基于决策树的驾驶员发生事故分类

中文摘要

当使用机器学习对发生过事故的驾驶员做分类时，可以使用各种算法，例如逻辑回归、支持向量机、决策树、随机森林等。这些算法可以使用驾驶员的个人信息以及驾驶情况等数据来预测驾驶员是否会发生事故。此外，还有一些研究使用了自监督一致性学习的方法来检测交通事故。本文使用决策树机器学习方法通过对驾驶员相关信息的分析，可以分类驾驶员是否会发生事故，从而提高交通安全性。总之，机器学习是一个非常有前途的技术，它可以帮助我们更好地分类驾驶员是否有事故发生，并为我们提供更好的解决方案。

关键词：机器学习，决策树，人员分类

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

[基于决策树的驾驶员发生事故分类 1](#_Toc138886290)

[中文摘要 1](#_Toc138886291)

[英文摘要 1](#_Toc138886292)

[第一章 绪论 3](#_Toc138886293)

[第二章 方法 4](#_Toc138886294)

[2.1 特征工程 4](#_Toc138886295)

[2.2 模型方法 6](#_Toc138886296)

[第三章 实验 6](#_Toc138886297)

[第七章 结论 8](#_Toc138886298)

# 第一章 绪论

决策树是一种常用的分类算法，可以用于驾驶员发生事故分类的任务。决策树通过对数据集进行递归划分，生成一棵树状的结构，每个内部节点表示一个属性测试，每个分支表示一个测试结果，每个叶节点表示一个类别标签。决策树的优点是易于理解和解释，可以处理数值和类别数据，可以处理缺失值和异常值，可以进行特征选择和剪枝等。[1][2][6][7]

驾驶员发生事故分类是指根据驾驶员的个人特征、驾驶行为、交通环境等因素，预测驾驶员在发生事故时的损伤程度，如无伤、轻伤、重伤或死亡。这是一个重要的研究课题，因为它可以帮助分析事故的原因和影响因素，提出有效的预防措施和干预策略，提高道路安全水平和社会福利。[5][8]

使用决策树对驾驶员发生事故分类的研究已经有了一定的进展。一些研究者使用不同的决策树算法和数据集[3][9][10]，探讨了影响驾驶员发生事故分类的各种因素，如年龄、性别、教育程度、车辆类型、天气条件、路面状况、交通流量等。这些研究为理解和预测驾驶员发生事故分类提供了一些有价值的见解和启示。

本文提出了一种基于决策树对驾驶员发生事故分类的新方法。本文的主要贡献和创新点有以下几点：(1)本文采用了一种综合性的数据预处理方案，包括数据清洗、缺失值填补、异常值检测、数据平衡化等步骤；（2）本文解释了决策树模型的结果，并根据不同的节点选择获取了最佳的特征组合。

# 第二章 方法

## 2.1 特征工程

特征工程是机器学习中非常重要的一步，它可以帮助我们从原始数据中提取有用的特征，以便更好地训练模型。在使用决策树对发生过事故的驾驶员做分类时，特征工程可以帮助我们选择最相关的特征，以便更好地预测驾驶员是否会发生事故。在特征工程中，我们可以使用各种技术来选择最相关的特征。例如，我们可以使用相关性分析来确定每个特征与标签之间的相关性。此外，我们还可以使用主成分分析（PCA）等技术来减少特征的数量，并确保每个特征都是独立的。[3][5]

首先本文先对数据进行特征清洗，我们发现在现有的数据集中依旧存在不少的空值，这对于模型训练而言是无法使用的，所以我们采用pandas库对缺失的数据信息进行删除，删除后得到的数据为：414条驾驶员信息，20033条工作时长信息，125条服务违章信息，105条安全违章信息，83条事故信息。

之后本文对数据进行特征预处理，主要包括数据变换以及衍生变量。由于数据中存在大量的日期数据以及文本信息，这些信息对于数值计算的决策树而言并不能很好的识别与处理，于是本文将日期数据进行计算换成成时长信息，用于统计年龄、工龄；而文本信息，本文采用数值化one-hot编码方法，对于类似学历、违章内容等，进行编码操作，转换成对应的数值信息，便于模型进行特征提取。

最后本文对特征进行选取整合。由于现有的数据集中包含了大量的冗余信息，我们进行相关性分析。首先是我们统计了相关工作时长的事故发生折线图，如图1所示：

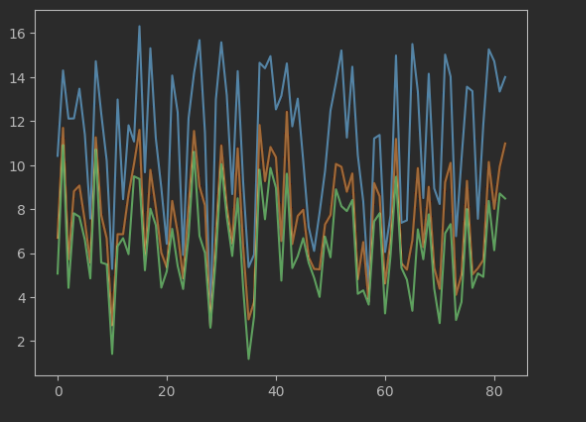


图1

可以发现在事故发生的情况下，对于不同的工作时长的范围波动很大，具备很强的不一致性，因此对于相关的工作时长，我们认为这与事故发生的关系并不显著，因此丢弃了时长的特征信息。之后我们对相关的违章信息与事故进行统计分析，如图2、3所示：

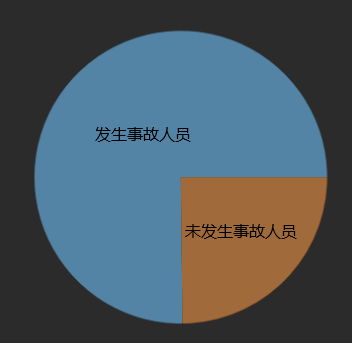
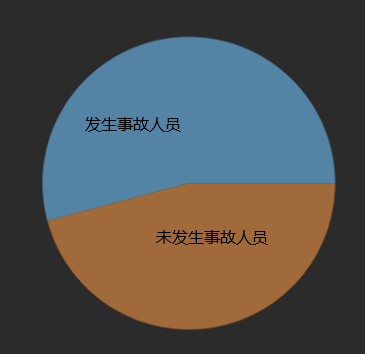
 

图2 图3

通过图2和图3可以发现，无论是服务违章还是安全违章的情况，都是发生事故的次数更多，我们有理由相信相关的违章与事故的联系十分紧密，因此将发生服务违章与安全违章作为特征。

## 2.2 模型方法

使用决策树作为分类器对发生过事故的驾驶员进行分类，需要进行特征工程。特征工程是指从原始数据中提取有用的信息，以便更好地预测未来的结果。在这种情况下，我们需要选择最相关的特征，以便更好地训练模型。决策树分类器的工作原理是将数据集分成多个小的子集，每个子集都包含相同数量的数据点。然后，它会在每个子集上构建一棵决策树，并使用这些决策树来预测未来的结果。使用决策树作为分类器对发生过事故的驾驶员进行分类需要进行特征工程，以便更好地预测未来的结果。[1][6][9][10]

# 第三章 实验

通过使用决策树对驾驶员进行发生事故分类，选择具有最高信息增益的特征作为测试特征，利用该特征对节点样本进行划分子集，会使得各子集中不同类别样本的混合程度最低,在各子集中对样本划分所需的信息(嫡)最少，本次实验使用GINI系数进行计算。

模型算法结构：

* 从根节点出发，根节点包括所有的训练样本。
* 一个节点（包括根节点)，若节点内所有样本均属于同一类别，那么将该节点就成为叶节点，并将该节点标记为样本个数最多的类别。
* 否则利用采用信息增益法来选择用于对样本进行划分的特征，该特征即为测试特征，特征的每一个值都对应着从该节点产生的一个分支及被划分的一个子集。在决策树中，所有的特征均为符号值，即离散值。如果某个特征的值为连续值，那么需要先将其离散化。。递归上述划分子集及产生叶节点的过程，这样每一个子集都会产生一个决策(子）树，直到所有节点变成叶节点。
* 递归操作的停止条件就是:

1. 一个节点中所有的样本均为同一类别，那么产生叶节点
2. 没有特征可以用来对该节点样本进行划分，此时也强制产生叶节点，该节点的类别为样本个数最多的类别
3. 没有样本能满足剩余特征的取值，此时也强制产生叶节点，该节点的类别为样本个数最多的类别

生成的决策树模型可视化如图4所示：

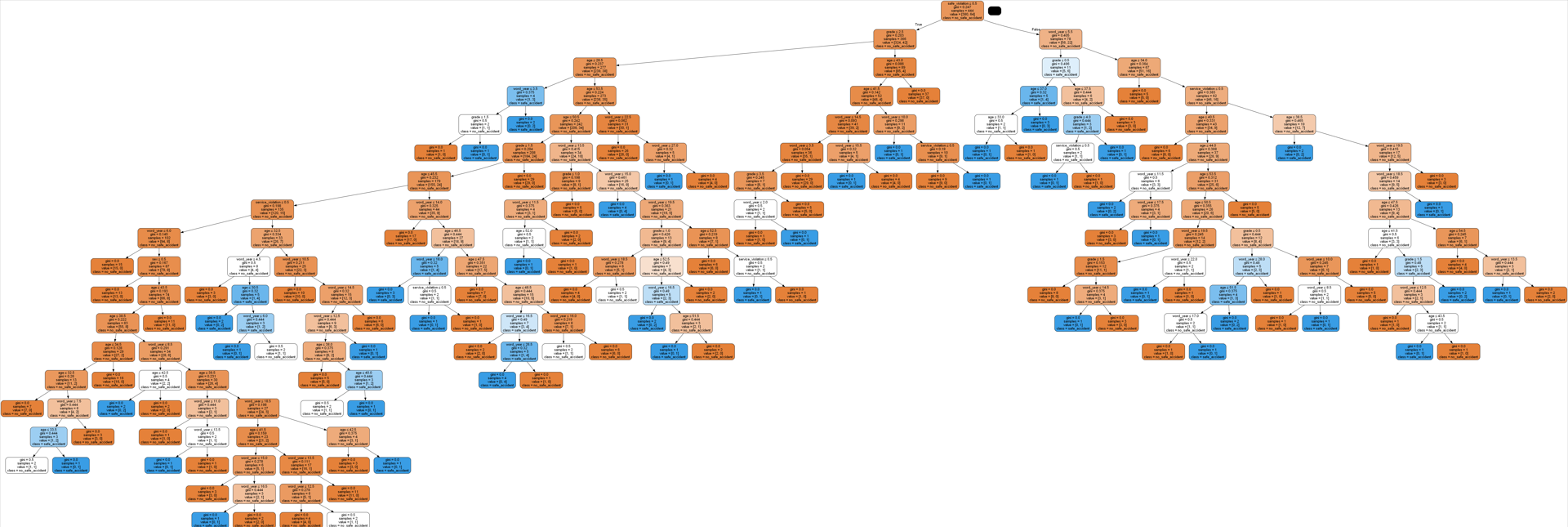


图4

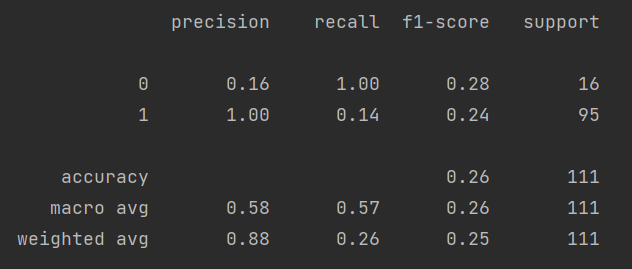


图5

由于决策树是阶级分类器，并不存在概率值的输出，只有分类结果的输出，所以无法绘制P-R曲线以及ROC曲线，本次实验直接展示了将数据集按1：3比例划分成训练集与测试集后，对于测试集的训练结果，可以看到分类结果良好。

# 第七章 结论

决策树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。它是一种树形结构（可以是二叉树或者非二叉树），其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果。决策树可以用于分类、预测、规则提取等领域，它的优点是简单易懂，不需要很多数据预处理，缺点是可能产生过拟合，即不能很好地推广到新的数据。决策树的构建过程是自上而下，分而治之的，核心问题是如何选择合适的属性对样本进行划分。决策树通过选择合适的属性对样本进行划分，直到达到某种终止条件或者达到最大深度。决策树的划分准则通常基于信息增益或者基尼指数等指标，目的是使得每次划分后样本集合的纯度或者均匀度增加。信息增益是基于信息熵和条件熵来计算的，它反映了特征对类别划分的贡献程度。

通过使用决策树，我们可以得到一个清晰可视化的分类规则，并且可以方便地进行解释和推理。使用决策树作为分类器对发生过事故的驾驶员进行分类的结果与结论取决于所选特征。例如，如果我们选择驾驶员的年龄、性别、驾龄、违章信息等作为特征，那么我们可以使用这些特征来构建一棵决策树，并使用这棵决策树来进行分类。如果我们选择其他特征，那么结果和结论可能会有所不同。总之，使用决策树作为分类器对发生过事故的驾驶员进行分类需要进行特征工程，以便更好地分类判断。

本文采用的综合性的数据预处理方案，包括数据清洗、缺失值填补、异常值检测、数据平衡化等步骤，能够很好的进行特征选择，并帮助模型取得了优秀的分类结果。

**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. **Driver impairment detection using decision tree based feature selection and classification，Mert Çetinkaya, Tankut Acarman，2023.3**
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, 2020.5**
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.，2021.11**
6. **A logistic regression model for predicting the severity of bicycle-motor vehicle crashes in China，Jian Sun, Xiaomeng Li, Yanyong Guo, et al.，2017.10**
7. **A logistic regression analysis of factors affecting injury severity of single-vehicle rollover crashes on rural highways in Iran，Mohammad Mehdi Khabiri, Amir Masoud Rahimi, 2016.11**
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**