**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**

**Key words**

**摘要**

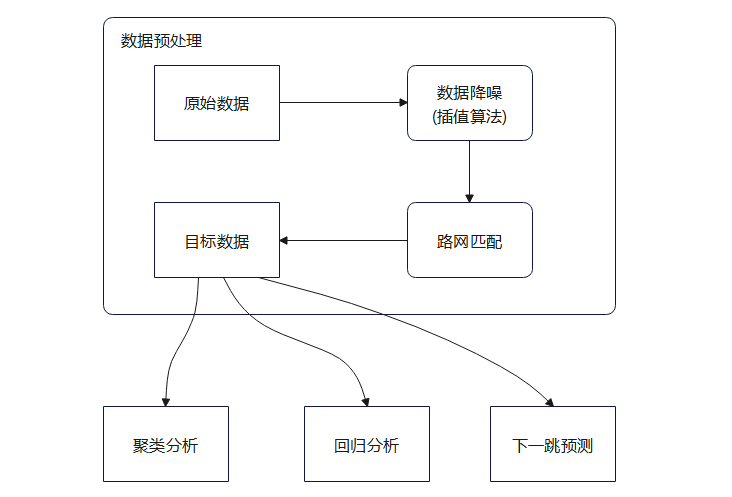
**关键词**

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，以提取有关运营、乘客行为、安全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，例如：

* 运行优化：通过分析列车运行数据，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。
* 安全管理：通过分析故障报告和设备监测数据，可以提前发现潜在的设备故障和安全隐患，并采取相应的维护和修复措施。
* 乘客满意度评估：通过分析乘客反馈数据和调查数据，可以评估乘客对轨道交通服务的满意度，并找出改进的方向。
* 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。

我们的课程报告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(路网匹配)

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测量数据可能会产生随机误差，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。为此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。

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

2.聚类分析(轨迹聚类)

2.1 轨迹相似度的定义

2.2 聚类算法的选取

2.3 聚类效果的评估

3.回归分析(ETA分析)

3.1 算法设计和研究点

3.2 算法实现

3.3 评估方案

4.分类任务(下一跳预测)

4.1 问题建模

4.2 算法实现

4.3 评估结果