基于机器学习的糖尿病识别

摘要

中国是世界上糖尿病患者最多的国家，病人达到1.1亿，每年有130万人死于糖尿病及其相关疾病。更严重的是，目前中国有超过6亿人处于糖尿病前期，处于这一阶段的人的血糖水平高于正常水平，如果不进行治疗，就有可能发展为2型糖尿病。传统的血糖监测方法需要多次抽血采样，而随着机器学习和大数据医疗的兴起，基于患者医疗体检的数据预测血糖，并判断是否有可能发展为糖尿病有了一定的可行性。

本文的研究目的在于利用机器学习预测患者的血糖值，并预测病人是否是糖尿病。为医生诊断提供参考，减少患者不必要的抽血检查。本文主要的工作包括以下几个方面：

（1）整合了多角度体检特征。如多种转氨酶，球蛋白含量，胆固醇的含量等。对患者数据进行预处理、均衡化，进行特征工程，更全面地描述患者身体指标。

（2）研究了多种回归预测模型，探究最优的参数配置。对包括传统模型如支持向量机、随机森林、XGBoost，最新模型如CatBoot 在内的多种模型及其参数进行了研究，通过多种指标比较了模型的性能，得到性能最佳的模型。

（3）模型融合进行血糖预测，通过stacking，boosting等方法进行了模型融合，进一步提高预测的准确度。

（4）特征重要性排序，在训练的过程中记录特征重要性，确定较重要的几个特征，为医生诊断提供额外的参考。

关键词： 糖尿病 血糖 机器学习

1.引言

（600~800字）

正常人的空腹血糖值为3.89～6.1mmol/L。如大于6.1mmol/L而小于7.0mmol/L为空腹血糖受损，如两次空腹血糖大于等于7.0mmol/L考虑糖尿病，建议复查空腹血糖，糖耐量试验。如果随机血糖大于等于11.1 mmol/L可确诊糖尿病。

目前中国有超过6亿人处于糖尿病前期，处于这一阶段的人血糖水平高于正常水平，也就是上述的6.1～11.1mmol/L之间。为了防止发展为2型糖尿病，需要常期去医院抽血验血糖，给患者带来了不便。随着机器学习和大数据医疗的兴起，基于患者医疗过往的体检数据预测血糖，并判断是否有可能发展为糖尿病有了一定的可行性。这将为医生决策提供参考，专注于大概率有可能发展为糖尿病的患者，同时减少普通患者的时间和经济成本。

本文主要的工作包括以下几个方面：

（1）整合了多角度体检特征。如多种转氨酶，球蛋白含量，胆固醇的含量等。对患者数据进行预处理、均衡化，进行特征工程，更全面地描述患者身体指标。

（2）研究了多种回归预测模型，探究最优的参数配置。对包括传统模型如支持向量机、随机森林、XGBoost，最新模型如CatBoot 在内的多种模型及其参数进行了研究，通过多种指标比较了模型的性能，得到性能最佳的模型。

（3）模型融合进行血糖预测，通过stacking，boosting等方法进行了模型融合，进一步提高预测的准确度。

（4）特征重要性排序，在训练的过程中记录特征重要性，确定较重要的几个特征，为医生诊断提供额外的参考。

本文的内容安排如下，第2章阐述几个不同的算法模型原理及关键参数的意义，为训练提供理论指导。第3章对数据进行预处理，包括对查看数据分布和图形化显示，缺失值填充、重要性排序及归一化等，为模型输入做好准备。第4章开始训练、评估模型，并详细介绍了其调参的过程和意义，并对实验结果进行分析。最后通过模型融合得到最终的结果。第5章对本文进行总结和延伸思考。第6章是参考文献。

2. 模型简介

（约800字）

机器学习的算法有很多，本文这里选取了性能较好的常见模型如RandomForest、XGBoost等，也尝试了最新推出的CatBoost模型，最后尝试了CNN神经网络模型。这里对各个模型做简要介绍。

（1）随机森林（RandomForest）

模型介绍

重要参数意义

（2）极限梯度提升（XGBoost）

模型介绍

重要参数意义

（3）CatBoost

模型介绍

重要参数意义

（4）卷积神经网络（CNN）

模型介绍

重要参数意义

3.数据预处理及特征工程

3.1 数据集介绍

本数据集采用某数据平台提供的6663例患者的体检数据，已经过脱敏处理。数据集包括41个特征，如下图所示。

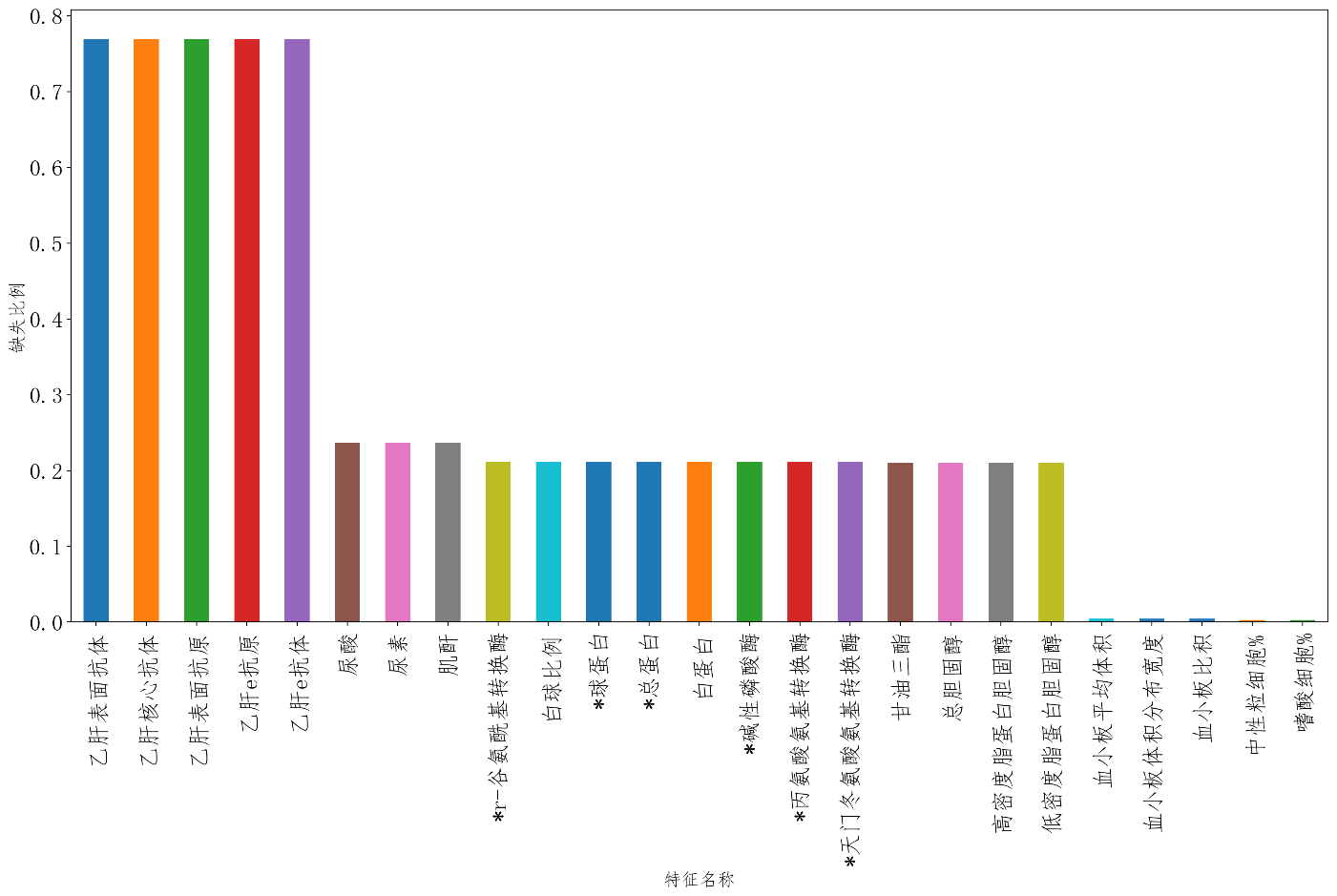
数据包含患者的基本信息如性别、年龄等，也包括详尽的生理病理学指标如氨基酸转换酶、球蛋白含量、淋巴细胞含量等。其中男性3268例，女性3395例，糖尿病患者1411例，占比21%。最后一列是要预测的血糖值。



3.2 数据及特征分析

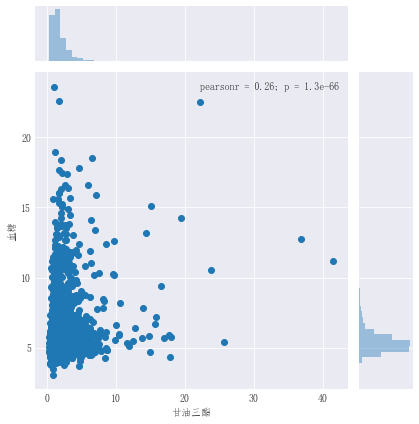
3.2.1 缺失值填充

首先查看数据的缺失值分布情况，对数据集每列中的缺失值占比进行统计，排序后将前25项特征进行可视化显示如下。由图可见，乙肝表面抗体, 乙肝核心抗体, 乙肝表面抗原, 乙肝e抗原, 乙肝e抗体等5个特征缺失值在70%以上，可以直接做删除处理。剩下的缺失值采用-999进行填充。



3.2.2 查看单个特征与血糖的相关性

既然有这么多个特征，一个比较自然的想法就是先看看单个特征和血糖之间的关系。最理想的情况下，如果存在线性的关系则说明该特征对血糖有非常直接的影响。随机选取了10个特征进行可视化展示，如下图所示，发现并没有比较明显的函数关系。但之后我们可以通过机器学习模型找出较为重要的特征。

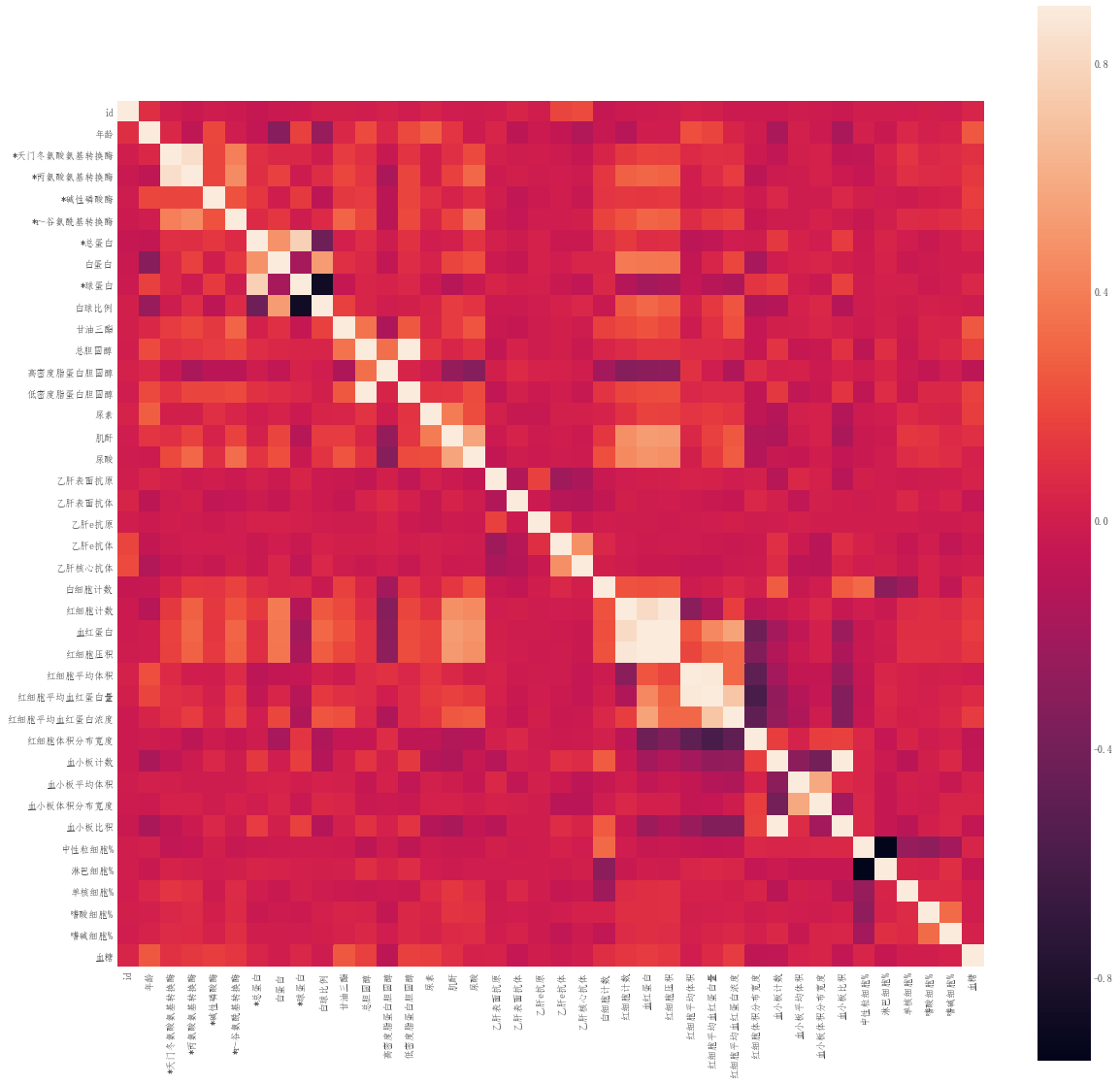


3.2.3 查看各个特征之间的相关性

既然单个特征与血糖之间没有明显的相关性，我们还需要查看两两特征之间是否存在，明显的相关性。如果两个特征之间高度相关，说明其中一个特征是不必要的，可以把它给删除掉，有利于降低输入的维度，减少拟合参数并提高模型的鲁棒性。

通过查看特征的协方差矩阵可以做到这一点。协方差是反映的变量之间的二阶统计特性，如果向量的不同分量之间的相关性很小，则所得的协方差矩阵几乎是一个对角矩阵。对于一些特殊的应用场合，为了使随机向量的长度较小，可以采用主成分分析（PCA）的方法，使变换之后的变量的协方差矩阵完全是一个对角矩阵，之后就可以舍弃一些能量较小的分量了。

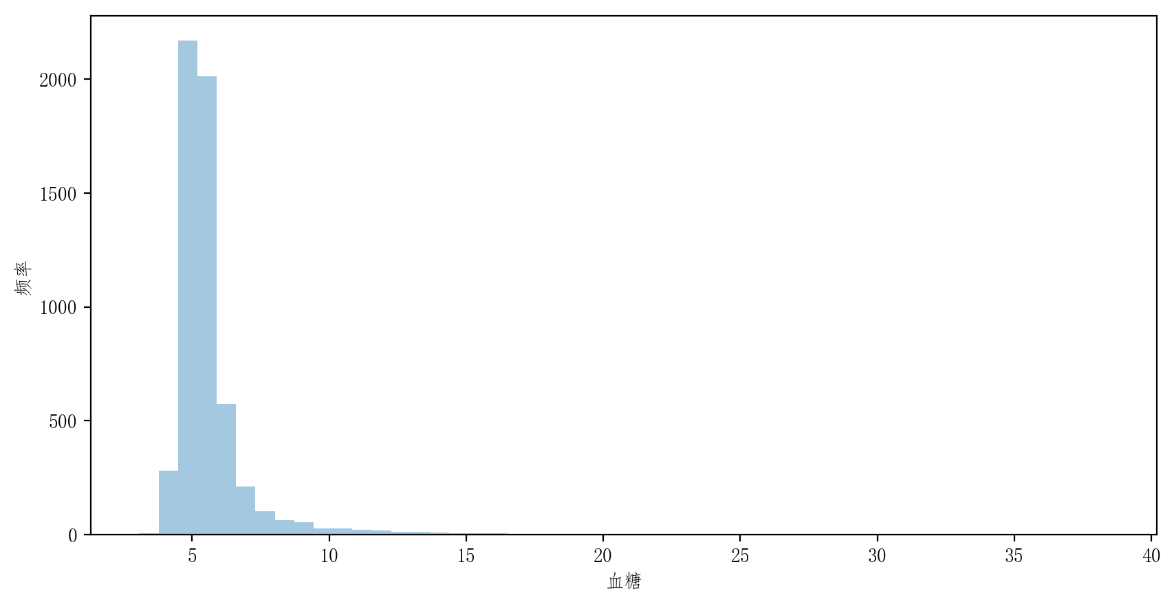
对协方差矩阵进行可视化显示如下图，颜色接近浅色说明正相关程度越高，颜色接近深色说明负相关程度越高，居中的颜色说明相关性不明显。可见协方差矩阵除对角线外的相关系数都非常小，说明两两特征直接不存在比较明显的相关关系，已经是比较独立的了，无需进一步约简。

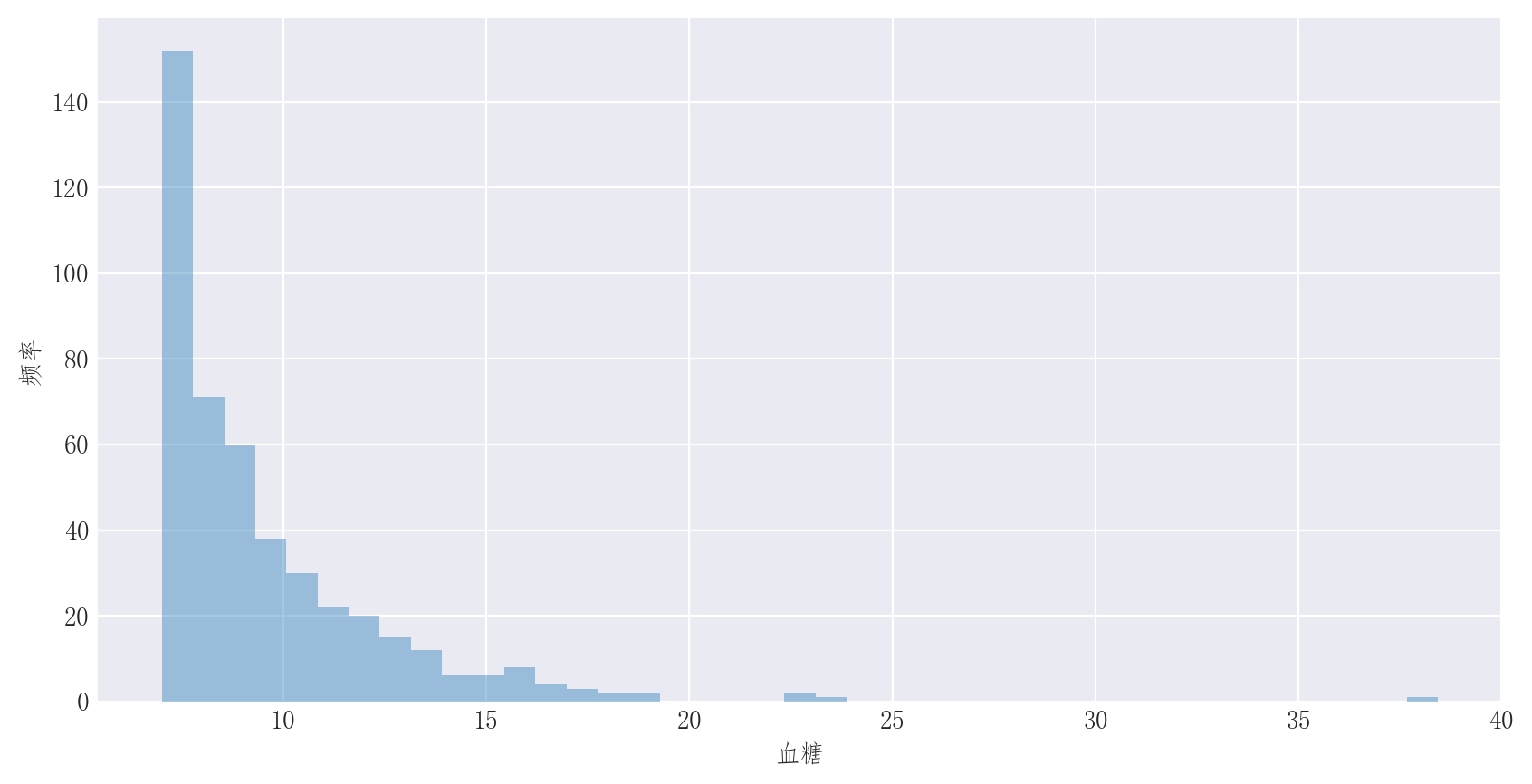


3.2.4 查看血糖分布

因为血糖是要预测的值，因此我们首先得知道训练的样本中的血糖分布大概是怎样的，是均匀分布、正态分布还是呈现其他特征。如果训练集呈现明显的不均衡的情况，我们就需要对其进行过采样或者欠采样等操作使得样本比较均衡，如此才能保证模型的预测比较全面准确。

对训练集中的血糖值进行频率统计并绘制直方图，如下图所示。横轴是血糖值，纵轴是出现的频次，可见血糖高度集中在4-10之间，但也有极少数偏差很大的点。对高血糖部分（7以上）进行放大显示如下图。可见随着高血糖的样本还是少数，需要对其进行过采样处理，尽量保证样本的均衡。



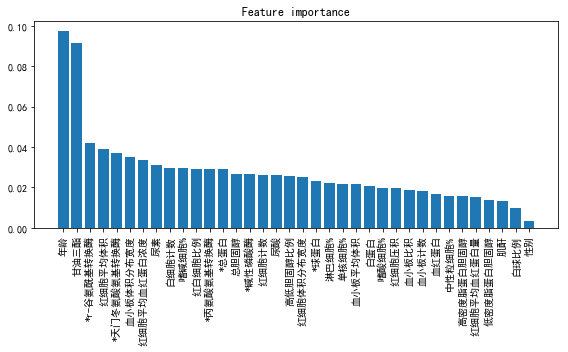


3.2.5 特征选择及重要性排序

作为特征分析的最后一步我们对特征的重要性进行一个排序，将重要性较低的特征抛弃可以起到降低模型复杂度的作用，这也称作特征选择。特征选择在维数很高时是非常有必要的。常用的选择方法有PCA, RFE, 基于模型的特征选择方法等。这样采用基于随机森林模型的特征选择方法。输出的特征重要性排序如下：

根据特征重要性排序可以看出，年龄和甘三油脂对血糖的影响比较明显，这也比较符合常理。年龄不必多说，甘油三酯据查是由于胰岛素的敏感性不够，所以需要更多的胰岛素来维持血糖正常水平，同时这个时候甘油会结合脂肪酸变成甘油三酯。甘油三酯>2.8mmol/L的成年人差不多八成在5-8年后成为糖尿病患者或者糖尿病高危人群。

由于只有30多个特征，因此我们暂时不对特征进行约减，将全部特征都用于学习。



3.3 数据预处理

3.3.1 缺失值处理

一般来说缺失值处理是数据处理的第一步。处理可以分为直接删除和填充两种。对于缺失值较多的特征或样本可以直接删除，如前面看到的乙肝表面抗体等5个特征。对缺失值较少的可以采用填充。填充又可以分为均值、中值、众数、负值填充等几种。经过比较发现，这里采用-999值作为填充效果较好。

3.3.2 类别特征处理

3.3.3 归一化处理

3.3.4 数据集拆分