**基于北京轨道交通数据的数据挖掘**

郭湘琦1 卢宇涵1 木枫凯1

1 (北京航空航天大学计算机学院 北京 100080)

(21373055@buaa.edu.cn)

**Data Mining Based On Beijing Rail Transit Data**

Guo Xiangqi1 Lu Yu Han1 Mu Fengkai1

1 (*School of Computer Science and Engineering , Beihang University, Beijing 100080*)

**Abstract**

This paper is a report on the final assignment of Beihang Data Mining Course, and the main work is to apply the clustering, classification, regression and other methods learned in the data mining course, using the traffic trajectory data of Beijing, to mine and analyze traffic travel patterns. We complete the four tasks required by the final assignment--data preprocessing (road network matching), cluster analysis (trajectory clustering), regression analysis (ETA estimation), and classification task (next-hop prediction). In the road network matching task, we use the road network mapping algorithm based on vector data, in the trajectory clustering task, we use the K-means clustering method, in the ETA estimation, we use the gradient boosted regression model, and in the next hop prediction task, we use the regression forest (a variant of random forest) method.

**Key words** cluster analysis, regression analysis, classification prediction

**摘要**

本文是北航数据挖掘课程的大作业课程报告，主要工作是用数据挖掘课程学习到的聚类、分类、回归等技术，结合北京交通轨迹数据，挖掘并分析交通出行规律。我们完成了大作业所要求的四个任务：数据预处理(路网匹配)、聚类分析(轨迹聚类)、回归分析(ETA估计)、分类任务(下一跳预测)。在路网匹配任务中，我们采用了基于矢量数据的路网映射算法；在轨迹聚类任务中，我们采用了K-means聚类方法；在ETA估计中，我们采用了梯度提升回归模型；在下一跳预测任务中，我们采用了回归森林(随机森林的一种变体)的方法。

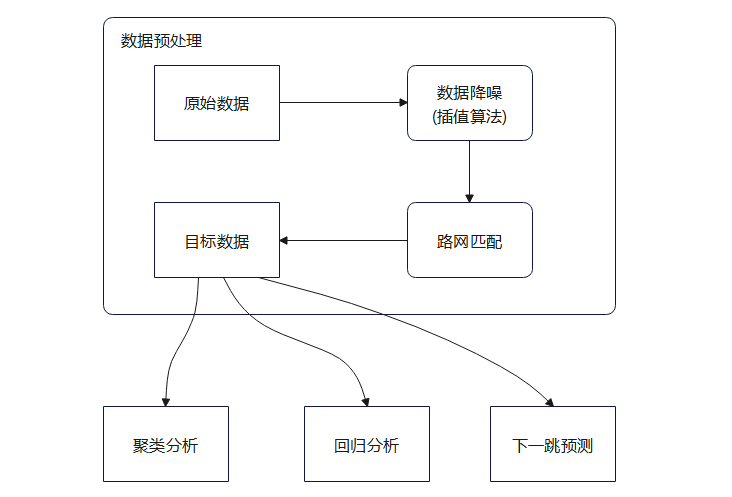
**关键词** 聚类分析 回归分析 分类预测

中图法分类号 TP391

轨道交通数据挖掘是指应用数据挖掘技术和方法对轨道交通系统中的数据进行分析和挖掘，以提取驾驶信息、车辆轨迹、有关运营、车辆行为模式等方面的有用信息和模式。轨道交通数据可以包括列车运行数据、乘客刷卡记录、车站人流数据、故障报告等。轨道交通数据挖掘的应用领域很广泛，在很多现实场景中具有很多应用：

* 运行优化：通过分析一些轨道交通的列车运行数据，可以优化列车的调度和运行策略，提高运行效率和准点率。
* 乘客行为分析：通过分析乘客刷卡记录和车站人流数据，可以了解乘客的出行模式、高峰时段和拥挤情况，从而进行合理的运力调配和乘客引导。
* 预测和规划：通过历史数据和趋势分析，可以预测未来的乘客需求和交通流量，从而进行合理的线路规划和资源配置。
* 到达时间估计：通过对大量的车辆行驶数据进行综合分析建模，可以对车辆预计到达时间进行非常准确的预估，方便乘客的出行安排，也带来了很大的商业价值。目前在一些打车软件中应用广泛。有不少打车软件还为司机提供提前接下一单的功能，这都是建立在对到达时间的精准预估的基础之上的。

我们的课程大作业是利用数据挖掘课程学习到的聚类、分类、回归等技术，结合北京交通轨迹数据，挖掘交通出行规律。我们的工作主要有以下几个部分:数据预处理、路网映射、聚类分析、回归分析，和下一跳预测(通过建模为分类问题来解决)。



1.数据预处理(路网映射)

轨迹数据是GPS形式的，不便于后续的数据处理工作，在使用这部分数据之前，我们对这部分数据进行了一定的数据的预处理。首先是GPS的噪声问题，由于卫星信号的传播延迟、接收器的硬件限制、大气条件、多径效应等各种因素。GPS测量的车辆轨迹数据通常会有随机误差，即会产生GPS噪声。GPS噪声可能会导致测量数据的不准确性和不稳定性。缓解这一问题的一种常见做法是应用地图匹配，即在数字化道路网络中将车辆轨迹与路段对齐。路网映射通常分为数据收集、数据预处理、以及路网建模(匹配)等几个大的步骤。在我们的课程作业中，数据收集这一步已经由课程组为我们统一安排，提供了相应的GPS数据。而在数据预处理的这一步，需要对数据进行降噪处理。我们采用了数据插补的方法，通过插值方法将数据扩充为更加光滑完整的形式。而在路网映射的这个环节，我们是应用了现有的路网映射算法。

1.1 数据插补(插值算法)

插值方法是一种通过使用已知数据点来估计缺失数据点的技术。以下是一些常见的插值方法：

* 线性插值：线性插值是一种简单的插补方法，它假设在两个已知数据点之间的数据值按线性关系变化。通过计算两个已知数据点之间的斜率，可以估计缺失数据点的值。这种方法适用于数据变化较为平滑的情况。
* 多项式插值：多项式插值是一种通过拟合一个多项式函数来估计缺失数据点的值的方法。这种方法可以使用已知数据点来构建一个多项式函数，并使用该函数来计算缺失数据点的值。多项式插值可以提供较高的精度，但对于高次多项式插值可能存在过拟合的问题。
* 样条插值：样条插值是一种光滑插值方法，它使用多个多项式函数来逼近数据。样条插值方法可以提供比线性插值更平滑的结果，并且可以通过调整插值的参数来控制插值曲线的光滑程度。常见的样条插值方法包括线性样条插值、二次样条插值和三次样条插值。
* 拉格朗日插值：拉格朗日插值是一种基于拉格朗日多项式的插值方法。它通过构建一个满足已知数据点的拉格朗日多项式来估计缺失数据点的值。拉格朗日插值方法简单易懂，但对于大量数据点可能存在计算复杂度较高的问题。

由于我们所使用的基础数据较多，不太合适使用拉格朗日插值方法，而车辆的轨迹实质上是大体上是一个分段函数，所以采用多项式拟合也不太适合，所以主要去尝试使用了线性插值和样条插值这两种方法。

1.2 路网映射

作为一个课程作业，我们很难在一个短时间内设计出一个新的路网映射算法，所以我们首先考察了现有的路网映射算法，现有的路网映射算法大体上可以分为两类：基于矢量数据的路网映射算法和基于栅格数据的路网映射算法。

1.基于矢量数据的路网映射算法主要是将道路网络信息表示为图（Graph）的形式，通过图论算法对道路网络进行分析和处理。常见的基于矢量数据的路网映射算法包括：

* 最短路径算法：通过计算道路网络中的最短路径，可以实现导航系统、路线规划等功能。常见的最短路径算法包括迪杰斯特拉(Dijkstra)算法、贝尔曼-福特(Bellman-Ford)算法等。
* 网络流算法：通过计算道路网络中的流量分配，可以实现交通流量优化、路口信号灯控制等功能。常见的网络流算法包括最大流算法、最小费用最大流算法等。
* 图匹配算法：通过将GPS轨迹数据与道路网络进行匹配，可以实现车辆定位、路径推荐等功能。常见的图匹配算法包括隐马尔科夫模型(HMM)算法等。

2.基于栅格数据的路网映射算法主要是将道路网络信息表示为栅格(Grid)的形式，通过栅格计算和分析对道路网络进行处理。常见的基于栅格数据的路网映射算法包括：

* 基于距离场的算法：通过计算道路网络中每个点到最近道路的距离，可以实现道路网络的建模和分析。常见的基于距离场的算法包括快速行进(Fast Marching)算法、水平集(Level Set)算法等。
* 基于统计学习的算法：通过对道路网络中的空间和属性信息进行统计学习，可以实现道路网络的分类、预测等功能。常见的基于统计学习的算法包括支持向量机、决策树等。

我们采用的是基于矢量数据的路网映射算法，并采用了最短路径的策略。最短路径算法实现简单，易于理解和实现，可以处理有向图、无向图、带权图和非带权图等多种情况，但缺点是时间复杂度较高，尤其是在处理大规模图时，可能会出现性能问题。

2.聚类分析(轨迹聚类)

聚类分析是一种无监督学习方法，用于将数据集中的对象分成不同的组或簇，使得同一组内的对象相似度较高，而不同组之间的对象相似度较低。聚类分析可以帮助我们发现数据中的内在结构和模式，以及识别相似的数据点。

聚类分析的基本思想是通过计算数据点之间的相似性或距离来确定它们之间的关系。常用的聚类算法包括K均值聚类、层次聚类、DBSCAN等。我们主要采用了K均值聚类作为我们的聚类方法。

2.1 轨迹相似度的定义

在聚类分析中相似性度量指的是选择适当的相似性度量方法来计算数据点之间的距离或相似性。常用的方法包括欧式距离、曼哈顿距离、余弦相似度等。但我们需要聚类的是一条一条的轨迹，每条轨迹又由若干个点所构成，顺次连接这若干个点就构成了整个的轨迹。因而，要对轨迹进行聚类，就不可能像对简单的数据点那样直接去计算距离。而要对所有的轨迹作数据的预处理。一般来说，有以下几种常见的处理办法：

* 轨迹点合并：将每条轨迹中的点合并成一个整体表示。这样，每条轨迹就可以表示为一个多维向量，其中每个维度代表轨迹中的一个点的特征。然后，可以使用常规的聚类算法（如K-means、DBSCAN等）对这些合并后的轨迹进行聚类。
* 轨迹距离度量：定义一种适合衡量轨迹之间相似度的距离度量方法。例如，可以使用动态时间规整（Dynamic Time Warping，DTW）算法来计算两条轨迹之间的距离。然后，可以使用基于距离的聚类算法(如层次聚类、密度峰值算法等)对轨迹进行聚类。
* 轨迹特征提取：从每条轨迹中提取出一组特征，然后使用这些特征来表示轨迹。例如，可以计算每条轨迹的长度、方向变化、速度变化等特征。然后，可以使用常规的聚类算法(如K-means、DBSCAN等)对这些特征进行聚类。

车辆的轨迹长度变化很大，表征轨迹的离散点的数量也在一个很大的范围中浮动，如果采用轨迹点合并的办法，将会很容易导致我们合并出来的点表征轨迹时出现很大偏差的情况，一些比较边缘的轨迹点将会对整个轨迹的合并表示造成很大的影响。而DTW算法的计算复杂度较高，在大规模数据集上可能会受到性能限制。因此我们权衡之下，我们采用了特征提取的方法。

我们要对每条轨迹进行特征提取(主要是位置信息和速度信息)。将提出出来的特征向量作为一条轨迹的表征。再将所有的特征向量当作n维几何空间中的点，然后进行聚类。

2.2 聚类算法的选取

在聚类算法选取中，我们选取了最为常用的K-means算法。K-means算法具有以下的优点：

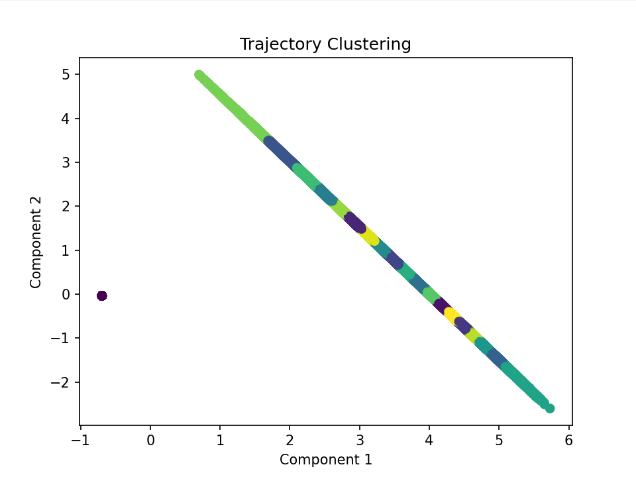
* 简单而高效：K-means算法是一种简单而易于理解的聚类算法，计算效率较高，尤其适用于大规模数据集。我们的数据集并不算小，在本地环境上运行极大地受到性能上的限制，这时K-means算法就成了一个相对更好的选择。
* 可扩展性：K-means算法可以很容易地扩展到处理大量数据和高维数据。我们要处理的轨迹数据特征是高维的，在这一点上K-means能够较好的支持。
* 可解释性：K-means算法产生的聚类结果相对直观，每个聚类簇都有一个代表性的中心点。

当然，K-means算法也存在着一些问题，比如对K值的选择没有准则可依循，聚类结果的好坏依赖于对初始聚类中心的选择，容易陷入局部最优解，对异常数据较为敏感，只能处理数值属性的数据，聚类结果可能不平衡等等。不过我们的数据上来自实际的GPS数据，异常数据出现的可能性是很小的，其次就是对于初始聚类中心和K值选择我们也进行了多次尝试，以尽可能地规避一部分K-means算法所带来的缺陷。

另外，由于我们的特征向量是高维的，在最后的可视化过程中，我们对原数据进行 PCA 降维处理，然后再进行可视化。

2.3 聚类效果的评估

下图是我们对原数据进行PCA降为处理再进行可视化得到的结果，可以看到，我们将轨迹数据成功地分为了若干簇。



3.回归分析(ETA分析)

3.1 算法设计

我们采用了梯度提升回归模型(Gradient Boosting Regression Model)来做ETA分析,梯度提升回归模型是一种集成学习方法，通过将多个弱预测器(如决策树)组合成一个强预测器来进行回归任务。该模型的主要思想是通过迭代训练多个弱预测器，并使得每个弱预测器都能够对前一个弱预测器的残差进行拟合。在每一轮迭代中，通过计算当前模型的梯度(即残差)和选取一个新的弱预测器来逐步拟合残差。最终，将所有的弱预测器进行加权求和，得到最终的预测结果。梯度提升回归模型的优点包括：

* 可以灵活地处理各种类型的数据，包括连续型、离散型和分类型数据；
* 可以自动处理特征之间的相互作用；
* 对异常值具有一定的鲁棒性。

当然，梯度提升回归模型也存在着如下的一些缺点：

* 对于大规模数据集和高维数据，训练时间可能较长；
* 对于噪声较多的数据，容易过拟合。

在我们处理的问题中，数据集规模尚可，训练时间在可接受的范围之内，GPS数据的噪声也较少，不太容易产生过拟合，故选择梯度提升回归模型是一个不错的选择。为了应用梯度提升回归模型，我们进行了多次尝试，以期选择合适的损失函数、决策树的最大深度、学习率等超参数，并进行交叉验证和调参来优化模型的性能。

3.2 算法实现

梯度提升回归模型的实现主要分为以下几个步骤：

1. 初始化参数：初始化回归模型的参数，例如回归系数和截距。
2. 定义损失函数：选择适当的损失函数来衡量模型的误差。
3. 计算梯度：使用梯度下降法来更新模型的参数。梯度表示损失函数对于每个模型参数的变化率，通过反向传播算法计算得到。
4. 更新参数：根据计算得到的梯度值，更新模型的参数。通过迭代更新参数，逐渐降低损失函数的值。
5. 模型训练：使用训练集对模型进行训练。重复执行步骤3和步骤4，直到达到收敛条件(如达到最大迭代次数或损失函数变化较小)。
6. 模型评估：使用测试集评估训练得到的模型的性能。我们采用了均方误差(MSE)、均方根误差(RMSE)。以及决定系数(R²)这三个评估指标来对我们的模型进行评估
7. 模型应用：根据需要，使用已训练好的模型进行预测。将输入特征提供给模型，计算预测值。

3.3 评估方案

我们采用了均方误差、均方根误差、以及决定系数这三个评估指标来对我们的模型进行评估，评估结果如下：

|  |  |
| --- | --- |
| 均方误差(MSE) | 29463.70229332763 |
| 均方根误差(RMSE) | 171.6499411398898 |
| 决定系数(R2) | 0.9999425532867326 |

**4.分类任务(下一跳预测)**

4.1 问题建模——如何将下一跳预测建模为一个分类问题

车辆的行驶过程中会产生很多的特征数据，一方面是车辆的特征数据：包括当前行驶的距离，当前行驶的速度，每一个记录点所在的经度和纬度；另一方面是当前行驶的道路的特征数据：道路的类别：是否是桥梁，是否是隧道，是否是高速公路，等等。此外还包括一些其他信息，例如当前是节假日还是非节假日。我们要做下一跳预测，本质上就是利用这些已知的信息，去不断地缩小车辆在下一个时间点可能出现的范围。每一个限制都是很重要的，例如车辆一个时间点还在高速公路上跑着，下一个时间出现在一个小路上的可能就是很低的；又如车辆当前的行驶速度很慢，一个时间跃迁到一个很远的距离也不太可能——这本质上就是一种分类，通过是否在高速公路上将车辆可能出现的地点划分为了两个集合。以此类推，对于其他的一些数据特征也是同样的道理，它们都能将车辆可能出现的地方分成几个类别，换言之就是打上标签，通过多种标签的重叠使用就可以通过分类的方法，确认车辆最有可能出现的位置。

4.2 算法实现

我们采用了回归森林(Random Forest Regression)的方法。回归森林是一种集成学习方法，它通过组合多个决策树来进行回归任务。它是随机森林算法的一种变体，基本步骤如下：

1.准备用于训练和测试的数据集。每个样本包含一组特征和相应的标签。。在我们要处理的车辆的下一跳问题中，最重要的特征数据就是当前的时间(hour-minute-second)，经度(lat)、纬度(lng)、车辆行驶速度(speeds)车辆当前行驶的距离(current\_dis)，当前是否为节假日(holidays)。

2.随机选择特征：从原始特征中随机选择一部分特征，这样可以减少模型对于某些特征的依赖性，增加模型的多样性。

3.构建决策树：对于每个随机选择的特征子集，使用训练数据构建一个决策树。决策树根据特征的取值进行分割，将数据划分为不同的子节点，直到达到终止条件(例如，节点中的样本数量小于某个阈值或达到了树的最大深度)。

4.集成预测：对于分类问题，随机森林中的每个决策树都会对样本进行预测，并根据投票或平均的方式来确定最终的预测结果。随机森林中的每个决策树会给出一个预测值，最终的预测结果是所有这些预测值的平均值。

5.模型评估：使用测试数据评估随机森林模型的性能。常用的评估指标包括准确率、精确率、召回率、F1值以及均方误差、平均绝对误差等。同ETA分析时的做法一样，我们采用了均方误差(MSE)，均方根误差(RMSE)，以及决定系数(R²)这三个评估指标来对我们的模型进行评估。

4.3 评估结果

|  |  |
| --- | --- |
| 均方误差(MSE) | 7.54288936369653e-06 |
| 均方根误差(RMSE) | 027464321152536305 |
| 决定系数(R2) | 09980588932748192 |

**参 考 文 献**

[1] Li X, Zhao K, Cong G, et al. Deep representation learning for trajectory similarity computation[C]//2018 IEEE 34th international conference on data engineering (ICDE). IEEE, 2018: 617-628.

[2] Fu Z, Hu W, Tan T. Similarity based vehicle trajectory clustering and anomaly detection[C]//IEEE International Conference on Image Processing 2005. Ieee, 2005, 2: II-602.

[3] Choong M Y, Chin R K Y, Yeo K B, et al. Trajectory pattern mining via clustering based on similarity function for transportation surveillance[J]. International Journal of Simulation-Systems, Science & Technology, 2016, 17(34): 19.1-19.7.

[4] Toohey K, Duckham M. Trajectory similarity measures[J]. Sigspatial Special, 2015,7(1): 43-50.

[5] Landau S, Ster I C. Cluster analysis: overview[J]. Á Á, 2010, 11(x12): x1p.

[6] Petzoldt T, Ngoc Q H S, Bogda K. Time to arrival estimates,(pedestrian) gap acceptance and the size arrival effect[C]//Driving Assessment Conference. University of Iowa, 2017, 9(2017).

[7] Tan C W, Park S, Liu H, et al. Prediction of transit vehicle arrival time for signal priority control: Algorithm and performance[J]. IEEE Transactions on Intelligent Transportation Systems, 2008, 9(4): 688-696.

[8] Ayhan S, Costas P, Samet H. Predicting estimated time of arrival for commercial flights[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 33-42.

[9] Liu K, Li Y, He F, et al. Effective map-matching on the most simplified road network[C]//Proceedings of the 20th International Conference on Advances in Geographic Information Systems. 2012: 609-612.

[10] Ochieng W Y, Quddus M A, Noland R B. Map-matching in complex urban road networks[J]. Brazilian Journal of Cartography (Revista Brasileira de Cartografia), 2003, 55(2): 1-18.

[11] Tang Y, Zhu A D, Xiao X. An efficient algorithm for mapping vehicle trajectories onto road networks[C]//Proceedings of the 20th International Conference on Advances in Geographic Information Systems. 2012: 601-604.

[12] Zhang Y, Haghani A. A gradient boosting method to improve travel time prediction[J]. Transportation Research Part C: Emerging Technologies, 2015, 58: 308-324.

[13] Ding C, Wu X, Yu G, et al. A gradient boosting logit model to investigate driver’s stop-or-run behavior at signalized intersections using high-resolution traffic data[J].Transportation Research Part C: Emerging Technologies, 2016, 72: 225-238.

[14] Zhang L, Alharbe N R, Luo G, et al.A hybrid forecasting framework based on support vector regression with a modified genetic algorithm and a random forest for traffic flow prediction[J]. Tsinghua Science and Technology, 2018, 23(4): 479-492.

[15] Cheng R, Zhang M M, Yu X. Prediction model for road traffic accident based on random forest[C]//4th International Conference on Education Science and Development. 2019.