建模实验报告

**2023**
**Machine Learning**
**Summary Sheet**

**Team Number**
26

# 基于特征提取和多分类支持向量机的人类行为分类模型

## 摘要

人类行为理解的一个重要方面是识别和监控日常活动。通过一系列不同传感器获得的数据，可对人类行为进行分类。

在本文中，我们使用了人类行为传感器收集的数据样本 csv 文件，包含属于 19 种类别的多个样本，每个样本包含由分属于 5 个传感器单元的总数为 45 个不同传感器采集的特征。根据这些特征，成功地建立了基于特征提取和多分类支持向量机的人类行为分类模型，将样本数据分为 19 类人类行为。

在本文的第一部分，我们分析了给定的数据，并提出了两种提取特征数据的方法，建立了多分类支持向量机模型，对数据进行分类。

在本文的第二部分，我们使用机器学习中的三个相关指标与不同核函数、不同分类器间的性能对比评估了模型的泛化能力。

在本文的第三部分，我们通过网格搜索调整模型中的关键参数，最终求得最优参数，以应对模型的过拟合问题，对模型进行优化。

最后，我们总结了模型，指出了模型中的不足之处，并提出了可行的改进方法，可提高模型的分类准确度、泛化能力、抗过拟合能力。

**关键词：支持向量机，核函数，方差，Pearson 相关系数，网格搜索**

# 目录

# 1 引言

## 1.1 背景介绍

人类行为的识别与分类是计算机领域的研究热点。活动识别是通过给定的一系列数据来识别一个人执行的动作，从而了解人们的行为活动。嵌入式智能设备以及可穿戴传感器数据都可以作为活动识别的信息来源。基于传感器数据的人类活动识别系统在医疗保健、动态监测、人机交互等方面有着广泛的应用。因此，如何通过传感器采集到的数据实现对人类行为的精准识别成了重中之重。事实上，人们更多地倾向于从运动的结构特征来识别动作。简单来说，就是需要根据所给的人类行为数据设计模型，提取多种行为特征并进行分类。再使用机器学习对处理过的数据进行训练，得到模型计算出的特征值。最后与可穿戴活动识别系统采集到的数据进行对比，从而判断该穿戴设备的数据是否可被确定为某一类人类行为。

## 1.2 待解决的问题

根据题目所述，我们需要建立模型来提取各种人类行为特征，并利用提取的特征与采集的数据进行对比，从而将数据归类为某一人类行为。由于穿戴式传感器每 5 秒采集一次实时数据，数据集非常庞大，所以需要采取合适的预处理方法对数据进行降维，以便设计更高效的分类算法。题目提供的数据量是有限的，使用的模型要尽可能适用于所给数据集之外的新鲜数据样本。因此，还需要制定多维度的评价指标来评估模型的泛化能力。最后，考虑到样本数据量少、模型复杂度高导致模型可能在训练集和测试集上表现不一致，所以需要对模型进行调优，克服过拟合问题。

# 2 问题分析

## 2.1 问题重述

问题的核心可以概括为：给定分属于 19 种类别的多个样本，每个样本包含由分属于 5 个传感器单元的总数为 45 个不同传感器采集的特征。根据这些特征，建立起能将样本数据分为 19 类的模型。

在此基础上，需要采取恰当的、可实现的策略，评估并提高模型对于不同数据的泛化能力，并解决模型的过拟合问题。

## 2.2 解决方案分析

(1) 给定的原始数据具有特征多、基数大的特点，无法直接用于模型的建立，因此需要进行数据预处理与特征提取；

(2) 问题的核心涉及到大量数据的分类，可采取机器学习的相关算法实现；

(3) 评估模型时，可以采取机器学习的常用评价指标，如准确率、精确率、召回率、F1 分数等指标进行评估。此外，还可以换用不同分类模型进行分类，对不同结果进行评价；

(4) 针对泛化与过拟合问题，可使用调整相关参数、参数正则化、特征降维等方法实现。

综上，本问题是可以通过基于机器学习的分类模型解决的。解决方案流程如图1：

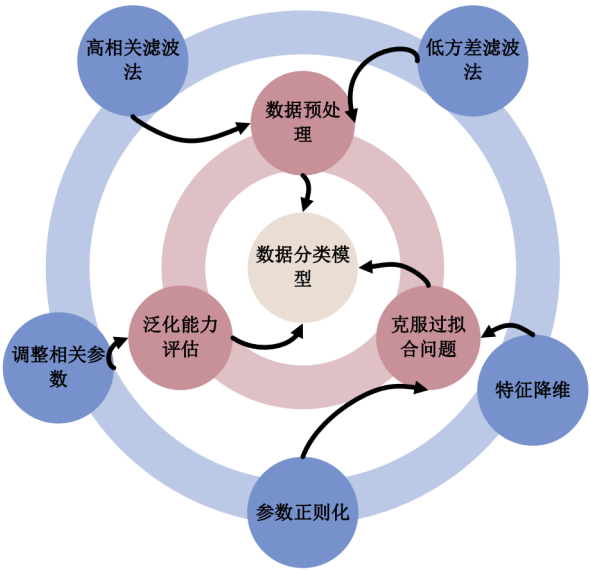


图 1: 解决方案流程图

# 3 模型

## 3.1 基本模型

为解决上述问题，我们使用低方差滤波与高相关滤波对原始数据进行了特征提取，并使用恰当的统计变量对数据进行压缩。此后，我们采用了一种基于多分类支持向量机算法的机器学习分类模型对数据进行分类，并进行模型评估与调优。

### 3.1.1 符号和定义

**定义一：数据特征——列定义**

对给定的数据矩阵，定义矩阵的每一列为数据的一个特征。在本问题中，样本总体的特征对应由 45 个传感器采集到的 45 列数据。这些特征的编号使用 $x_0$，$x_1 \cdots x_{44}$ 表示。

**定义二：数据类别、测试者、时间片段与单个样本——行定义**

根据原数据文件名称与编号，作出如下定义：本问题拟分类 19 种人类活动，使用 1-19 进行编号；每项活动数据来源于 8 位测试者，使用 $p_0$，$p_1 \cdots p_8$ 进行表示；每个测试者共采集了 60 个时间片段（使用 $s_{01}$，$s_{02} \cdots s_{60}$ 表示）的数据，每一片段的数据中包含了相关传感器在 5 秒内共计 125 个特定时刻采样的数据，使用 $l_0$，$l_1 \cdots l_{124}$ 表示.

**定义三：总数据集、训练集与测试集**

将给定的所有数据的集合称为总数据集。其中，实际参与到模型训练中的数据集合称为训练集，实际参与到模型测试中的数据集合称为测试集。在本问题中，对每一个类别中，取 $p_1$-$p_6$ 为训练集，$p_7$、$p_8$ 为测试集。

**定义四：Pearson 相关系数公式** Pearson 相关系数用于度量两个变量之间的线性相关程度，其值介于-1 与 1 之间。

Pearson 相关系数的计算公式如式 (1) 所示。

$$\rho_{X,Y} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{\sum(Y-\overline{Y})^2}} \tag{1}$$

### 3.1.2 问题假设

假设给定的 19 类数据均具有明显特征，即可进行有效分类，且模型可在较短时间内对问题进行求解，训练集与数据集内的数据均符合格式规范。

### 3.1.3 模型基础

(1) 特征过滤方法

- 低方差滤波：

由于数据集非常庞大，并且不是每一个特征都具有高区分度，因而要对数据进行特征选择。低方差滤波的原理为：首先按照发散性度量各个特征。即计算样本中每一个特征值对应的方差，选定一个期望阈值，将方差与阈值进行比较。方差体现了样本在特

征上是否表现出显著差异。若某一特征值方差过小，低于阈值，则说明该特征无法显著地区分所有数据。因此，通过保留高于阈值的所有特征，可以降低数据集特征维度，提高模型精确度。

- 高相关滤波：

  为了减少数据集的复杂程度，需要进一步对数据集特征进行降维处理。数据集之间的相关性是衡量它们相关程度的指标。如果两个特征高度相关，这意味着它们具有相似的趋势，并且可能携带相似的信息。因此，当相关系数超过某一阈值时，可以舍弃其中一个特征，从而降低数据集的维度，同时提高模型性能。

(2) 归一化与标准化

- 最大最小归一化

在数据预处理时，为了消除量纲影响，需要对数据进行标准化或归一化处理。最大最小归一化将数据中的最大值和最小值进行归一化处理，处理后的数值处于 $[0, 1]$ 之间。我们在特征工程中对数据进行归一化，其计算方法如式2所示。

$$x^{'} = \frac{x - min}{max - min} \tag{2}$$

- Z-Score 标准化

Z-Score 标准化是一种中心化方法，基于原始数据的均值和标准差进行的标准化。标准化后数据均值为 0，方差为 1。我们在机器学习中对数据进行 Z-Score 标准化处理，计算方法如下：

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_i - \mu)^2} \tag{3}$$

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

(3) 支持向量机算法（SVM）

支持向量机（SVM）是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。SVM 使用铰链损失函数计算经验风险并在求解系统中加入了正则化项以优化结构风险，是一个具有稀疏性和稳健性的分类器。SVM 可以通过核方法进行非线性分类，是常见的核学习方法之一。

标准 SVM 是基于二元分类问题设计的算法，无法直接处理多分类问题。利用标准 SVM 的计算流程有序地构建多个决策边界以实现样本的多分类，通常的实现为“一对

多”和“一对一”。一对多 SVM 对 m 个分类建立 m 个决策边界，每个决策边界判定一个分类对其余所有分类的归属；一对一 SVM 是一种投票法，其计算流程是对 m 个分类中的任意 2 个建立决策边界，即共有

$$\frac{m(m-1)}{2} \tag{5}$$

个决策边界，样本的类别按其对所有决策边界的判别结果中得分最高的类别选取。一对多 SVM 通过对标准 SVM 的优化问题进行修改可以实现一次迭代计算所有决策边界。在此，我们采用了“一对一”SVM。

考虑到 SVM 模型的数学原理较为复杂，本文将不对其详细原理进行说明，仅介绍影响模型建立的三个重要参数。

- 核函数

常用的核函数如表1所示：

表 1: 常用核函数

| 名称 | 解析式 |
|---|---|
| 线性核函数 kernel= ’linear’ | kernel=$< x, x^{'} >$ |
| 多项式核函数 kernel= ’poly’ | kenel=$(\gamma < x, x^{'} > +r)^{d}$ |
| 径向基核函数 kernel= ’rbf’ | kernel=$exp(-\gamma \left\|x - x^{'}\right\|^{2})$ |
| sigmod 核函数 kernel= ’sigmod’ | kernel=$tanh(\gamma < x, x^{'} > +r)$ |

- 正则化参数 C

正则化的强度与 C 成反比，必须严格为正。引入正则化系数 C，可以理解为允许划分错误的权重（越大，越不允许出错），当 C 较小时，允许少量样例划分错误。C 越大，说明越不能容忍出现误差, 容易过拟合。C 越小，容易欠拟合。C 过大或过小，泛化能力变差。

- 核系数 $\gamma$

$\gamma$ 是用于非线性支持向量机的超参数。最常用的非线性核函数之一是径向基函数 (rbf)。rbf 的 $\gamma$ 参数控制单个训练点的影响距离。

$\gamma$ 值较低表示相似半径较大，这会导致将更多的点组合在一起。对于 $\gamma$ 值较高的情况，点之间必须非常接近，才能将其视为同一组 (或类)。因此，具有非常大 $\gamma$ 值的模型往往出现过拟合情况。

### 3.1.4 解决方案与结果

(1) 数据预处理与特征提取

首先，我们使用 pandas 分析了总数据集，未发现存在缺失值与格式不合法的值。之后，对训练数据进行最大最小归一化处理，发现其存在如图2所示特征：

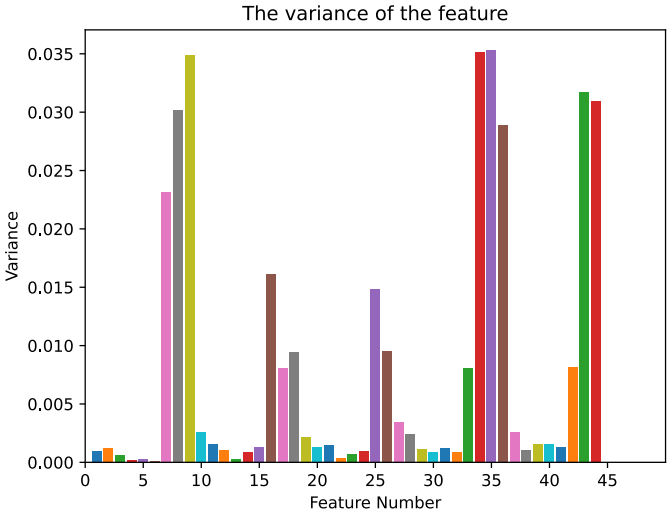


图 2: 低方差滤波结果

取 0.005 作为阈值，使用低方差滤波法，可筛选出编号为 $x_6$，$x_7$，$x_8$，$x_{15}$，$x_{16}$，$x_{17}$，$x_{24}$，$x_{25}$，$x_{32}$，$x_{33}$，$x_{34}$，$x_{35}$，$x_{41}$，$x_{42}$，$x_{43}$，$x_{44}$ 的 16 组特征集。

随后，我们采用高相关滤波法进行进一步过滤。经过多次尝试与相关案例调研，确定阈值为 0.8。计算上述 16 组特征集的 Pearson 相关系数，得到如图3所示特征：

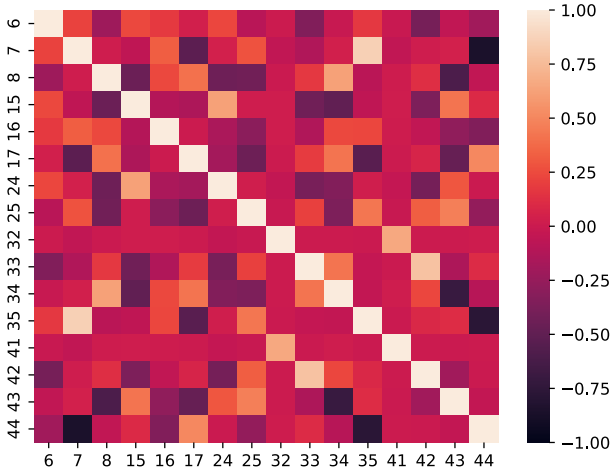


图 3: 高相关滤波结果

最终，筛选出 13 组显著特征：$x_6$，$x_7$，$x_8$，$x_{15}$，$x_{17}$，$x_{24}$，$x_{25}$，$x_{32}$，$x_{33}$，$x_{34}$，$x_{41}$，$x_{43}$，$x_{44}$。

从训练集与测试集中分别选取出这 13 组显著特征，并提取每个时间片段中数据的均值与方差这两个特征，进行数据压缩，同时细化特征，构建出含有 26 个特征，6840 个数据的训练集与含有 26 个特征，2280 个数据的测试集，并进行 Z-score 标准化，以便进行模型训练。

(2) 模型训练

使用 scikit-learn 库内的 svm 分类器进行模型构建。由于数据具有特征维度数量远小于样本个数的性质，选择核函数为‘rbf’，构建偏向非线性、高维度的分类模型。取参数 C = 64，gamma = 0.0078125，得到如表2所示的分类结果：

表 2: 分类结果

| | |
|---|---|
| 训练集得分 | 0.9931286549707602 |
| 测试集得分 | 0.823245614035087 |
| 测试集错误个数 | 403 |
| 测试集错误占比 | 0.17675438596491228 |

### 3.1.5 结果分析

训练集得分基于预测结果的准确率指标得到，表示在给定的训练集上，我们的分类模型准确率达到了 99.3 %，可有效地对具有大量特征的给定数据进行分类。在测试集上，模型的分类准确率达到了 82.3 %，具有较为优秀的性能。在测试集上的错误分类结果中，分入错误类别后的错误数据编号分布如图4所示：

由此可见，误差出现在类别 7 与类别 8 的情况占比较高，模型对于类别 7 与类别 8 的误识别率较高。

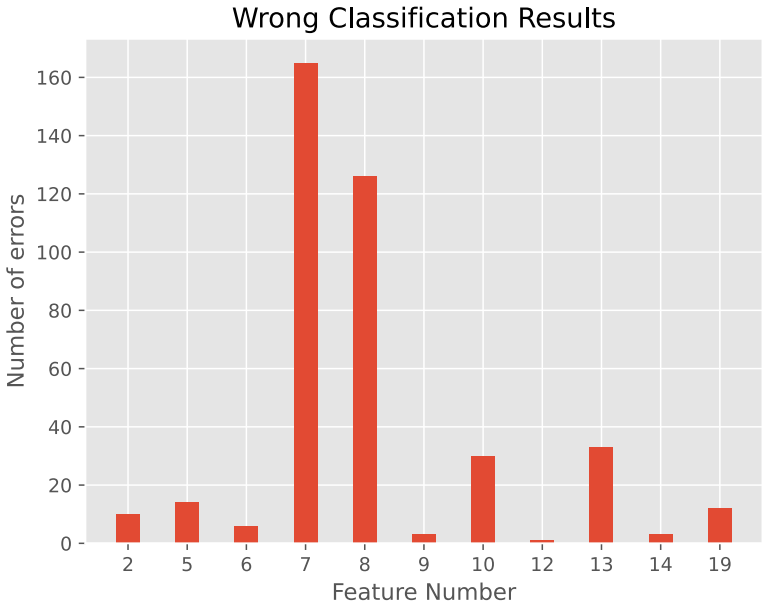


图 4: 错误数据编号分布

#### 3.1.6 模型优点与缺点

- 优点:

  模型采用 SVM 多分类算法，具有较好的泛化性、唯一的全局最优解和鲁棒性等特点，在解决非线性、有限样本等分类问题中表现出特有的优势。此外，SVM 参数少，SVM 的调包、优化较为简便. 因此选用 SVM 的方法对读取的数据进行学习，输出分类模型。

- 缺点:

  由于 SVM 算法不适用于大量数据，因此我们对数据的特征进行了压缩，这势必会损失一部分有效信息。此外，SVM 算法具有较高的时间复杂度，其时间复杂度介于 $O(n_{features} \times n_{samples}^2)$ 和 $O(n_{features} \times n_{samples}^3)$ 之间，因此，本模型适用于一般配置计算机、机器学习模型，在数据量较多时会降低其运行效率。

### 3.2 针对泛化能力的模型改进

#### 3.2.1 附加条件

由于数据成本较高，我们需要在有限的数据集下使模型具有良好的泛化能力。因此，需要按照具体的衡量指标研究和评估这个问题。

**附加定义一：混淆矩阵**

以 2×2 的混淆矩阵为例，如图5所示。

| Confusion Matrix | | 真实值 | |
|---|---|---|---|
| | | P | N |
| 预测值 | $P'$ | TP | FP |
| | $N'$ | FN | TN |

图 5: 混淆矩阵示例图

- P 表示正例；
- N 表示负例；
- FP 表示实际为负但被预测为正的样本数量；
- TN 表示实际为负且被预测为负的样本数量；
- TP 表示实际为正且被预测为正的样本数量；
- FN 表示实际为正但被预测为负的样本数量；
- $P'$ 表示所有被预测为正的样本数量；
- $N'$ 表示所有被预测为负的样本数量。

### 3.2.2 解决方案与结果

针对模型的评价，结合本模型中使用的数据经过 Z-score 标准化，采用如下三个指标对模型进行评估。

- 准确率：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

- 精确率：(此处取宏平均精确率进行计算)

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

- 召回率：

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

模型的准确率、精确率、召回率如表3所示：

表 3: 模型评价结果

| **评价指标** | **数据结果** |
|---|---|
| 准确率 | Accuracy = 82.325 % |
| 精确率 | Precision = 88.335 % |
| 召回率 | Recall = 82.325 % |

### 3.2.3 泛化能力对比

(1) 采取不同核函数进行模型训练，依据指标对模型进行评估. 在此仅举采取线性核函数 kernel=‘linear‘进行训练的结果。

表 4: linear 模型训练结果

| **评价指标** | **训练结果** |
|---|---|
| 准确率 | Accuracy = 81.271 % |
| 精确率 | Precision = 83.282 % |
| 召回率 | Recall = 81.272 % |

由表4可见，选择 kernel=‘rbf’提升了本模型的泛化能力。

(2) 采取不同算法进行分类模型构建，在此仅举采取随机森林分类器进行模型构建的结果。由表5可见，多分类 svm 在针对偏向非线性、多特征数据进行分类时，具有良好的泛化能力。

表 5: 随机森林分类器训练结果

| 评价指标 | 训练结果 |
|---|---|
| 准确率 | Accuracy = 82.105 % |
| 精确率 | Precision = 85.861 % |
| 召回率 | Recall = 82.105 % |

(3) 在未经 Z-score 标准化处理的原始测试集中，随机添加 2 % 噪声进行验证，得出准确率为 81.172 %，由此可见，模型具有良好鲁棒性。

需要指出的是，不同分类模型的评估指标虽有差别，但差别较小，反映出所给的待分类数据存在偏向于线性相关的类别影响因子。

此外，为提高模型的泛化能力，在构建模型进行参数调整时，我们采取了优化后的方法，具体将在 3.3 中与过拟合问题的改进方法一并介绍。

## 3.3 针对过拟合问题的模型改进

### 3.3.1 解决方案和结果

SVM 处理数据要求对其进行标准化，我们采取 Z-score 标准化方法，对于过拟合具有一定的修正能力。另外，在进行特征提取时，先进行低方差滤波，再进行高相关滤波，提高了模型偏向线性与偏向非线性两种数据分布方式的泛化能力，并在一定程度上放置了过拟合情况。但此两项并非主要改进措施。

对于多分类 SVM，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 的目标函数采用结构风险最小化策略，但是由于允许误分类的存在核引入了核函数，SVM 仍会有过拟合的问题。

我们构建的 SVM 模型是核函数 + 软间隔的支持向量机，那么，有以下原因导致 SVM 过拟合:

(1) 核函数导致过拟合。可通过调整非线性核函数 rbf 的核系数参数 $\gamma$ 进行修正;

(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，相当于硬间隔 SVM。而当 C 过小时，模型容易出现欠拟合情况。

在机器学习中，可以绘制验证曲线获取最优超参数，但验证曲线只能每次获取一个最优超参数。如果多个超参数有很多排列组合，就可以使用网格搜索寻求最优超参数的组合。

综上，采取网格搜索的方法进行参数调整，确定适用于给定数据分类的最佳参数 $(C, \gamma)$，使用 scikit-learn 中的 GridSearchCV 工具实现，在区间 $[2^{-5}, 2^{15}]$ 内搜索 C，在区间 $[2^{-9}, 2^{3}]$ 内搜索 $\gamma$。流程如图6所示。

最终，寻找到最优内核为 rbf，最佳的 $(C, \gamma)$ 参数为（64，0.0078125），有效地克服了因过度训练而导致的过拟合问题。

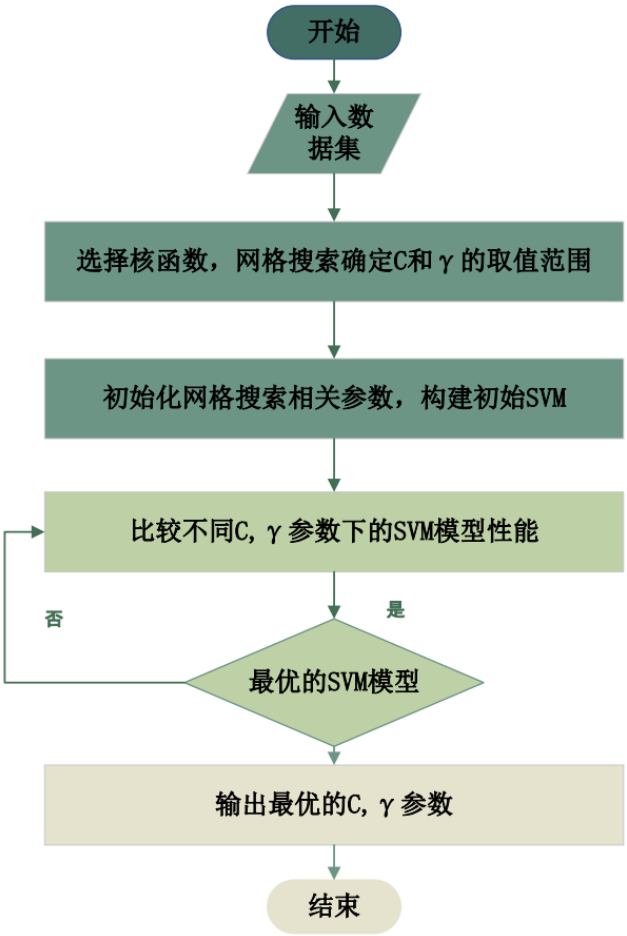


图 6: 模型调优流程图

# 4 结论

## 4.1 问题总结

为解决人类行为数据的分类问题，我们进行了如下的模型构建：

- 基于低方差滤波与高相关滤波的特征提取
- 基于多分类支持向量机的分类器
- 对模型泛化能力的量化评估指标与对比评估
- 对模型过拟合问题的有效解决方案

## 4.2 模型中使用的方法

- 算法：

  低方差滤波

  高相关滤波

  多分类支持向量机

  混淆矩阵、准确率、精确率与召回率

  网格搜索

- 工具：

  scikit-learn

  numpy,matplotlib,pandas

## 4.3 模型应用

本模型可应用于具有多维特征数据的人类行为分类问题，并可推广至给定数据集，且数据集具有切实可分类特征情况下的分类问题。

# 5 未来工作

## 5.1 模型的不足

由于时间限制与计算机硬件设备的有限可行性，我们在模型的构建与调优上仍然存在不足之处：

- 未进一步提取特征：仅对每一个时间段提取了均值与方差两个特征进行升维;
- 未依据每项特征进行单独分类：可牺牲计算量以进行更为良好的特征提取;
- 模型的高复杂度：占用机器内存大、消耗时间长;

## 5.2 可行的改进方案

1. 在原有基础上再通过一些降维方法（如依据每项特征进行单独分类的预训练），进一步删去一些无用特征，继续降低模型复杂度，提高准确率。
2. 选择更多的统计量（如极差、峰值、偏斜度、均方根频率等），对数据进行更好的压缩-升维描述
3. 及时停止。对于训练过度问题，当模型的准确度不发生变化时及时停止训练，可以有效防止过度训练。
4. 使用深度学习方法，利用支持深度学习的软硬件环境（如 Pytorch、cuda、cudnn 等），增加数据量进行深度学习。

# 6 参考文献

[1] Fan, Rong-En, et al., "LIBLINEAR: A library for large linear classification.", Journal of machine learning research 9.Aug (2008): 1871-1874.

[2] Bishop, Pattern recognition and machine learning, chapter 7 Sparse Kernel Machines

[3] Altun K, Barshan B, Tunçel O. Comparative study on classifying human activities with miniature inertial and magnetic sensors[J]. Pattern Recognition, 2010, 43(10): 3605-3620.

[4] A Tutorial on Support Vector Regression" Alex J. Smola, Bernhard Schölkopf - Statistics and Computing archive Volume 14 Issue 3, August 2004, p. 199-222.

[5] Platt "Probabilistic outputs for SVMs and comparisons to regularized likelihood methods"

[6] Ahmed M, Antar A D, Ahad M A R. An approach to classify human activities in real-time from smartphone sensor data[C]//2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR). IEEE, 2019: 140-145.

[7] 陆俊儒. 基于支持向量机的高维不平衡数据二分类方法的研究. 哈尔滨工业大学，2018

[8] Crammer and Singer On the Algorithmic Implementation ofMulticlass Kernel-based Vector Machines, JMLR 2001.

[9] 刘佳. 支持向量机在不平衡数据分类中的研究与应用. 厦门大学，2021

[10] 刘东启. 基于支持向量机的不平衡数据分类算法研究. 浙江大学，2017

# 7 附件

## 7.1 源程序

参见附件-demo。

## 7.2 数据集

即问题给定数据集，在原问题中可获取。