunke H, Improving graph edit distance approximation by centrality  measures[C] 22nd International Conference on Pattern Recognition, Stockholm. Sweden, 2014: 3910-3914.  [19] Blumenthal D B, Bougleux S, Gamper J, Brun L. Ring Based Approximation of Graph Edit  Distance[C] Structural，Syntactic，and Statistical Pattern Recognition. S+SSPR 2018. Lecture Notes in Computer Science, vol 11004: 293-303.  [20] Serratosa, Francesc. Fast computation of Bipartite graph matching[J]. Pattern Recognition Letters, 2014,45: 244-250.  [21] Prakash D S, Shankar S R. Error-Tolerant Graph Matching using Node Contraction[J]. Pattern Recognition Letters, 2018, 116(DEC.1):58-64.  [22] Prakash D S, Shankar S R Error-tolerant approximate graph matching utilizing node centrality information[J]. Pattern Recognition Letters, 2020,133(0):313-319.  [23] Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations[C]. Proc of Berkeley Symposium on Mathematical Statistics & Probability.1965: 281-297.  [24] Estivill-Castro V. Yang J. Fast and robust general purpose clustering algorithms[C]. In  Proceedings of Pacific Rim International Conference on Artificial Intelligence.2000:208-218.  [25] NG, R. T. A Method for Clustering Objects for Spatial Data Mining[J]. IEEE transactions on knowledge & data engineering, 2001, 14(5):1003-1016.  [26] Zhao L, Shi G, Yang J. An adaptive hierarchical clustering method for ship trajectory databased on DBSCAN algorithm[C]. IEEE International Conference on Big Data Analysis. IEEE, 2017.[27] S. Ghassempour, F. Girosi, A. Maeder. clustering multivariate time series using hidden markov models[J]. International Journal of Environmental Research & Public Health, 2014, V11(3): 2741-2763.[28] Chamroukhi F, Same A, Aknin P, et al. Model-based clustering with Hidden Markov Model regression for time series with regime changes[C]. Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011.  [29]江玉玲，熊振南，唐基宏．基于轨迹段DBSCAN的船舶轨迹聚类算法[J]．中国航海，2019(3).  [30]吐尔逊·买买提，谢建华．基于DBSCAN的农机作业轨迹聚类研究[J]．农机化研究，2017,39(4):7-11.  [31]周培培，丁庆海，罗海波，等．基于DBSCAN聚类算法的异常轨迹检测[J]．红外与激光工程2017,46(5):230-237.  [32]陈锦阳，宋加涛，刘良旭，等．基于改进Hausdorff距离的轨迹聚类算法[J].计算机工程，2012,38(17):157-161.  [33]曹妍妍，崔志明，吴健，等，一种改进Hausdorff距离和谱聚类的车辆轨迹模式学习方法[J]，计算机应用与软件，2012,29(5);38-40.  [34]王健宗，孔令炜,黄章成等.图神经网络综述[J].计算机工程, 2021.  [35] Y. Lecun, L. Bottou. Gradient-based learning applied to document recognition[J].  Proceedings of the IEEE,1998,86(11):2278-2324.  [36] C. Peng, X. Wang, J. Pei, et al. A Survey on Network Embedding[J]. IEEE Transactions  on Knowledge and Data Engineering, 2017, PP (99):1-1.  [37] B. Perozzi，R. Al-Rfou, S. Skiena. DeepWalk: Online Learning of Social  Representations[J].  [38] A. Grover, J. Leskovec. node2vec: Scalable Feature Learning for Networks[C].  Knowledge Discovery and Data Mining, 2016:855-864.  [39] A. Sperduti, A. Starita. Supervised neural networks for the classification of structures[J].  IEEE Transactions on Neural Networks, 1997,8(3): 714.  [40] M. Gori, G. Monfardini, F. Scarselli. A new model for learning in graph domains[C].  IEEE International Joint Conference on Neural Networks.2005.  [41] F. Scarselli, M. Gori, A. C. Tsoi, et al. The Graph Neural Network Model[J]. IEEE transactions on neural networks, 2009,20(1): 61-80.  [42] J. Bruna, W. Zaremba, A. Szlam, et al. Spectral Networks and Locally Connected Networks on Graphs[C]. International Conference on Learning Representations, 2014.  [43] M. Defferrard, X. Bresson, P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering[C]. Neural Information Processing Systems, 2016:3844-3852.  [44] T. N. Kipf, M. Welling. Semi-Supervised Classification with Graph Convolutional Networks[C]. International Conference on Learning Representations, 2016.  [45] A. Micheli. Neural Network for Graphs: A Contextual Constructive Approach[J]. IEEE Transactions on Neural Networks, 2009,20(3): 498-511.  [46] Z. Wu, S. Pan, F. Chen, et al. A Comprehensive Survey on Graph Neural Networks[J].  IEEE transactions on neural networks, 2021,32(1): 4-24. | |

**二、论文工作实施计划**

|  |  |  |
| --- | --- | --- |
| **论文工作的进度与安排** | | |
| **起止时间** | **进度安排** | **备注** |
| 2022.12-2023.2 | 确定论文的方向和资料收集 |  |
| 2023.3-2022.4 | 完成开题报告和论文提纲 |  |
| 2023.5-2023-9 | 完成论文初稿 |  |
| 2023.10-2023.11 | 论文修改 |  |
| 2023.12-2024.2 | 完成论文 |  |
| 2024.3-2024.4 | 论文送审并准备毕业答辩 |  |

**三、指导教师综合意见**

|  |
| --- |
| **对报告人论文的选题价值、文献综述、研究或开发设计、可行性等的综合意见**  该生对选题与老师进行了充分讨论，符合计算机技术专业毕业论文要求，参考了许多文献，论文选题具有一定的理论意义和较的实用价值，主要基于图编辑距离的聚类算法及其应用。本选题是学生所学专业知识的延续，符合学生专业发展方向，对于提高学生的基本知识和技能，对于提高学生的研究能力有益。研究方法和研究计划合理，难度合中，学生能够在预定时间内完成该论文的设计。同意开题。  指导教师：  年 月 日 |

**四、开题报告论证记录**

|  |
| --- |
| **包括时间、地点、参加人、建议、提问及回答要点等**  记录（签名）：  年 月 日 |

|  |
| --- |
| **对报告人确定的论题的选题价值、可行性、是否同意撰写论文等做出结论**  主持人（签名）：  年 月 日 |
| **院学位评定分委员会意见**  主席（签名）：  年 月 日 |
| **校学位评定委员会办公室审核意见**  盖章：  年 月 日 |

**五、报告会结论**

**六、计划变动情况（无变动此页不填）**

|  |
| --- |
| **计划变动原因：**  计划变动申请人（签字）：    年 月 日 |
| **导师意见**    导师（签字）：  年 月 日 |
| **院学位评定分委员会意见**    主席（签字）：  年 月 日 |
| **校学位评定委员会办公室意见**    盖章：  年 月 日 |