基于多层分类器模型的人类行为分类

摘要

人类行为理解的一个重要方面是对日常活动的识别和监控。近些年来基于人机交互系统的人体活动识别技术的研究引起了人们的广泛关注。本文将基于可穿戴设备检测的数据建立数学模型，并对 19 种不同的人类行为进行分类。并评估模型的有效性与普适性。

针对问题一：首先，考虑到数据的复杂性，我们将数据整合到了一个 csv 文件中。随后对特征重要性进行了计算，并挑选出了特征权值在 0.02 以上的 20 个特征参与后续分类算法。由于分类对象较多，使用单一的分类算法并不能完美的实现分类。同时，人类活动行为本身就存在一定的相似性，对于相似的行为，我们必须做针对性的研究。所以，我们团队提出了基于广义判别分析的多层分类器模型。利用第一层随机森林分类器对所有人类活动进行分类。对于结果中容易混淆的行为，再使用广义判别分析提取非线性特征，并将其交给第二层 SVM 分类器进行深入分类。最后，我们与 KNN，逻辑回归,lightgbm 等单一分类算法进行了比较，结果显示，我们的多层分类器模型的精确率高达 92%, 远远超过了 KNN 与逻辑回归，lightgbm 的精度。

针对问题二：在第一问中，我们的模型验证只使用了一份验证集，预测精度可能偏高。在此，为了进一步评估模型的泛化能力，我们采用了 K-折交叉验证的方法来训练模型。我们将数据集分为 5 份，分别作为训练集和验证集进行训练。通过绘制混淆矩阵，计算召回率，F1-score 等对模型的泛化能力进行评估。结果显示，我们团队提出的基于广义判别分析的多层分类器泛化能力良好。

针对问题三：为了探究并克服模型的过拟合问题，我们在第二问的基础之上，绘制了基于广义判别分析的多层分类器的学习曲线，通过曲线可以看出，我们的模型具有良好的普适性，并不存在过拟合问题。

关键字：多层分类器　随机森林　广义判别分析　SVM

一、 Introduction

1.1 问题背景

对日常活动的识别和监控是提高对人类行为理解的一个重要方面。如今，在计算机快速发展的基础上，可穿戴式活动识别系统也得到了快速发展升级，改善了许多关键领域的生活质量，如动态监测、家庭康复和跌倒检测。我们可以使用基于惯性传感器的活动识别系统通过个人报警系统实现多种功能，比如远程监测和观察老年人、跌倒检测和分类、医疗诊断和治疗、在家或学校远程监测儿童、康复和物理治疗、生物力学研究、人体工程学、运动科学、芭蕾舞和舞蹈、动画、电影制作、电视、现场娱乐、虚拟现实和电脑游戏。位于身体不同部位的微型惯性传感器和磁力计可以产生大量有关人类活动的数据流，这些数据流可以可以用来更好理解人类行为，并可用于对人类活动进行分类。

1.2 问题提出

鉴于以上背景，本文需要建立数学模型解决以下问题：

1. 设计一组特征和有效的算法，以便根据这些人体传感器的数据对 19 种人类行为进行分类。
2. 在有限的数据集下使模型具有良好的泛化能力。我们需要专门研究和评估这个问题。并设计一个可行的方法来评估模型的泛化能力。
3. 研究并克服过度拟合问题，以便我们的分类算法可以广泛应用于人们的行为分类问题。

1.3 Data Preprocessing/Feature Extraction

本文数据来自于身体不同部位的微型惯性传感器和磁力计对人体活动的测量。共采集了 8 位受试者在进行 19 项不同活动的传感器数据。每个受试者每次活动的总信号持续时间为 5 分钟。传感器单元被校准为以 25 Hz 采样频率采集数据。将 5 分钟的信号分成 5 秒的片段，从而为每个活动获得 480（=60×8）个信号片段。

我们最终得到了 9120 份文本文件，详细记载了不同传感器在相同时间间隔里测得的不同受试者进行不同活动时的数据。我们将这些文本数据整合到一个 csv 文件中，并加上了受试者编号以及活动类型两列。

通过使用基于随机森林的特征重要性排序算法，筛选出重要特征，以便于后续分类。

1.4 Our Model

数据预处理之后，根据筛选出来的特征，我们团队设计了一种可行的算法方案来对19 种人类行为进行分类。由于数据量特别庞大，并且人类活动本身就具有一定的相似性，若使用单一的机器学习算法对 19 种人类行为进行直接分类，很容易造成相似行为的混淆，导致分类精度下降。为解决这一问题，我们团队提出了基于广义判别分析的多层分类器算法。

算法流程如下：

①数据采集与预处理，特征提取。

②第一层分类识别全部活动。

③广义判别分析相似活动特征。

④第二层分类识别相似活动以及双层分类器加权融合。

算法流程框图如下：

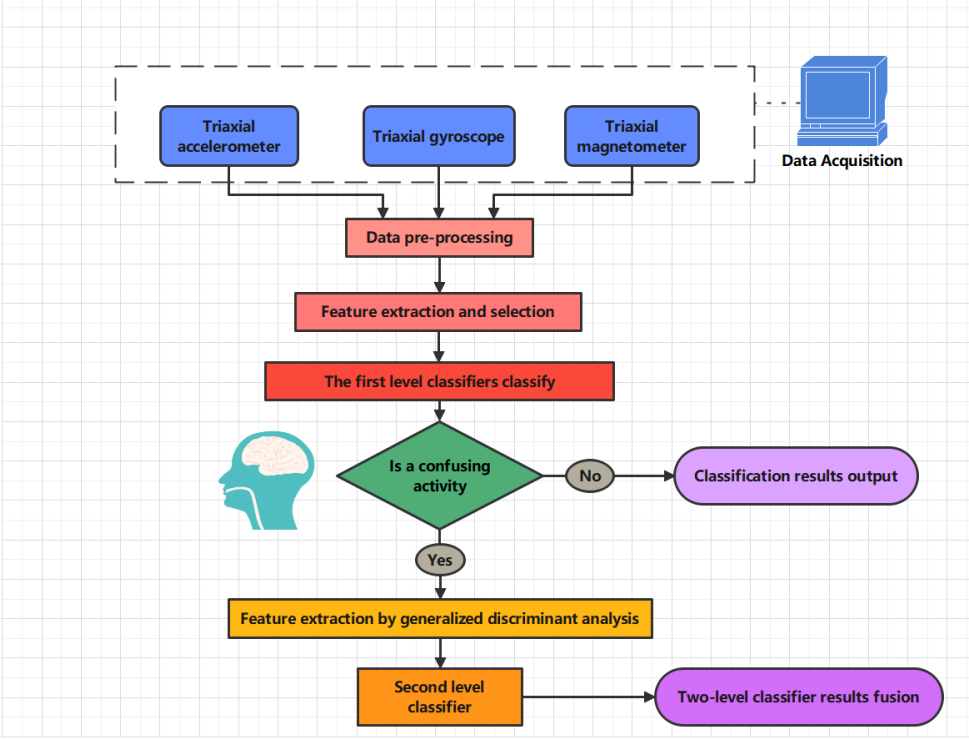


图 1　基于广义判别分析的多层分类器

二、 The Description of the Problem

2.1 问题一分析

由于问题涉及到的数据量极为庞大，特征指标也非常多。所以我们首先对数据进行处理之后，再进行特征提取。考虑到题目要求的分类是一个 19 类的细分任务，因此，直接使用单一的分类算法效果肯定大打折扣。因此，我们提出了基于广义判别分析的多层

分类器算法。第一层分类器我们采用随机森林算法对全部活动进行分类，对于分类结果中容易分类错误的类别，我们进一步使用广义判别分析提取这些类别的非线性特征，随后采用第二层分类器 SVM 深入分类。

2.2 问题二分析

题目要求对模型泛化能力进行评估，根本就是对模型要对新数据有很好的预测能力，我们采用测试集的指标表现评估模型的泛化性能，首先采用 k 折交叉验证的评估方法，以原始数据作为测试集对其进行分类预测，通过评价指标量化模型在不同方面的表现，从而得出使用模型 1 的结果和实际的偏差，评估模型泛化能力。

2.3 问题三分析

针对所建模型可能存在的过拟合的情况，我们通过绘制模型学习曲线判断模型是否存在过拟合的问题。

三、 Models

3.1 Basic Model

3.1.1 Terms, Definitions and Symbols

为了方便后面的算法，将 19 种人类行为 Sitting、Standing、Lying on back、Lying on right side、Ascending stairs、Descending stairs 、Standing in an elevator still、Moving around in an elevator、Walking in a parking lot、Walking on a treadmill with a speed of 4 km/h in flat position and 15 deg、inclined positions、Walking on a treadmill with a speed of 4 km/h in 15 deg inclined positions、Running on a treadmill with a speed of 8 km/h 、Exercising on a stepper、Exercising on a cross trainer、Cycling on an exercise bike in horizontal position、Cycling on an exercise bike in vertical position 、Rowing 、Jumping 、Playing basketball 按顺序简称为 A1、A2、A3、... 等

3.1.2 Assumptions

假设 1：假设人体传感器的数据皆正确无误；
假设 2：
假设 3：

3.2 特征提取

3.2.1 基于随机森林的特征重要性排序算法

随机森林是利用多个决策树进行训练并且预测的一种集成学习算法。为衡量样本中各项数据特征与阿尔兹海默症诊断间的相关性强弱，对共计 48 项数据特征分别按照式 (1) 计算基于随机森林的特征重要性指数（PIM），并依据 PIM 值大小对数据特征重要性开展排序，具体流程如下:

①构造 M 棵决策树;

②当前决策树 $k_{tree}=1$ 时，得到对应袋外数据 OOB_k;

③计算当前决策树对 OOB_k 的预测误差 $errOOB_k$;

④将 OOB_k 中第 i 种数据特征的随机扰动设置为 OOB_k^i;

⑤对于每一颗决策树，$k_{tree}=2,\cdots M$, 重复步骤②到 ④;

⑥根据式 (1) 计算数据特征的重要性

$$PIM=\sum_{i}^{M}(errOOB_k^i-errOOB_k) \tag{1}$$

式中:M 为构造的决策树数量，$errOOB_k^i$ 和 $errOOB_k$ 分别表示对第 i 种统计参量添加扰动后的袋外数据和未添加扰动的袋外数据在 k_{tree} 棵决策树情况下的预测误差。

3.2.2 重要特征提取

通过使用该算法对数据集中的 45 个特征进行分析，可以得到不同特征对于 19 种分类结果的不同重要程度。我们选取重要性大于 0.02 的特征作为重要特征，参与后续基于广义判别分析的多层分类器分类。而对于重要性较弱的特征，我们予以筛除，避免这些特征可能携带的干扰信息对我们分类的准确程度造成影响。

对重要特征的权值绘制的柱状图如下:

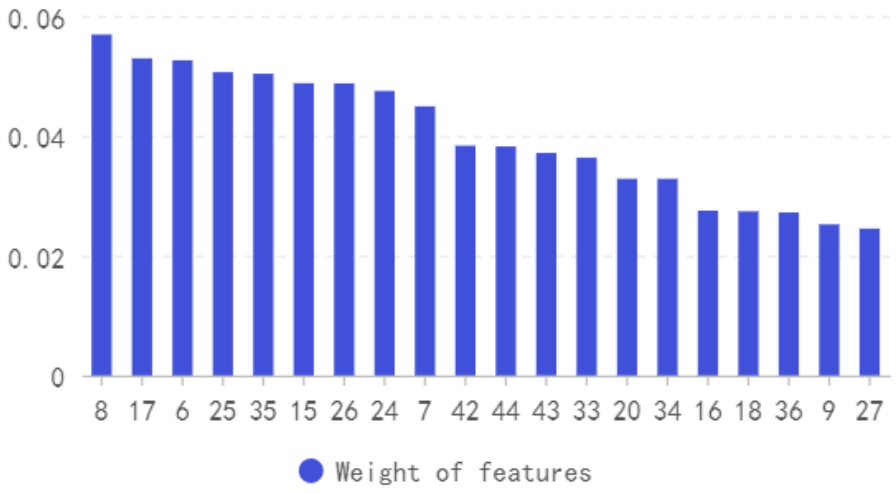


图 2　重要特征权值图

3.3 问题一模型建立与求解

3.3.1 基于随机森林的人类行为第一层分类器

在上述步骤已将能够分类人类行为的主要特征提取出来了，考虑到这些数据之间的关系以及用于训练的样本是离散的，且数据量巨大，因此考虑采用随机森林的算法进行网络训练，其一般的算法流程如图图 3所示。为了初步对多类活动进行识别，第 1 层分类器选用在监督学习中性能表现优异的随机森林分类算法. 随机森林 (RF,Random forest) 作为 Bagging 的进一步扩展与优化，在以决策树为基学习器构建 Bagging 集成基础上引入了随机属性选择的方法。随机森林是由一系列未剪枝的决策树（这里指分类回归树）$\{h(\boldsymbol{x}, \boldsymbol{\Theta}_k)\}$ 组合成的分类器，其中 Θ_k 为独立同分布随机向量，且每棵树对输入向量 X 所属的最受欢迎类投一票。它改善了单一决策树的不足，且不容易造成过拟合的现象。具体流程及流程图如下：

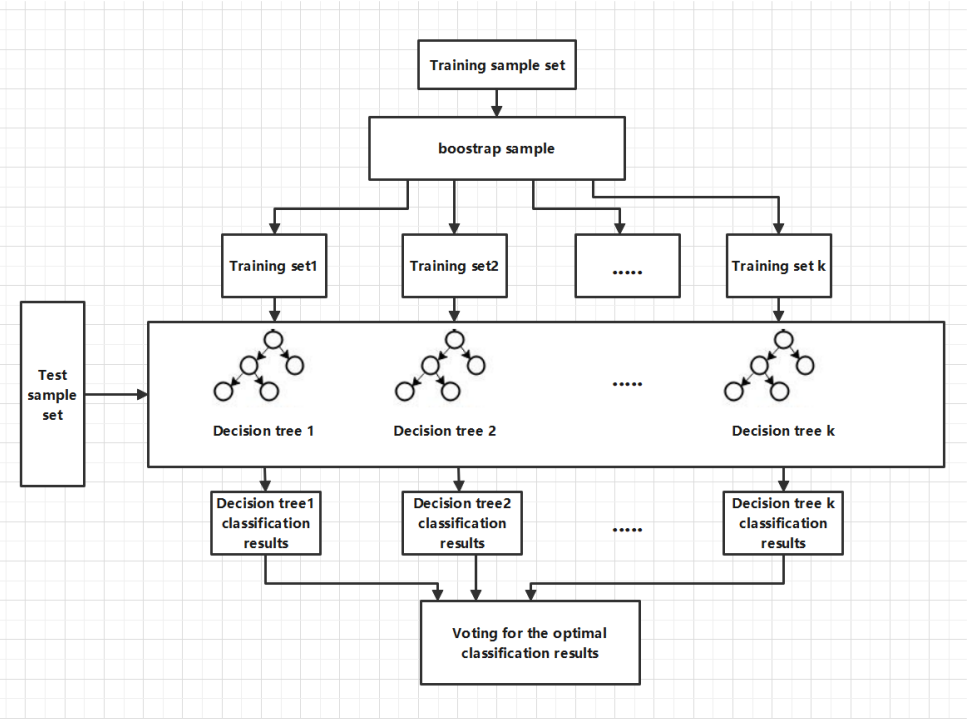


图 3　随机森林算法流程图

Step1: 从经过特征选取的人体传感器数据 D 中通过 Bootdstrap 采样选取 k 个子训练样本 $(D1, D2, ..., D_k)$，建立 k 棵决策树。

Step2: 在分类树的每个节点上随机地从 n 个指标中选取 m 个，按照节点不纯度最小原则从 m 个候选指标中选择最优特征对节点进行分类生长，让决策树充分生长直到每个叶子节点的不纯度（即 Gini 指数）达到最小，同时不对决策树进行剪枝。

Step3: 重复步骤 Step2 遍历预建的 k 棵决策树，由 k 棵决策树形成随机森林。

Step4: 依据生长好的 k 棵决策树来预测新的未知样本。待测样本的分类结果根据 k 棵决策树投票的多数投票结果来决定。其分类公式为:

$$f(x_t) = \text{ majority vote } \{h_i(x)\} \quad (i = 1, 2, \cdots, k)(3) \tag{2}$$

式中：majority vote 为多数投票结果。

先用 matlab 使用随机森林模型，以决策数的个数为自变量，误差为因变量做出决策数个数与误差之间的关系图，如下图所示

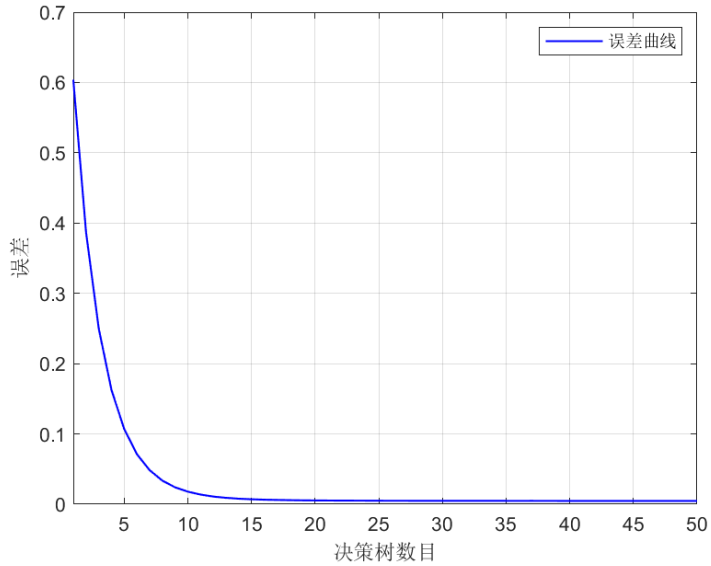
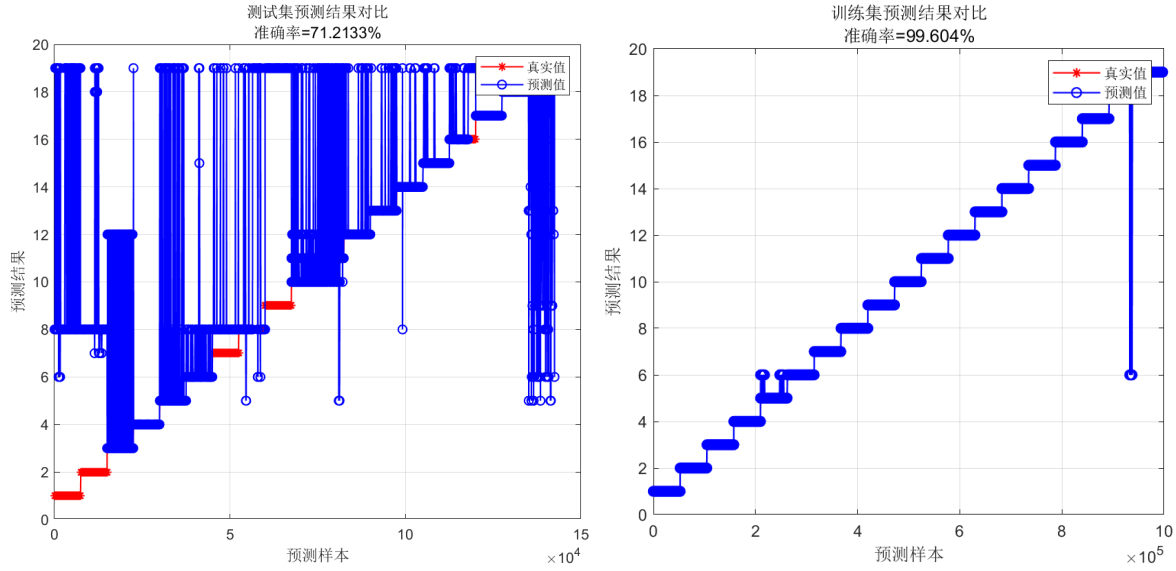


图 4　随机森林模型中决策数个数与误差关系图

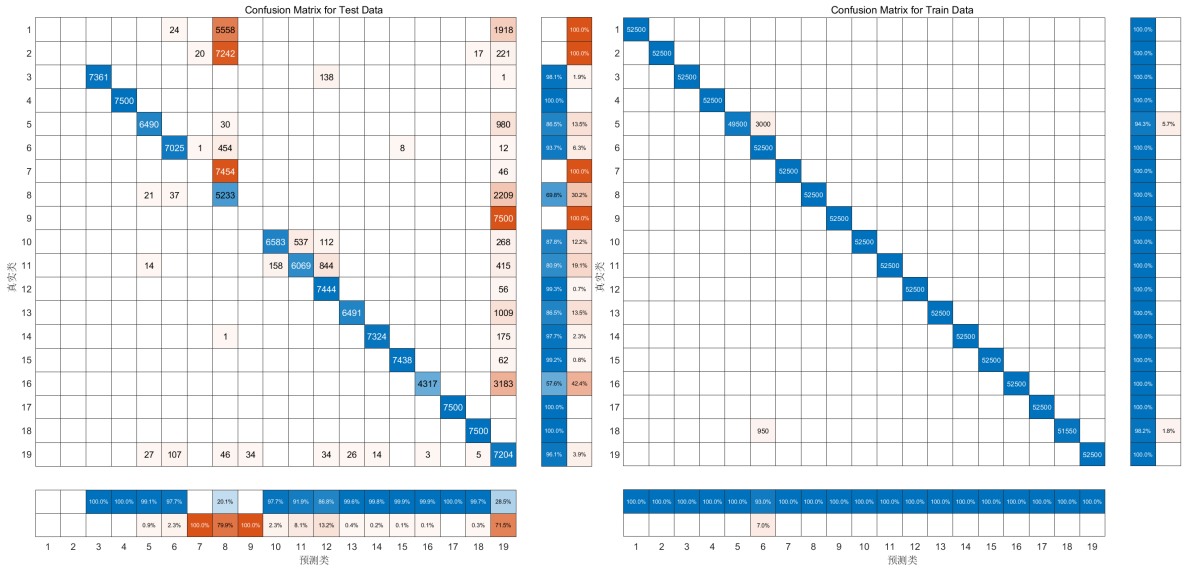
由上图我们可知在该数据集中，基于电脑性能与模型准确度，以及模型的可靠性分析，确定决策树的个数为 50 颗，为了更好地分析随机森林模型的分类效果，我们设立了测试集检验我们的训练模型，使用 MATLAB 以编号 p1-p7 为训练集，p8 为测试集进行随机森林训练与测试，得到如下图所示结果：



(a) 随机森林模型识别测试集结果对比图　　**(b) 随机森林模型识别训练集结果对比图**

图 5　随机森林模型识别结果对比图

为了更好的分析上述随机森林识别结果，使用 MATLAB 做出上述两个结果的混淆矩阵图，如下图所示：



(a) 测试集随机森林训练结果混淆矩阵图　　**(b) 训练集随机森林训练结果混淆矩阵图**

图 6　测试集与训练集随机森林训练结果混淆矩阵图

观察上面四个图，本文认为动作识别结果与真实结果不一致主要原因在于这些动作都具有相似的特征，在算法识别过程中极易混淆，例如上下楼梯、站立和坐下，除了上述这种情况的动作，其他动作的预测率都接近 100%，也就是说，这些动作可以在这一层分类模型中被识别并分类到真实的动作中，而剩下未被正常识别的动作主要分为两大块，像 A1、A2、A7 都被模型分类成 A8，A9、A16 都也同样被模型分类成 A19。为方便接下来的细分类模型，将这些混淆的动作分为两大类如所示

表 1　混淆动作分类表

	易混淆动作
混淆 I 类	A1、A2、A7、A8
混淆 II 类	A9、A16、A18

3.3.2 基于广义判别分析的特征映射

为了解决上述 2 类混淆动作相似的问题，运用广义判别分析进一步提取非线性特征，广义判别分析是对线性判别分析的非线性扩展，通过非线性映射，将输入特征映射到更高维的特征空间，从而在高维特征空间进行 Fisher 判别分析。

令 $\emptyset$ 为输入特征空间到高维特征空间 F 的非线性映射:

$$\emptyset : R^d \to F, x \to x^{\emptyset} \tag{3}$$

定义类间散度矩阵 $S_B\emptyset$ 和总类内散度矩阵 $S_W\emptyset$:

$$S_W^{\emptyset} = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i}\left[\emptyset\left(x_i^j - m_i\right)\right]\left[\emptyset\left(x_i^j - m_i\right)\right]^T$$
$$S_B' = \frac{1}{N}\sum_{i=1}^{c} N_i\left(m_i - m\right)\left(m_i - m\right)^r \tag{4}$$

其中 N 代表一组高维特征向量维度，C 代表类别个数，mi 代表第 i 组高维特征的均值，$m_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\Phi\left(x_j\right)$; m 代表所有类别高维特征的均值，$m = \frac{1}{N_i}\sum_{i=1}^{C}\sum_{j=1}^{N_i}\Phi\left(x_i^j\right)$.

定义判别准则 J(w) :

$$J(w) = \max\left(\frac{\left(w^{\emptyset}\right)^T S_B^{\emptyset} w^{\emptyset}}{\left(w^{\emptyset}\right) S_W^{\emptyset} w^{\emptyset}}\right) \tag{5}$$

求解投影向量 w_{opt} 就是特征值问题的解.

$$\lambda S_W^{\emptyset} w^{\emptyset} = S_B^{\emptyset} w^{\emptyset} \tag{6}$$

由于特征空间 F 维度较高无法直接求解，因此引用内积核函数 RBF 作为映射函数 k:

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{7}$$

其中 x,y 代表对应特征值,σ 为常数, 代表非线性化程度.

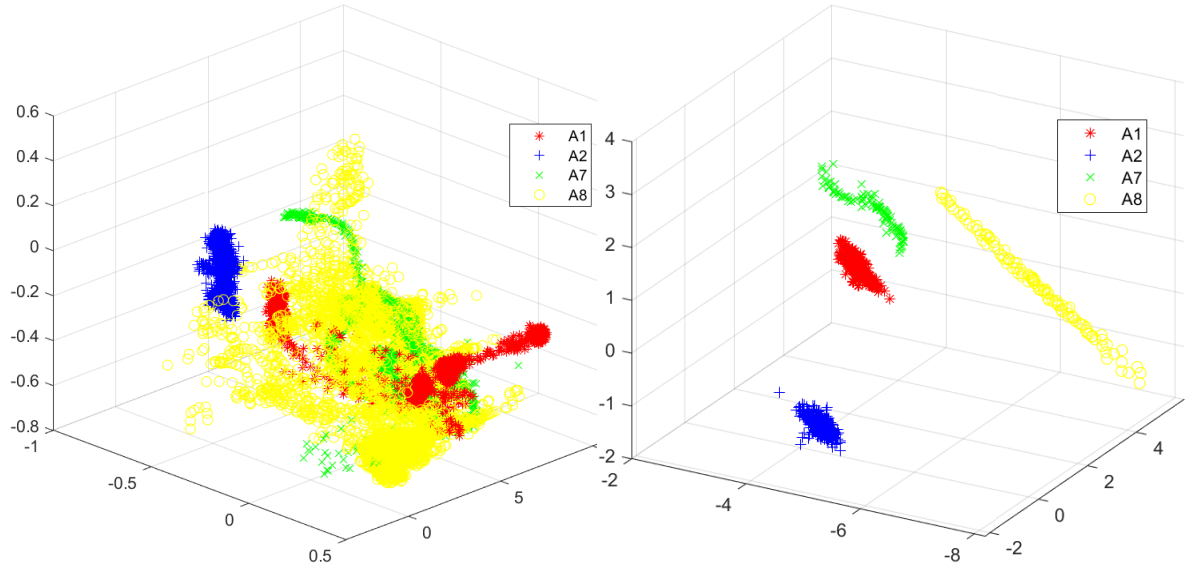
由再生核理论可知，高维空间任一解可以被表示为该空间训练样本的线性组合:

$$w^{\emptyset} = \sum_{i=1}^{c}\sum_{i=1}^{N} a_{ij}\emptyset\left(x_{ij}\right) \tag{8}$$

则样本在最佳投影方向上的投影为:

$$w^{\emptyset}\emptyset(x) = \sum_{i=1}^{c}\sum_{i=1}^{N} a_{ij}k\left(x_{ij}, x\right) \tag{9}$$

以混淆 I 类为例，先对数据集中的指标提取三个相关性较大的指标：RA_ygyro、LA_yacc、RA_xmag 作为做 X 轴、Y 轴以及 Z 轴，利用 MATLAB 得到散点图，如图图 11c所示，再利用 SAS 对这三个指标进行广义判别分析得到如图图 11b所示。



(a) 原始特征散点图　　(b) 广义判别分析特征散点图

图 7　广义判别分析特征前后散点图

3.3.3 基于 SVM 的人类行为第二层分类器

为了进一步细分混淆动作，将混淆动作细分为具体的某个动作，本文引入 SVM 向量机作为划分混淆动作的细分类模型，SVM 分类器原理就是取超平面，令不同类别间的特征距离最大化从而实现分类效果。如图所示，分类间隔宽度越宽 (即最大化)，训练集的局部干扰所引起的影响越低。因此可以认为最后一种分类方式的泛化性能和通用性是最佳的。SVM 的模型可以表述为:

$$y = \text{sign}\left(w^T x + b\right) \tag{10}$$

式中,x 为特征向量,w 为权重向量,y 为标记向量，sign(y) 则是符号函数。

当 y=1 时，样本为正样本；当 y=-1 时，样本为负样本，即

$$\begin{cases} w^T x + b > 0, y = 1 \\ w^T x + b \leqslant 0, y = -1 \end{cases} \tag{11}$$

如图 (2) 所示,SVM 通常通过令分类间隔最大化来求得最优分类超平面。假定训练集输入为 x(i) 向量集合，输出为 y(i) 向量集合，分类间隔则是全集合样本到该超平面最小距离的两倍，即式中，m为样本个数。

$$\gamma = \min_{i=1,\cdots,m} 2y^{(i)} \left(\frac{w^T x^{(i)} + b}{\|w\|} \right) \tag{12}$$

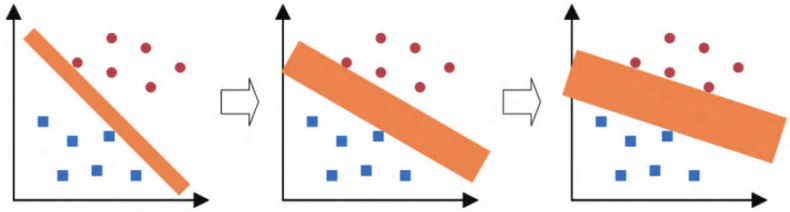


图 8　向量机分类流程图

数学上会将所有符合式 (4) 要求的样本点（即样本点到分类超平面的欧氏距离最小）定义为支持向量，那么该样本集必满足以下两种情况：若样本为正，则 $w^T x^{(i)} + b = 1$ 若样本为负，则 $w^T x^{(i)} + b = 1$，如图所示。

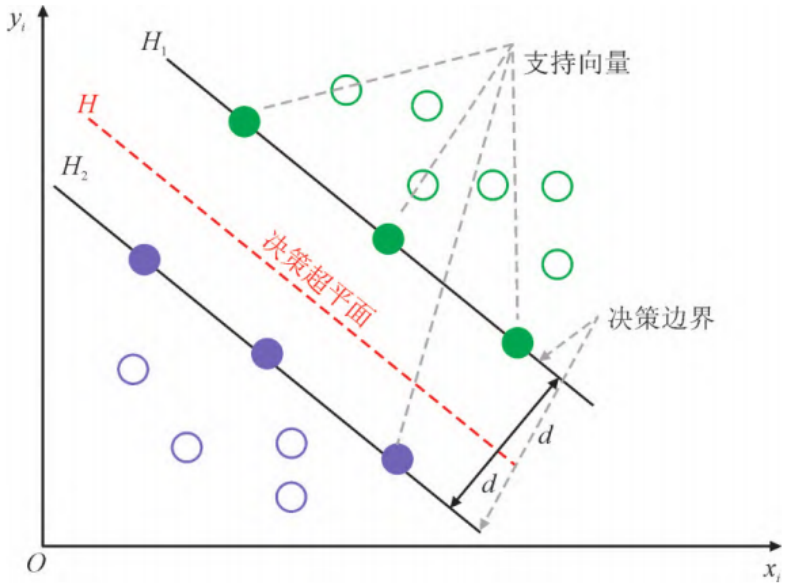


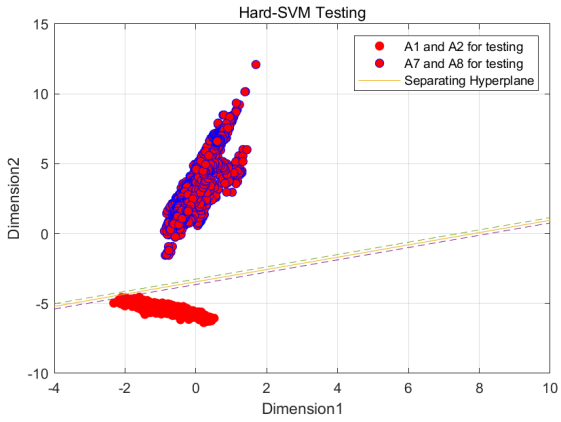
图 9　向量机分类原理图

因此, 样本集中的特征样本在判别方程乘以相应系数时应满足：

$$y^{(i)}\left(w^T x^{(i)} + b\right) \geqslant 1 \tag{13}$$

本文使用 MATLAB 对上述模型进行细分类处理，将已经进行广义判别分析的指标作为原始数据输入到 SVM 中，以混淆 I 类为例，因为 A1、A2 联系较为紧密，A7、A8 的联系也较为紧密，所以先将混淆 I 类分为 A1、A2 以及 A7、A8 两个大类，再进行第二次细分，就可以通过两层 SVM 向量机将混淆 I 类细分成较为 A1、A2、A7 和 A8 这四类活动。

(a) SVM 初步细分类训练集结果图

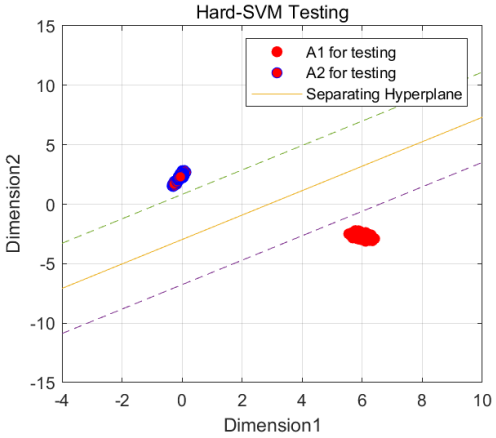


(b) SVM 初步细分类测试集结果图

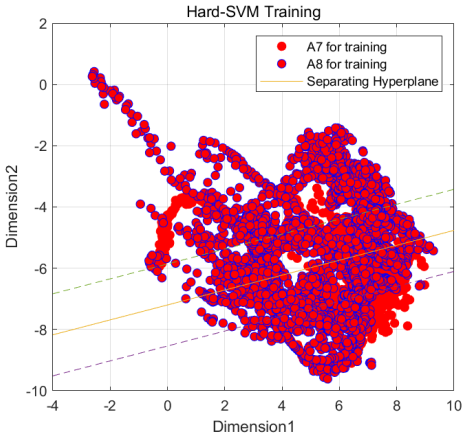
图 10　SVM 初步细分混淆 I 类结果图

通过上述步骤，已经将混淆 I 类的数据通过 SVM 向量机分类成 A1、A2 和 A7、A8 两个大类，为了更细致地分类，本文再对这两个大类进行一次细的分类，将上述两个大类具体分成具体的活动类。如图所示：

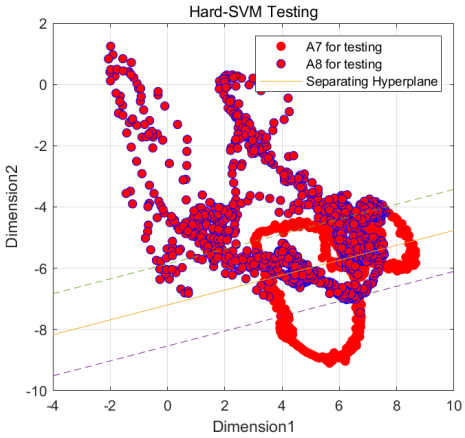
(a) SVM 细分类 A1、A2 训练集结果图



(b) SVM 细分类 A1、A2 测试集结果图



(c) SVM 细分类 A7、A8 训练集结果图

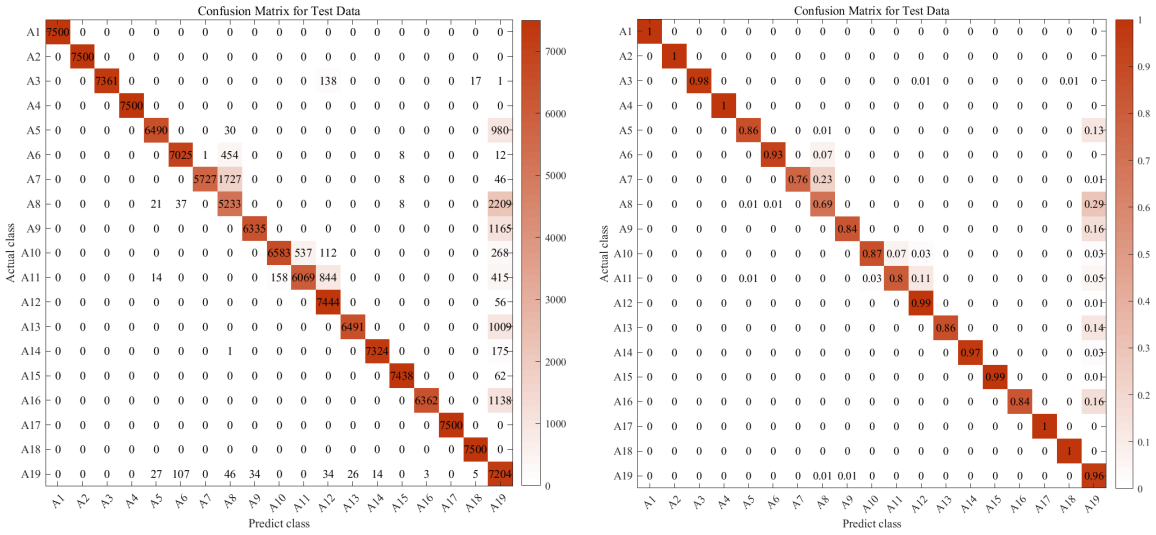


(d) SVM 细分类 A7、A8 测试集结果图

图 11　SVM 细分类混淆 I 类结果图

通过图 11d的相关步骤，我们就能够将混淆 I 类的所有数据通过上述 SVM 向量机细致分类为具体的活动类，虽然 SVM 向量机细分类 A7、A8 的效果不显著，但比初次随机森林的分类效果已经好很多了。

本文同样也对混淆 II 类进行了相同的操作，输入到了第二层分类向量机得到了 4 种相似动作的识别概率，与第 1 层分类器识别概率进行加权平均得到 4 种相似活动的最后识别结果。混淆矩阵如图 12b所示，可以看出原来的易混淆的动作都提升了不少，总体正确率从 71% 提升到了 91%。



(a) 双层分类器修正结果混淆矩阵图　　**(b) 双层分类器修正结果准确率混淆矩阵图**

图 12　双层分类器混淆矩阵图

本文还对其他分类机器学习模型进行了训练如表，与双层分类器模型进行对比，事实证明，这种双层分类器能显著提高分类的准确率。

表 2　各机器学习模型分类准确率

Classification algorithm	Logistic regression	Lightgbm	KNN	double-layer classifier
Accuracy	52%	62%	65%	91%

3.4 问题二的模型建立与求解

题目要求对模型泛化能力进行评估，根本就是对模型要对新数据有很好的预测能力，我们采用测试集的指标表现评估模型的泛化性能，首先采用 k 折交叉验证的评估方法，以原始数据作为测试集对其进行分类预测，通过评价指标量化模型在不同方面的表现，从而得出使用模型 1 的结果和实际的偏差，评估模型泛化能力。

3.4.1 泛化能力

指模型对未知数据的预测能力。从理论上对泛化能力进行分析。如果学到的模型是 $\hat{f}$ 那么用这个模型对未知数据测得的误差即为泛化误差（generalization error）：

$$R_{\text{exp}} = E_P[L(Y, \hat{f}(X))] = \int_{x,y} L(y, \hat{f}(x))P(x,y)dxdy \tag{14}$$

泛化误差也就是所学习到模型的期望风险。

3.4.2 k 折交叉验证

交叉验证是用来验证分类器性能的一种统计分析方法。基本思想是将原始数据进行分组，一部分作为训练集，另一部分作为验证集，首先用训练集对分类器进行训练，再利用验证集来测试训练得到的模型，以此作为评价分类器的性能指标。K 折交叉验证 (KCV) 将原始数据分成 K 组，不重复地抽取 1 个子集作为一次验证集，将其余的 K-1 组子集数据组合在一起作为训练集. 如图所示。

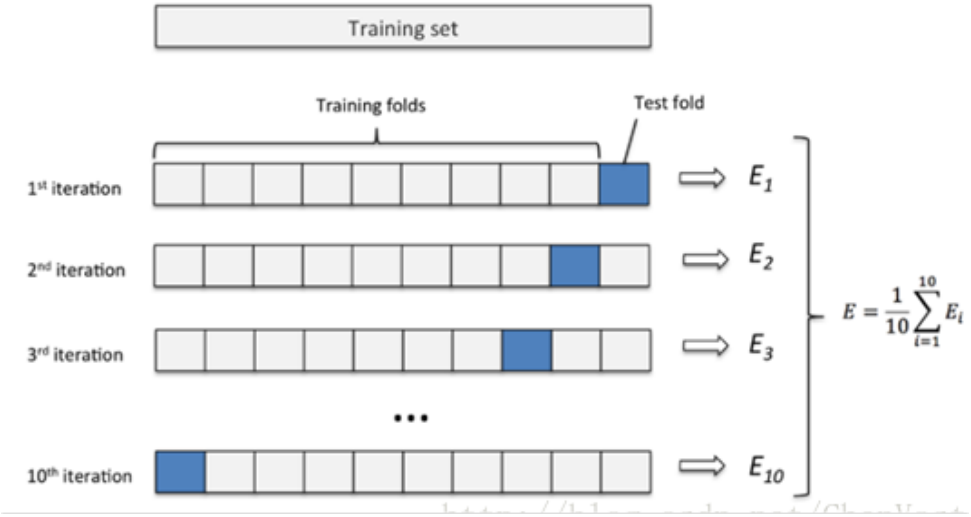


图 13　k 折交叉验证原理

通过分组训练会得到 K 个模型，用这 K 个模型验证集准确率平均数作为 K 折交叉验证分类器的性能指标。K 折交叉验证能避免过学习和欠学习状态的发生，最后得到的结果也比较具有说服性。

3.4.3 性能度量指标的选取

针对建立的模型本质为多分类模型，我们选择混淆矩阵直观感受模型分类精确度，在混淆矩阵当中得到的更高级分类指标 Accuracy 和 F1-Score 作为判断分类模型总体的标准。并通过绘制 ROC 曲线，弥补在实际的数据集中经常会出现类不平衡的问题，得到模型分类正确的概率值。

3.4.4 混淆矩阵

在机器学习领域，混淆矩阵（Confusion Matrix），又称为可能性矩阵或错误矩阵。混淆矩阵是可视化工具，特别用于监督学习，在无监督学习一般叫做匹配矩阵。在图像精度评价中，主要用于比较分类结果和实际测得值，可以把分类结果的精度显示在一个混淆矩阵里面。

混淆矩阵的结构一般如下图表示的方法。

Predict / Real	0	1
0	TN	FP
1	FN	TP

图 14　混淆矩阵结构

混淆矩阵要表达的含义：

混淆矩阵的每一列代表了预测类别，每一列的总数表示预测为该类别的数据的数目；

每一行代表了数据的真实归属类别，每一行的数据总数表示该类别的数据实例的数目；每一列中的数值表示真实数据被预测为该类的数目。

True Positive（TP）：真正类。样本的真实类别是正类，并且模型识别的结果也是正类。

False Negative（FN）：假负类。样本的真实类别是正类，但是模型将其识别为负类。

False Positive（FP）：假正类。样本的真实类别是负类，但是模型将其识别为正类。

True Negative（TN）：真负类。样本的真实类别是负类，并且模型将其识别为负类。

混淆矩阵是对分类问题的预测结果的总结。使用计数值汇总正确和不正确预测的数量，并按每个类进行细分，这是混淆矩阵的关键所在。混淆矩阵显示了分类模型的在进行预测时会对哪一部分产生混淆。它不仅可以了解分类模型所犯的错误，更重要的是可以了解哪些错误类型正在发生。正是这种对结果的分解克服了仅使用分类准确率所带来的局限性。

经过交叉验证后，输出的混淆矩阵如图所示：

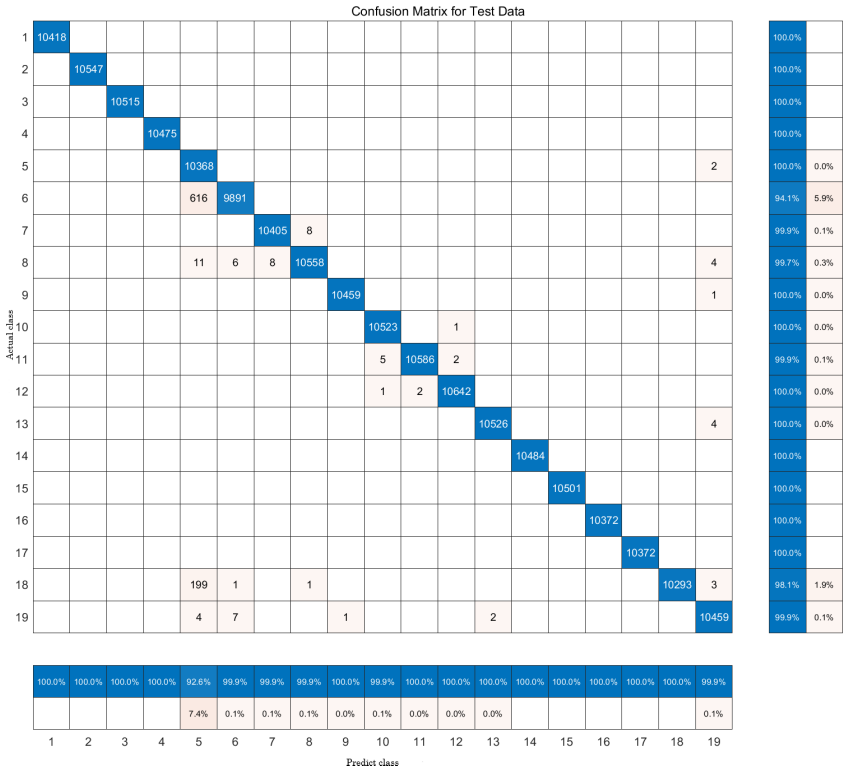


图 15　混淆矩阵图

由混淆矩阵可以看出模型经过交叉验证后精准度高达 99%，证明模型的分类预测的能力很好。同时我们通过混淆矩阵计算出以下评价指标，以便更好评估模型的泛化能力。

3.4.5 准确率和 F1-score

accuracy：

精确率是最常用的分类性能指标。可以用来表示模型的精度，即模型识别正确的个数/样本的总个数。一般情况下，模型的精度越高，说明模型的效果越好。

$$Accuracy = (TP + TN)/(TP + FN + FP + TN) \tag{15}$$

F1-score:

召回率和精确率之间往往存在此消彼长的关系，当模型能找出更多的正样本时，往往也会导致将更多的负样本分类为正样本，即 recall 高时，precision 往往较低，而 precision 高时，recall 往往较低。为了在这两个指标之间取得平衡，我们选取 F1 指标，它是上述两者的调和平均数。它是精确率和召回率的调和平均数，最大为 1，最小为 0。

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{16}$$

模型经过交叉验证后的精确度和 F1-score 如图所示：

Evaluation indicators	accuracy	F1-score
double-layer classifier	99.3%	97.2%

由表可以看出模型精确度和 F1-score 都超过 97%, 说明模型在验证集上表现优异，证明了双层分类模型具有很好的泛化能力。

3.4.6 ROC 曲线

ROC 曲线是根据一系列不同的二分类方式（分界值或决定阈），以真阳性率（灵敏度）为纵坐标，假阳性率（1-特异度）为横坐标绘制的曲线。传统的诊断试验评价方法有一个共同的特点，必须将试验结果分为两类，再进行统计分析。ROC 曲线的评价方法与传统的评价方法不同，无须此限制，而是根据实际情况，允许有中间状态，可以把试验结果划分为多个有序分类。因此，ROC 曲线评价方法适用的范围更为广泛。

ROC 曲线将灵敏度与特异性以图示方法结合在一起，可准确反映某分析方法特异性和敏感性的关系，是试验准确性的综合代表。ROC 曲线不固定分类界值，允许中间状态存在，利于使用者结合专业知识，权衡漏诊与误诊的影响，选择一更佳截断点作为诊断参考值。提供不同试验之间在共同标尺下的直观的比较，ROC 曲线越凸越近左上角表明其诊断价值越大，利于不同指标间的比较。曲线下面积可评价诊断准确性。

模型 ROC 曲线如图所示：

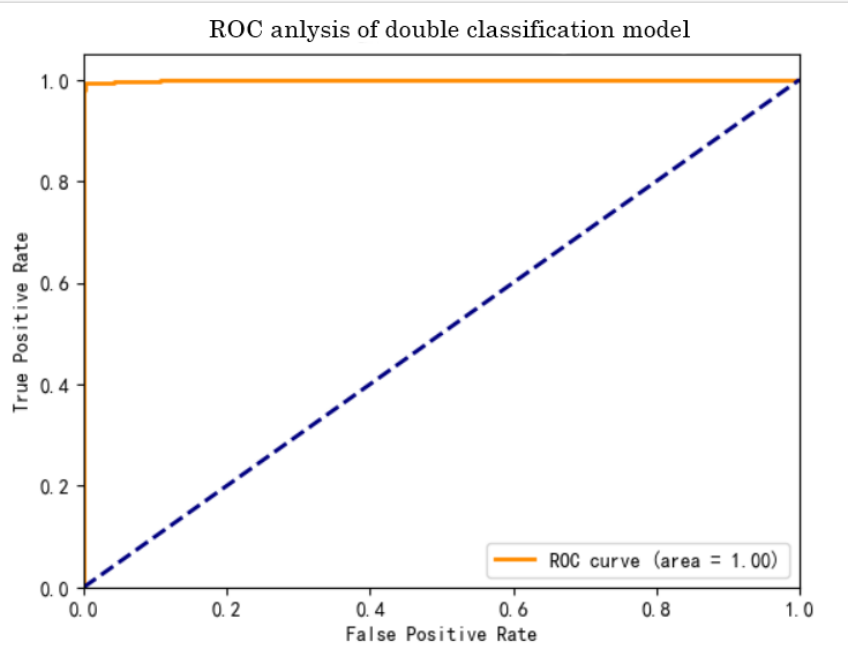


图 16　ROC 曲线图

由图中 roc 曲线变化趋势，我们可以判断模型输出准确概率接近 99%. 具有很好的泛化能力。

3.5 问题三的模型建立与求解

3.5.1 过度拟合的研究

过拟合（overfitting）是指在模型参数拟合过程中的问题，由于训练数据包含抽样误差，训练时，复杂的模型将抽样误差也考虑在内，将抽样误差也进行了很好的拟合。

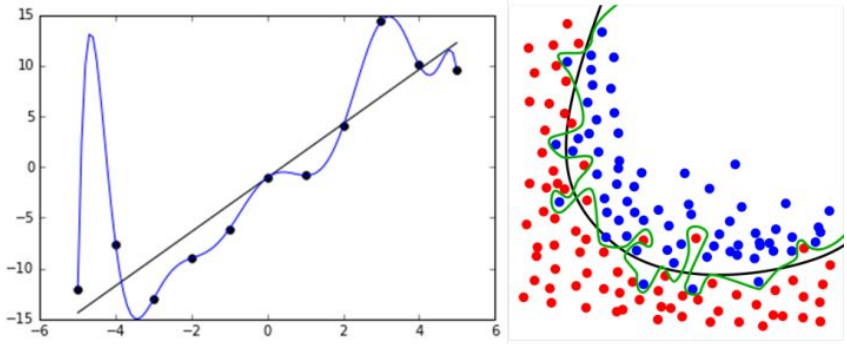


图 17　过拟合示意图

具体表现就是最终模型在训练集上效果好；在测试集上效果差。模型泛化能力弱。

为什么要解决过拟合现象？这是因为我们拟合的模型一般是用来预测未知的结果（不在训练集内），过拟合虽然在训练集上效果好，但是在实际使用时（测试集）效果差。同时，在很多问题上，我们无法穷尽所有状态，不可能将所有情况都包含在训练集上。所以，必须要解决过拟合问题。

为什么在机器学习中比较常见？这是因为机器学习算法为了满足尽可能复杂的任务，其模型的拟合能力一般远远高于问题复杂度，也就是说，机器学习算法有拟合出正确规则的前提下，进一步拟合噪声的能力。

3.5.2 基于学习曲线模型拟合问题判定

学习曲线就是通过画出不同训练集大小时训练集和交叉验证的准确率，可以看到模型在新数据上的表现，进而来判断模型是否方差偏高或偏差过高，以及增大训练集是否可以减小过拟合。

模型学习曲线如图所示:

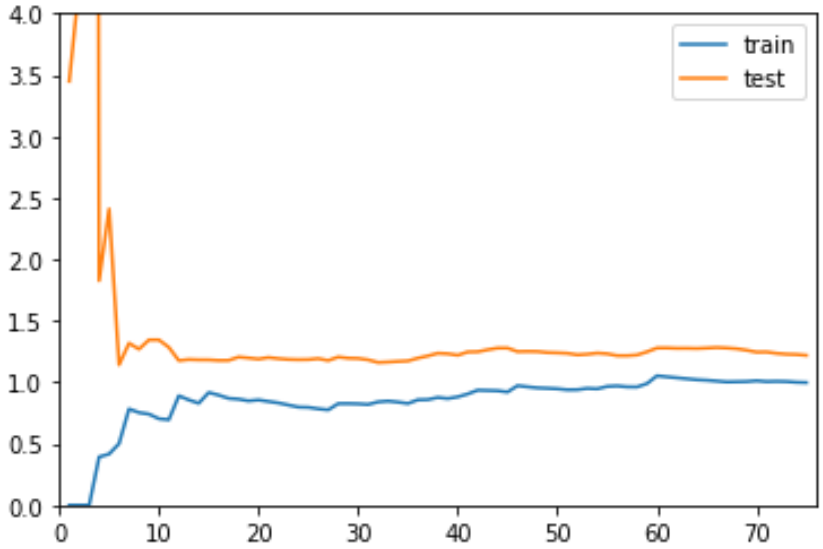


图 18　学习曲线图

由学习曲线可以看出，测试数据集的误差曲线距离训练数据集的误差曲线较近，且趋于稳定，说明模型泛化能力强，对对于新的数据预测误差较小，不存在过拟合现象。表明我们的分类算法可以广泛应用于人们的行为分类问题。

四、 模型评价与推广

4.1 模型优点

1. 问题一采用双层分类器融合模型，相比单一机器学习模型分类性能更好，准确率大幅度提高。
2. 问题二在对模型泛化能力进行评估的过程中，针对多分类模型特性，采用 k 折交叉验证，选取多重指标，全面客观反映模型泛化性能。
3. 问题三在研究过拟合问题时，采用学习曲线进行可视化展示并提出自己的分析。

4.2 模型改进

1. 数据处理相对简单，一些重要的数据可能丢失。
2. 可以多考虑一些指标，建立多方面考虑的智能诊断模型。

参考文献

[1] 邱阳, 李盛, 金亮, 张咪咪, 王杰. 基于统计特征混合与随机森林重要性排序的桥梁异常监测数据识别方法[J]. 传感技术学报,2022,35(06):756-762.

[2] 冯昊, 李树青 基于多种支持向量机的多层级联式分类器研究及其在信用评分中的应用[J]. 数据分析与知识发现,2021,5(10):28-36.

[3] 李辉, 李瑞祥, 张耀威, 乐燕芬，施伟斌 多层分类器模型的相似人体活动识别 [J] 小型微型计算机系统，202142(04):861-867

[4] 路佳佳 基于交叉验证的集成学习误差分析 [J/OL]. 计算机系统应用：:1-8[2022-12-05].DOI:10.15888/j.cnki.csa.008898.

[5] 许志兴, 吴俊华, 唐晓纹 基于粗集的多层分类器的设计与实现[J]. 计算机工程与应用,2006(08):184-186.

[6] 张洋, 姚登峰 人类的行为识别分类方法综述 [C]. 中国计算机用户协会网络应用分会 2019 年第二十三届网络新技术与应用年会论文集.,2019:44-47.DOI:10.26914/c.cnkihy.2019.004425.

[7] 周智强 面向股票价格预测的深度学习过拟合问题研究及其优化[D]. 深圳大学,2020.DOI:10.27321/d.cnki.gszdu.2020.000809.

[8] 吕志浩 多任务学习组合预训练模型泛化能力的研究 [D]. 华东师范大学,2022.DOI:10.27149/d.cnki.ghdsu.2022.001283.

附录 A　第二问基于 K-Means 算法的聚类代码

```
clc
clear
load gaojia.mat
[idx,cmeans3,cen]=kmeans(gaojia,2,'Replicates',1000);
figure(1)
silhouette(gaojia,idx)
color=['r','g','b'];
ptsymb = {'bs','r^','md','go','c+'};
figure(2)
for i = 1:3
    clust = find(idx==i);
    plot3(gaojia(clust,1),gaojia(clust,2),gaojia(clust,3),ptsymb{i});
    hold on
end
syms x
plot3(cmeans3(:,1),cmeans3(:,2),cmeans3(:,3),'ko');
syms y
plot3(cmeans3(:,1),cmeans3(:,2),cmeans3(:,3),'kx');
grid on
hold off
data = gaojia;
syms j
data=mapminmax(gaojia,0,1);
[n,p]=size(data);
figure(3)
syms o
K=8;D=zeros(K,2);
for k=1:K
[lable,c,sumd,d]=kmeans(data,k,'dist','sqeuclidean');
sse1 = sum(sumd.^2);
D(k,1) = k;
D(k,2) = sse1;
end
plot(D(2:end,1),D(2:end,2))
syms p
hold on;
plot(D(2:end,1),D(2:end,2),'or');
title('CaO-Al2O3-SrO 不同K值聚类偏差图')
xlabel('分类数(K值)')
ylabel('簇内误差平方和')
```

附录 B　第四问的 matlab 绘图代码

```
load matlab.mat
x=2011:1:2021;%x轴上的数据，第一个值代表数据开始，第二个值代表间隔，第三个值代表终止
```

```
x1=2013:1:2021;
x2=2011:1:2019;
x3=2012:1:2021;
figure(1)
plot(x,y1,'-Xg',x,y2,'-*r',x,y3,'-ob'); %线性，颜色，标记
 title(['CN trilinear chart'])
legend('CN-CN','CN-Dementia','DN-MCI'); %右上角标注
xlabel('year') %x轴坐标描述
ylabel('ecog composite index')


figure(2)
plot(x1,y4,'-Xg',x1,y5,'-*r'); %线性，颜色，标记
 title(['SMC double line diagram'])
legend('SMC-CN','SMC-MIC'); %右上角 标注
xlabel('year') %x轴坐标描述
ylabel('ecog composite index')

figure(3)
plot(x3,y6,'-Xg',x3,y7,'-ob',x2,y8,'-*r'); %线性，颜色，标记
 title(['LMCI trilinear chart'])
legend('LMCI-CN','LMCI-MCI','LMCI-Dementia'); %右上角标注
xlabel('year') %x轴坐标描述
ylabel('ecog composite index')

figure(4)
plot(x2,y9,'-Xg',x3,y10,'-ob',x,y11,'-*r'); %线性，颜色，标记
title(['EMCI trilinear chart'])
legend('EMCI-CN','EMCI-MCI','EMCI-Dementia'); %右上角标注
xlabel('year') %x轴坐标描述
ylabel('ecog composite index')

figure(5)
plot(x3,y12,'-*r'); %线性，颜色，标记
 title(['AD Line Chart'])
legend('AD-Dementia'); %右上角标注
xlabel('year') %x轴坐标描述
ylabel('ecog composite index')
```