基于逻辑回归的驾驶员发生事故分类

中文摘要

本项目使用了机器学习的方法来对发生过事故的驾驶员进行分类，目的是预测他们是否有高风险再次发生事故。项目使用了一个包含了驾驶员的个人信息（如年龄，性别，驾龄等）和工作信息（如工作时长，违章操作，事故等）的数据集。项目采用了逻辑回归算法来建立分类模型，并使用了网格搜索来优化模型的参数。项目的结果表明模型具有较好的性能和泛化能力。本项目为发生过事故的驾驶员提供了一个有效的风险评估工具，可以帮助他们提高安全意识和防范能力。

关键词： 机器学习，目标分类，逻辑回归

英文摘要

This project uses machine learning methods to classify drivers who have experienced accidents, with the aim of predicting whether they have a high risk of recurring accidents. The project used a dataset that includes personal information of drivers (such as age, gender, driving experience, etc.) and work information (such as work hours, violations, accidents, etc.). The project uses the Logistic regression algorithm to establish the classification model, and uses grid search to optimize the parameters of the model. The results of the project indicate that the model has good performance and generalization ability. This project provides an effective risk assessment tool for drivers who have experienced accidents, which can help them improve their safety awareness and prevention capabilities.

Keywords: Machine learning, Target classification, Logistic regression

目录

[基于逻辑回归的驾驶员发生事故分类 1](#_Toc138883548)

[中文摘要 1](#_Toc138883549)

[英文摘要 1](#_Toc138883550)

[第一章 绪论 3](#_Toc138883551)

[第二章 方法 4](#_Toc138883552)

[2.1 数据清洗 4](#_Toc138883553)

[2.2 特征提取整合 5](#_Toc138883554)

[2.3 逻辑回归分类 6](#_Toc138883555)

[第三章 数据集 7](#_Toc138883556)

[第四章 实验 7](#_Toc138883557)

[第七章 结论 9](#_Toc138883558)

# 第一章 绪论

道路交通事故是一种严重的社会问题，每年造成数百万人的死亡和伤残。驾驶员的安全行为是影响事故发生的关键因素之一，因此，对驾驶员进行分类和分析，有助于识别不同类型的驾驶员在不同情境下的行为特征和风险倾向，从而制定针对性的安全教育和干预措施，提高道路交通安全水平。然而，目前对驾驶员分类的研究主要基于主观的问卷调查或人为的实验室模拟，缺乏对客观的真实事故数据的分析和利用。[1][3][5][6][8]

机器学习是一种利用数据进行自动学习和预测的技术，近年来在各个领域得到了广泛的应用和发展。[2][3][4]机器学习可以处理大量复杂的数据，挖掘其中的潜在规律和模式，提供有效的分类和预测结果。机器学习在交通领域也有许多应用，例如交通流量预测、交通信号控制、交通事故检测等。然而，机器学习在驾驶员分类方面的应用还比较少见，尚缺乏系统的方法和评估。[6][7]

驾驶员的安全是交通管理的重要问题之一，事故的发生不仅会造成人员伤亡和财产损失，还会影响交通效率和社会稳定。[9][10[因此，对驾驶员的风险评估和预防措施是必要的。然而，传统的风险评估方法通常基于人工的经验和判断，缺乏客观性和准确性，也难以适应大规模和复杂的数据环境。为了解决这些问题，本项目提出了一种基于机器学习的方法来对发生过事故的驾驶员进行分类，目的是预测他们是否有高风险再次发生事故。机器学习是一种利用数据来自动学习和优化模型的技术，具有强大的数据处理和分析能力，可以从复杂的数据中提取有用的信息和知识。我们收集了一些真实的交通事故数据，包括事故发生的时间、驾驶员等信息，以及事故后驾驶员的伤情、责任、违法等信息。我们使用了常用的机器学习算法—逻辑回归分类，对驾驶员进行了多维度的分类，如安全型、服务型等，并比较了算法的分类性能和准确度。我们希望通过本文能够为驾驶员分类提供一种新的思路和工具，为道路交通安全管理提供一些参考和建议。[4][6][7]

# 第二章 方法

## 2.1 数据清洗

数据清洗是指对数据进行重新审查和校验的过程中，发现并纠正数据文件中可识别的错误，按照一定的规则把错误或冲突的数据洗掉，包括检查数据一致性，处理无效值和缺失值等，数据清理一般是由计算机而不是人工完成。利用数理统计、数据挖掘和预定义清理规则等有关技术将“脏数据”处理掉，从数据源中检测并消除错误、不一致、不完整和重复等数据，为满足要求提供高质量的数据。数据清理的标准模型是将数据输入到数据清理处理器，通过一系列步骤清理数据，然后以期望的格式输出清理过的数据。[2][4][6][9]

在本次实验中，首先统计了不同表格的数据缺失情况，对于缺失的数据我们选择性的将部分数据进行替代或直接删除数据。例如驾驶员信息中对于学历信息的缺失，我们认为学历信息对于驾驶员是否发生事故有着很大的影响，直接删除会导致不良影响，于是我们采用分类，将缺失的部分划分成新的一类学历，也便于增强模型的鲁棒性。而对于一些我们认为无关的信息例如拿证时间，则直接删除。

之后对于一些非数值化的参数，我们进行了参数更新，例如生日，工作日期，我们计算了具体时长进行替代，便于模型进行参数选择与训练。

## 2.2 特征提取整合

由于原有数据集包含了很多表，我们需要对其中属性进行区分合并成我们需要的数据进行模型训练，因此需要我们对表中的数据进行选择合并。在驾驶员信息表中，我们选取了性别，学历，年龄，工龄；对工作时长表，我们通过统计分析，发现工作时长对于事故的影响并不具备代表性，如图1所示：

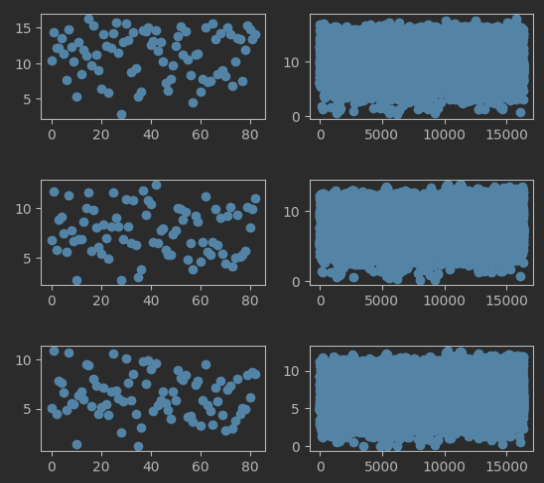


图1

左侧为发生事故的工作时长，右侧位未发生事故的工作时长，发现并没有显著区别，因此不考虑工作时长影响；我们还统计了发生违章情况时发生事故的情况，具体如图2、3所示：

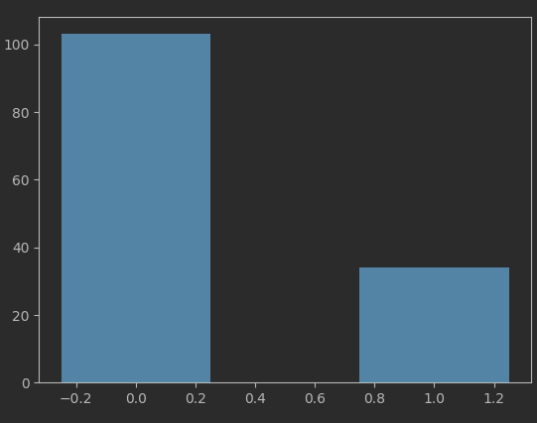
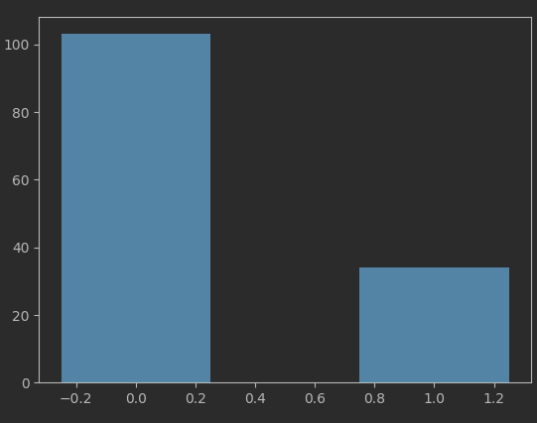
 

图2 图3

图2中展示了服务违章对应的事故发生，图3展示了安全违章对应的事故发生，都具有一定的影响力，于是我们选择了将是否发生违章情况作为特征进行训练。

## 2.3 逻辑回归分类

逻辑回归是一种广泛使用的分类算法。它是一种线性模型，用于将输入变量与输出变量之间的关系建模为概率。逻辑回归使用逻辑函数来执行任务。逻辑函数将任何实值输入映射到0和1之间的值。在二元分类中，逻辑回归模型输出一个介于0和1之间的概率值，表示输入属于类别1的概率。如果该概率大于或等于0.5，则将输入分类为类别1；否则，将其分类为类别0。逻辑回归的训练过程是通过最大化似然函数来完成的。似然函数是给定模型参数和训练数据时数据生成过程的概率。最大化似然函数等价于最小化交叉熵损失函数。[6][9]

在本次实验中，使用逻辑回归做分类任务，模型输入为选取的6项特征，输出为分类结果的概率值。

# 第三章 数据集

数据集采用课题数据集，共有444条个人信息，对444人进行分类。为了确保模型正常训练，实验将数据集中的部分文字或日期数据进行更新，进行数字化。

# 第四章 实验

本次实验在windows系统下，使用pycharm软件开发。

实验中对总计444名驾驶人员进行是否发生事故检测，选择了6项特征如图4：年龄，性别，工龄，学历，服务违章，安全违章进行模型训练，使用自己编写的逻辑回归模型进行分类。

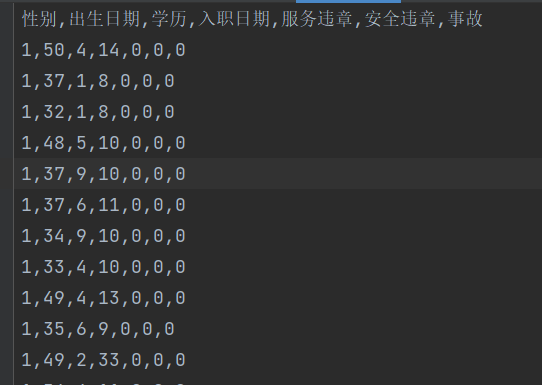


图4

将清洗整合好的特征数据统一存放在cvs文件中。

在模型分类中，我们绘制了模型的分类的P-R曲线以及ROC曲线，具体如图5、6所示：

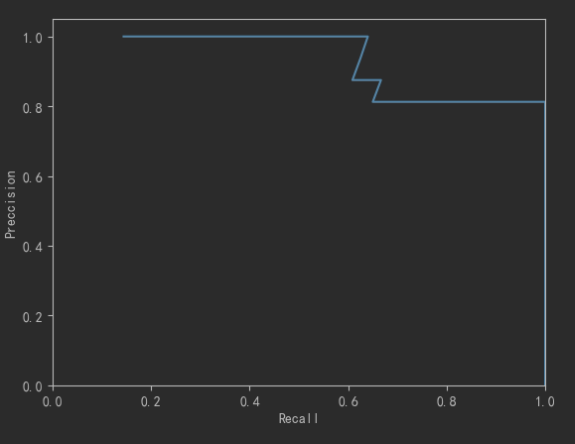


图5

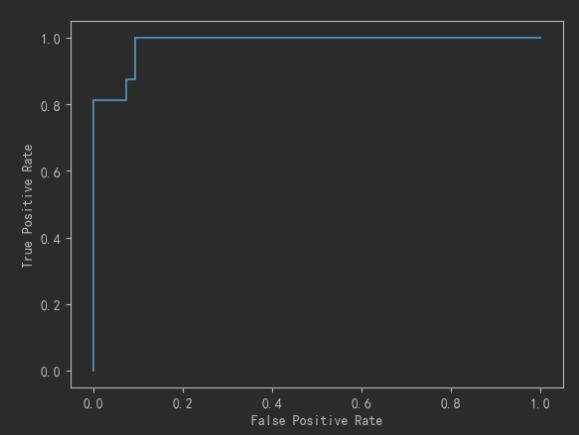


图6

根据图5、6可以发现P-R曲线是以召回率（Recall）为横坐标，精确率（Precision）为纵坐标的曲线，用于评估分类器在不同阈值下的性能。ROC曲线是以假正例率（False Positive Rate）为横坐标，真正例率（True Positive Rate）为纵坐标的曲线，用于评估分类器在不同阈值下的性能。通过P-R曲线和ROC曲线，我们了解分类器在不同阈值下的性能。我们取0.2-0.6的阈值可以获得最好的效果。

# 第七章 结论

逻辑回归是一种常用的机器学习方法，它可以对发生事故驾驶员进行分类，从而分析事故原因和预防措施。逻辑回归通过建立线性函数与对数几率之间的关系，来拟合数据中类别之间的差异。逻辑回归不仅可以输出分类结果，还可以输出分类概率，这对于评估风险和制定策略很有帮助。逻辑回归通过极大似然估计法来求解模型参数，使得数据出现的可能性最大。逻辑回归的损失函数是一个凸函数，可以通过各种优化算法来求解最优解。通过使用逻辑回归，我们可以得到一个简单而有效的分类模型，并且可以方便地进行解释和推理。但是其在驾驶员分类方面的应用还比较少见，尚缺乏系统的方法和评估。本文通过采取相应的特征选择进行逻辑回归训练，并根据P-R曲线和ROC曲线进行分析其中的性能。

但是简单的逻辑回归模型存在着不能很好地处理非线性问题，因为其决策边界是线性的；特征空间很大时，性能不好；容易欠拟合，分类精度不高；逻辑回归本身无法筛选特征，需要结合其他方法进行特征选择等问题，而本文中则根据相应的特征选择方法进行优化，发现了其中关联度较大的特征并进行训练。

本文使用机器学习对发生事故的驾驶员进行了分类，从多个维度对驾驶员的行为特征和风险倾向进行了分析和比较。我们发现，不同类型的驾驶员在事故发生的时间、环境、服务等方面有显著的差异，也有不同的伤情、责任、违法等后果。本文为驾驶员分类提供了一种新的思路和工具，为道路交通安全管理提供了一些参考和建议。未来的研究可以进一步扩大数据量和范围，优化机器学习算法和参数，探索更多的分类维度和指标，以提高驾驶员分类的效果和意义。

**参考文献**

1. **Factors influencing accident severity: an analysis by road accident type，Laura Eboli, Carmen Forciniti, Gabriella Mazzulla，2020.8.28**
2. **Using logistic regression to estimate the influence of accident factors on accident severity，Ali S Al-Ghamdi，2002.11**
3. **Predicting individuals’car accident risk by trajectory, driving events and driver characteristics，Yan Li, Xianghao Xu, Zhiyuan Liu, et al.，2021.2**
4. **A comparative study of logistic regression and artificial neural networks for driver injury severity analysis in traffic accidents，Mehmet Eren, Mehmet Çelik, 2019.5**
5. **A logistic regression model for predicting the severity of motorcycle crashes in Malaysia，Mohd Syazwan Solah, Mohd Hafzi Md Isa, Radin Zaid Radin Umar, et al.，2020.12**
6. **A literature review of machine learning algorithms for crash injury severity prediction，Kenny Santos, João P. Dias, Conceição Amado，2022.2**
7. **Prediction in Traffic Accident Duration Based on Heterogeneous Ensemble Learning，Yuexu Zhao, Wei Deng，2022.1**
8. **A logistic regression model for predicting the severity of pedestrian crashes in urban areas of Iran，Mohammad Mehdi Khabiri, Amir Masoud Rahimi, 2016.4**
9. **A logistic regression model for predicting the severity of run-off-road crashes in Victoria, Australia，Samantha L Cockfield, Brian N Fildes,2001.7**
10. **Logistic regression analysis of factors influencing the injury severity of single-vehicle crashes in Japan，Yoshitsugu Hayashi, Toshiyuki Yamamoto, 2000.1**