选拔论文

题目分析

要点：数据分类

- 19项人类活动（A1-A19），每一项活动4男4女

任务需求：

1. 设计模型提取19种行为特征并进行分类
2. 评估、提高模型泛化能力
3. 克服过拟合，拓展算法的应用范围

数据文件结构

- 19项活动a01-a19
- 8位测试者p1-p8
- 60个时间片段s01-s60
- 5个单元（T,RA,LA,RL,LL），每个单元上有9个传感器（xacc,yacc,zacc加速度计，xgyro,ygyro,zgyro陀螺仪，xmag,ymag,zmag磁力计）
- 5个单元与9个传感器等效于45个不同传感器
- 每个文件含有某一活动的某一个人的某一个5秒信号段内，45个传感器捕捉到的125个数据
- 共有19 * 8 * 60 * 45 * 125 = 5130,0000个数据

题目抽象

- 多维数据（5个维度： 活动类型、参与者、时间片段、传感器、传感器的多组数据）的特征提取及分类
- 数据量庞大，选取模型必须是能处理大量数据并提取有效特征的

思路

运用模型：神经网络/svm

IO：

- 输入：多维数据
- 输出：根据多维数据求得的某个向量
- （类似于，输入5.09,5.03,5.04,4.99，输出5.00+-0.15，根据输出，可以判断5.11属于这个数据范围，而5.40不属于这个数据范围）

第一问：根据部分数据设计出模型，实现“输入输出”（数学原理介绍 + 代码实现）

第二问：使用剩下的数据验证模型，采取某些评价指标（指标的数学原理介绍 + 代码实现）评估，给数据添加噪声，如果还能准确识别，说明模型泛化能力好
和第三问相似的思路：加入噪声验证

第三问：
参考

> 而对于我们选用的SVM来说：
> 在完全线性可分的数据集下，支持向量机没有过拟合问题，因为它的解是唯一的。而在非线性不可分的情况下，虽然SVM的目标函数采
> 应用的SVM模型都是核函数+软间隔的支持向量机，那么，有以下原因导致SVM过拟合：
> 1）选择的核函数过于powerful
> 2）要求的间隔过大，即在软间隔支持向量机中C的参数过大时，表示比较重视间隔，坚持要数据完全分离，当C趋于无穷大时，相当于
> 那么我们可以重点关注一下参数C，引入松弛变量或者往模型中加强正则化。

克服过拟合问题后，使用剩下的数据加入噪声验证

实现

SciKit-Learn库

- 分类模型（传参可优化）
- 评估指标

参考文献和资源

1

2

3

4

5

6

7

8

9

[11]周林寰. 一类支持向量机在线算法及其应用[D].大连理工大学,2021.DOI:10.26991/d.cnki.gdllu.2021.001233.