# 基于决策树算法的心脏病诊断数据集分类

# 背景问题和数据描述

根据WHO统计，世界上每年有近1700万人死于心血管疾病，如何快速、有效地对心脏病进行诊断始终是生命科学领域研究的重点问题之一，随着机器学习技术的兴起，其在医疗领域的应用已经越来越广泛，2018年全球人工智能应用大赛（南京）设置了智能医疗项目，其中就包含了“实时监测心血管疾病的智能算法”。

本数据集包含了来自克利夫兰医学中心的270个病人实例数据，用户可基于此建立心脏病诊断模型。本数据集的数据内容、格式和取值范围如表1和表2所示

表1 数据集格式



表2 数据内容及取值



另外包括对原数据集进行标准化处理和归一化处理的数据。

1. **解决思路和具体方案**

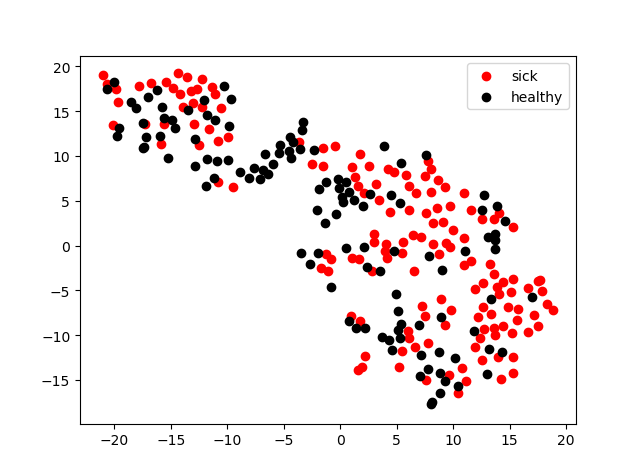
原数据集中包含 269 个样本，除去作为分类标签的患病情况，每个样本可以看作一个长度为 13 的一维向量，样本的维度并不高，我们尝试将数据集的样本进行二维空间的可视化展示。采用了t分布随机近邻嵌入（t-SNE）的方法，将样本数据的维度从4维降至2维后，在二维坐标系中进行散点图展示，结果如图1-3所示。

图1 原始数据的tSNE可视化展示

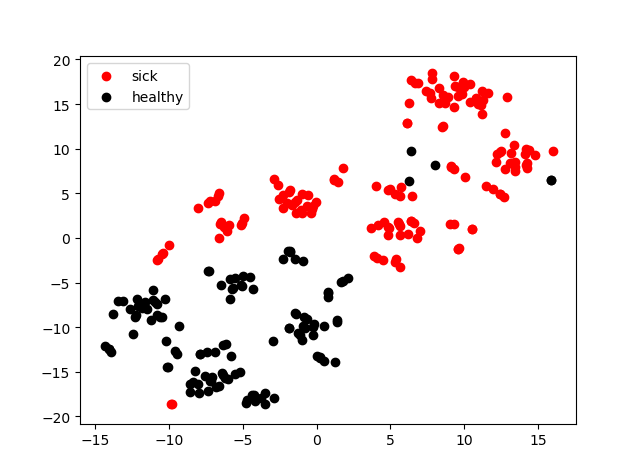


图2 归一化数据的tSNE可视化展示

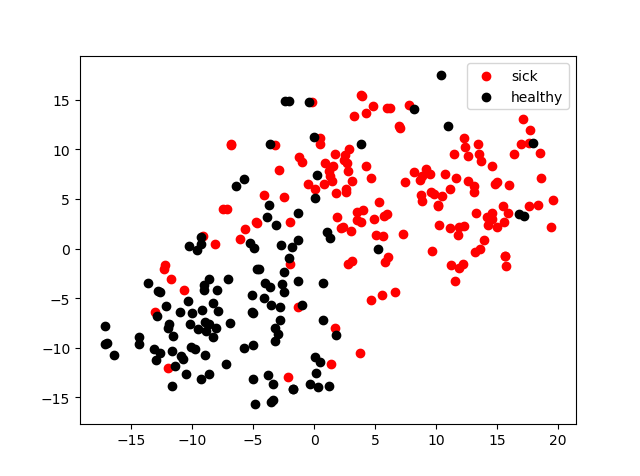


图3 标准化数据的tSNE可视化展示

各图中 sick 和 healthy 所对应的红色和黑色数据点分别为患病和未患病样本。可以看出，数据集的类别数量比较平衡，且对数据集进行标准化和归一化后的区分效果要比原始数据更好。

采用随机方式将数据集打乱后，按4:1的比例划分训练集和测试集，在训练集上使用决策树算法进行分类，然后在测试集上进行效果的评估。

1. **实现步骤和代码**

引入必要的python库：

import pandas as pd

from sklearn.tree import DecisionTreeClassifier

from sklearn import tree

from sklearn.metrics import accuracy\_score

from sklearn.model\_selection import train\_test\_split

import pydotplus

import graphviz

from sklearn.manifold import TSNE

import matplotlib.pyplot as plt

使用pandas读取数据：

data = pd.read\_csv('heart\_disease5597/data.csv').values

y = data[:, -1]

使用tSNE对数据进行降维：

x\_embed = TSNE(n\_components=2).fit\_transform(data)

对数据集进行训练集和测试集的划分：

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x\_embed, y, test\_size=0.2)

使用训练集训练决策树：

clf = DecisionTreeClassifier(random\_state=0, max\_depth=8).fit(x\_train, y\_train)

在测试集上进行预测并计算准确率：

y\_pred = clf.predict(x\_test)

print(accuracy\_score(y\_pred, y\_test))

使用graphviz绘制可视化的决策树图形：

dot\_data = tree.export\_graphviz(clf,out\_file=None,feature\_names=['AT1','AT2'],class\_names=['sick', 'healthy'], filled=True, rounded=True)

graph = pydotplus.graph\_from\_dot\_data(dot\_data)

graph.write\_pdf("result1.pdf")

1. **处理结果和分析**

对原始数据、归一化数据、规范化数据都进行了决策树分类和测试。结果如图4-6和表所示。

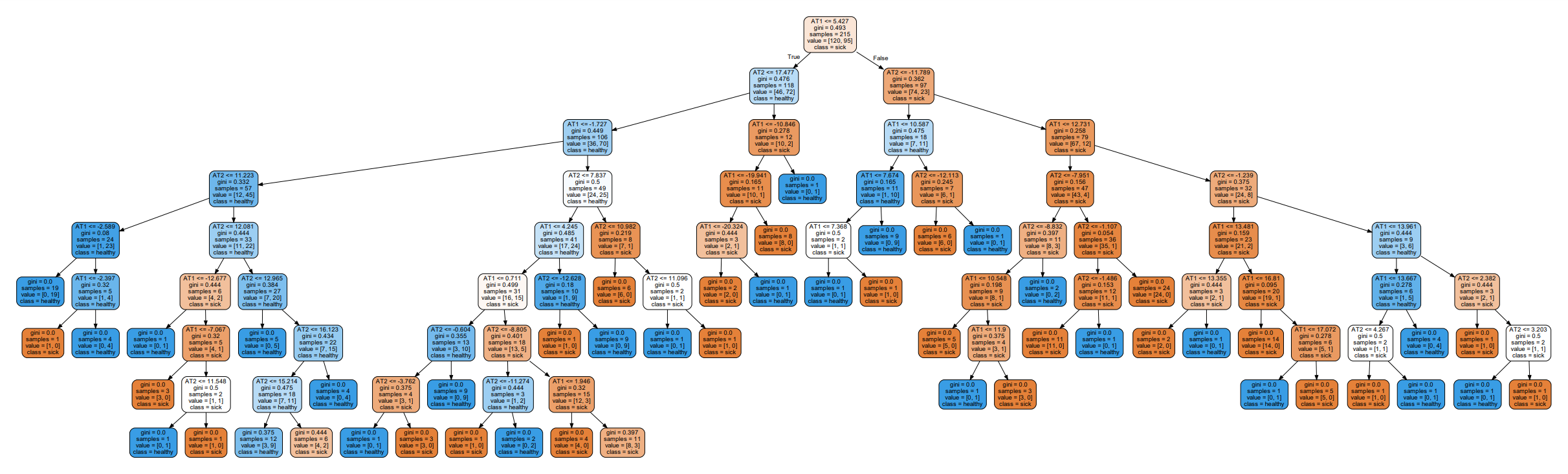


图4 原始数据训练的决策树



图5 归一化数据训练的决策树

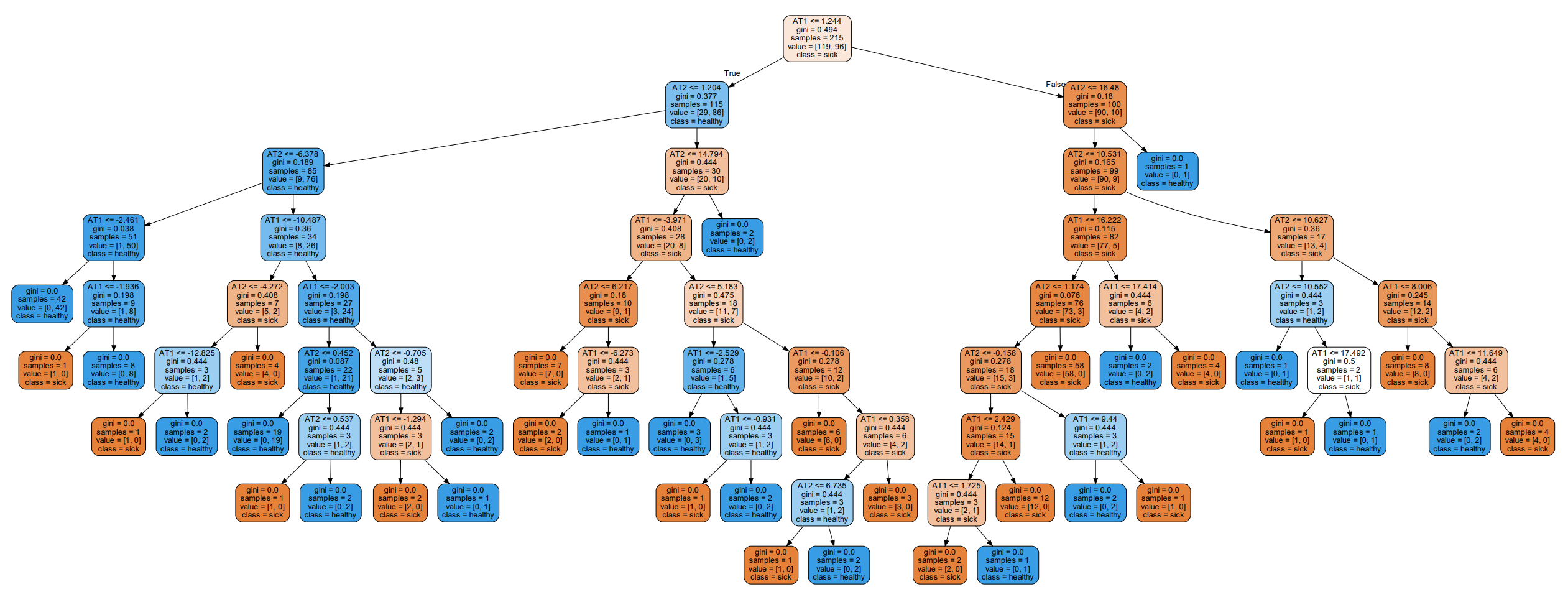


图6 标准化数据训练的决策树

表3 各数据集分类准确率的对比

|  |  |
| --- | --- |
| 数据集 | 准确率 |
| 原始数据 | 0.6111 |
| 归一化数据 | 0.9444 |
| 标准化数据 | 0.7593 |

可以根据上述图表得出结论，对原始数据进行t-SNE降维并进行归一化处理能够得到更好的分类准确率，而且准确率达到了90%以上。

# 个人工作总结

在此次课程设计中，本人对于决策树算法、t-SNE降维算法的原理及实现都有了初步的了解。