
北京航空航天大學
BEIHANG UNIVERSITY

# 数据挖掘课程 2023 年春

# 期末大作业

| | |
|---|---|
| 院（系）名称 | 人工智能研究院 |
| 团队名称 | 主楼挖金 |
| 团队学生 | 修曾琪 张家瑞 魏少杭 |
| 指导教师 | 庄福振 |

2023 年 X 月

# 目录

# 第一章 背景

## 1.1 背景描述

北航自建校伊始便始终与共和国的航空航天事业紧密相连，新时代强军强国梦，北航亦不会缺席，作为北航学子的我们也必须将所知所学应用到军事中，为国家领土和发展安全做出杰出贡献。

目前，我国周边局势呈现日趋动荡的现象，同时新时代的战争环境已经变得越来越复杂，海陆空、多层次、立体性成为了现代战争的三大突出特征。严峻的战争形式已经对我国防控系统的智能化、信息决策提出了更高的要求，作战指挥必须综合考虑多个维度的信息来有效识别意图并做出正确决策。

为了更好地保障我方指挥员作战指挥能力，我们将开发一个意图识别模型，来使得指挥员更快速、更精准地做出决策。

## 1.2 任务描述

本项目给定的数据集包括了 train.csv 和 test.csv 两个大文件，每个文件中每一条数据均包含了多个变量，时间戳、目标唯一编号、目标唯一名称、目标意图、目标类型、敌我识别结果、目标当前位置的经纬度和海拔、xyz 各方向速度、速率、干扰标记、分组。我们需要根据时间、经纬度、海拔、各方向速度和速率、干扰标记、分组等基本特征数据进行合理的特征工程，并训练分类器来精准识别敌机的意图。

# 第二章 原理

在本项目中我们在 baseline 随机森林分类器的基础上添加了 SVM 分类器，再将随机森林和 SVM 组合进行软投票分类。以下是相关原理简述。

## 2.1 随机森林

随机森林作为 Bagging 集成方法的变体，是一个包含多个决策树基学习器的分类器，其输出的类别由个别树输出类别次数最多的类别决定。在一个大数据集上构造随机森林时，对于森林中的每一棵决策树通过有放回抽样的方式采样训练集，并随机抽取数据集

中的特征进行训练。这种随机采样、集成方法能够很好地解决决策树过拟合现象，并提高了模型在分类任务上的表现性能。

## 2.2 支持向量机

支持向量机是一种监督学习的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最大的边距超平面。当训练数据可分时，通过硬间隔最大化，学习一个线性的分类器，即线性可分支持向量机；而当训练数据近似线性可分时，通过软间隔最大化，学习一个线性的分类器，即线性支持向量机；而当训练数据不可分时，通过使用核技巧以及软间隔最大化，学习非线性支持向量机。

## 2.3 投票

投票分类器是组合概念上不同的机器学习分类器，并使用多数投票或加权平均预测概率来预测类别的标签。投票分类器包括了硬投票和软投票两类。硬投票，即多数投票法，根据少数服从多数的原则；若是有并列的最高票，则按照升序顺序选择。软投票，即加权投票法，相比于硬投票增加了权重参数，使用加权平均概率来预测类别标签。

# 第三章 方法

## 3.1 数据统计分析与观察

### 3.1.1 “干扰”、“诱扰”分析与处理

通过观察训练数据集和测试数据集，可以看到，当一架飞机为“干扰”目的时，它的 interfere_flag=1，而非“干扰”目的的飞机，它的 interfere_flag=0。因此，我们可以认为 interfere_flag 与“干扰”目的是强相关的。

此外，“诱扰”目的的飞机的 formation 基本上都是“横队”，因此我们认为“横队”的 formation 与“诱扰”也是强相关的。

综合上述分析，我们在测试阶段分类器输出预测结果后，再构造映射规则，将 interfere_flag=1 的飞机标记为“干扰”目的，formation 为“横队”的飞机标记为“诱扰”目的，以提高分类准确性。

### 3.1.2 group 属性对意图识别的信息增益

我们对不同意图的飞机在 group 上的概率分布进行画图分析，如图 3.1 所示：

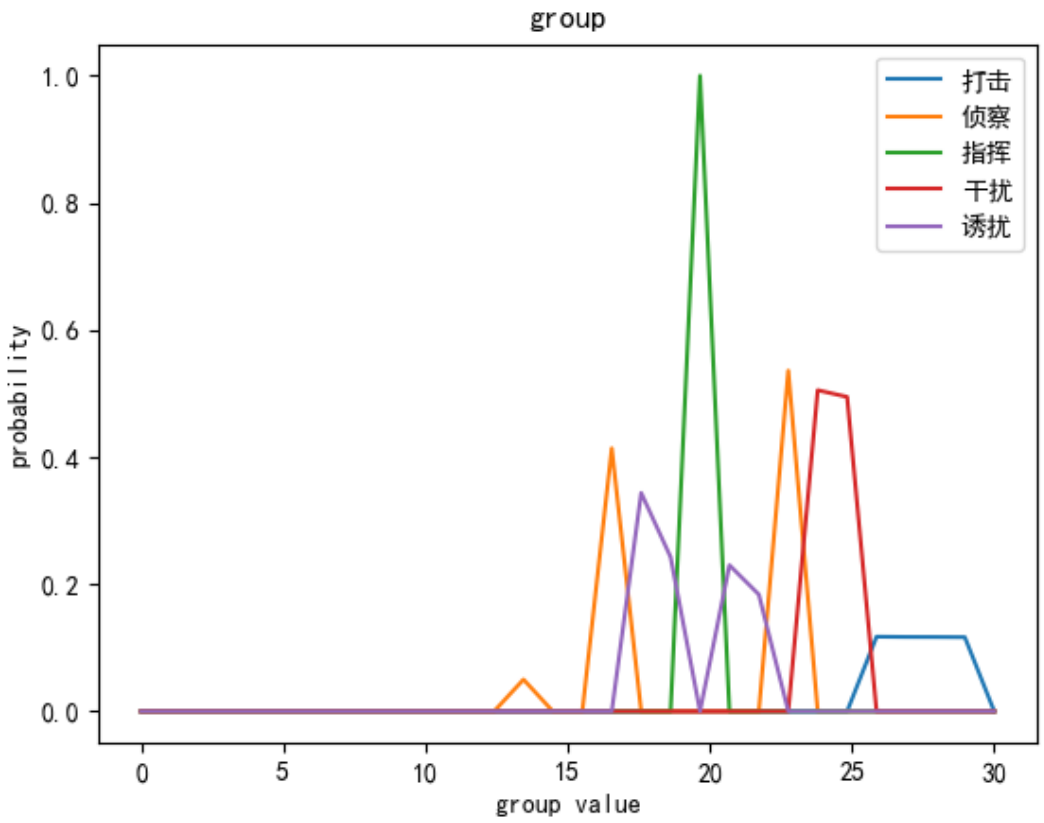


图 3.1 训练集中飞机意图类型关于 group 值的概率分布

从统计分析结果来看，对于不同的 group 值，有且仅有一类飞机意图与之对应。所以，我们可以将 group 作为特征，这将有助于我们在分类的时候将不同的类区分得更明显。

## 3.2 数据特征组合

### 3.2.1 加速度（速率变化率）

我们认为，敌人为了实现不同意图，往往会采用不同类型的飞机。不同类型的飞机在有限的 5 个连续的雷达探测时间点跨度内往往会采取不同程度的踩油门加速、刹车减速策略。为此，我们希望通过利用已有数据求出加速度，来反映飞机的实时行动过程。具体而言，对于一架飞机的全部时序序列，其时间间隔为 1s，那么我们可以直接通过相

邻时间点的速率值相减得到速率变化率。由于在 1s 时间间隔内，飞机方向变化不大，所以我们近似认为速率变化率等于加速度值。我们将求得的加速度值作为每一个时间点的新的特征。

### 3.2.2 vx, vy, vz 和速率

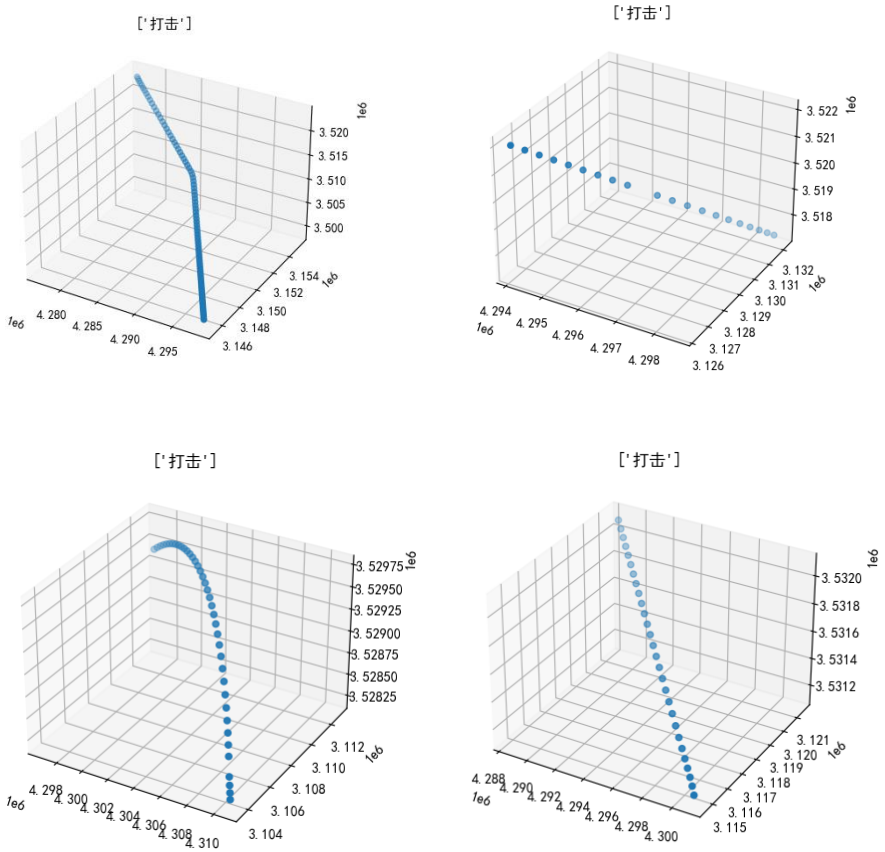


图 3.2 打击意图的典型轨迹示意图

我们认为，敌机若有不同的意图，往往会呈现具有不同特征的轨迹。于是我们分析了不同类型的意图所对应的轨迹。

如图 3.2 所示，我们使用 python 脚本画出了训练集中打击意图的运动轨迹，主要呈现以下特点：第一，打击的速率较快；第二，打击飞行高度不断降低。

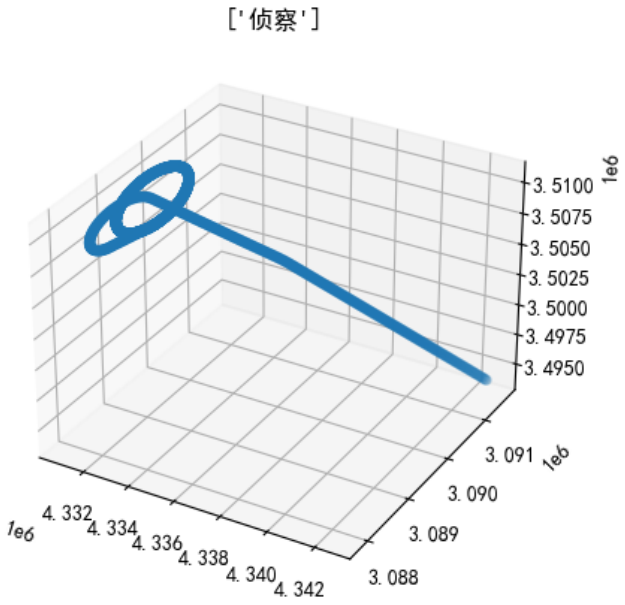


图 3.3 侦察意图的典型轨迹示意图

如图 3.3 所示是典型的侦察机的轨迹，可以看到侦察机的飞行速率比较缓慢，并且行动轨迹出现盘绕旋转的特点，这说明了我们需要使用 vx、vy、vz 等不同方向速度以及速率作为特征。

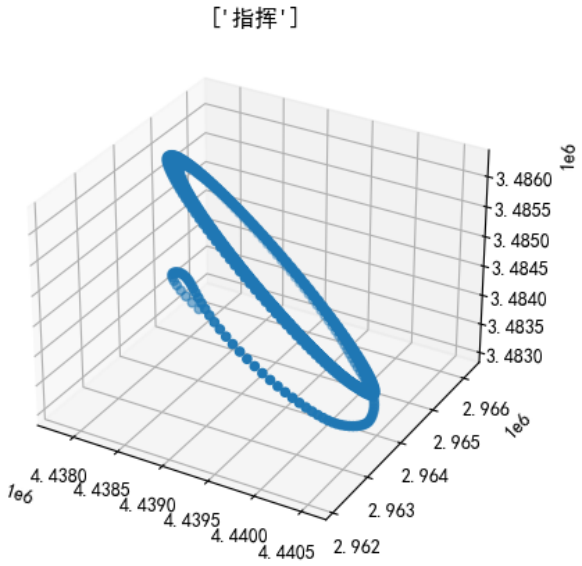


图 3.4 指挥意图的典型轨迹示意图

如图 3.4 所示是典型的指挥机的轨迹。可以看到，指挥机呈现速度较快的巡回轨迹。综合以上分析，我们可以使用 vx、vy、vz 以及速率值作为输入进行有效地分类。

## 3.3 Padding 操作分析与改进

在 baseline 中数据集获取部分，使用了 padding 操作，目的是为了将同一架飞机的最后几个不满足 5 个点的轨迹点进行补全，以使得能够充分利用已经获取到的数据来构造成完整形状的数据集。但是 baseline 中的 padding 草率地使用 0 来补全，这是不恰当的。我们认为，尽管后续进行了标准化操作使得不同维度的特征维度通道上按照在均值上达到 0 的结果，但是直接用 0 来补全后面 4 个点的所有维度的特征是不对的，因为例如坐标 x、y、z 在连续 5 个时间点构成的一个时间段内往往均值不为 0 甚至相差很远，而这时突然在缺少的时间点内补全为 0，就造成了不符合实际情况的数据变化情况。

在实际操作中，我们取消了 padding 操作，原因有二：第一，我们认为每一架飞机在雷达探测的范围内点数已经够多，也就是说已有的小段数够多，不必去构造虚假的特征值来增加样本；第二，如果进行了 padding，那么必然会造成训练数据的不真实性，可能因为人为的预先定义造成了分类器学习到有偏差的特征规律，进而导致在测试集上的分类准确率降低。实验证明，如果保留补全 0 的 padding 操作，测试集上的准确率为 0.9622，F1 score 为 0.9614；当我们取消了 padding 之后，我们的 acc 为 0.9677，F1 score 为 0.9659。这也印证了取消 padding 操作是恰当的。

## 3.4 数据集上 shuffle 操作分析

在原始的训练集中，是按时间顺序排列的，同时我们在加载数据的时候也是按顺序加载的，可能会在训练阶段使得模型认为样本与时间特征具有相关性，造成预测结果因为隐含的时间关系出现偏差。而在测试集上，我们每一条数据都是独立通过模型进行预测的，那么尽管不进行 shuffle 而保留了时间顺序，但时间顺序不会对预测结果造成任何影响。

## 3.5 数据标准化操作分析

标准化操作能够将每一个维度按照数据集中所有样本该维度的值进行标准化，达到标准正态分布，即均值为 0，方差为 1 的分布。标准化操作的作用主要是将不同维度特征的值变化范围拉伸到一致性。具体而言，在需要根据不同维度的特征计算样本距离时，

如果不同维度特征的值变化范围不一致或者尺度不一致，会影响模型的训练和预测性能，主要影响如 SVM 和 KNN 算法。从这个方面来说，决策树和随机森林等不需要计算样本距离的算法不会受标准化操作太大影响。

如果不进行标准化，则模型无法真正捕捉到各个特征维度的变化规律。这时，如果训练集样本和测试集样本在某些维度上出现了分布上的较大差异，训练集上训练的模型将无法迁移到测试集上预测，所以测试集上的模型性能会大幅下降。从这个角度来说，决策树和随机森林也同时会受到较大影响。

## 3.6 最优超参数的确定

### 3.6.1 随机森林超参数

随机森林中主要包含了基本决策树个数（即 n_estimators）和树最大深度（即 max_depth）两个超参数。

基本决策树个数越多，则可以提高模型的准确性和稳定性。随机森林中的每个决策树都是独立建立的，增加决策树的数量可以减少模型的方差，提高模型的泛化能力；然而增大基本决策树个数会增加内存占用和计算成本。

树最大深度越深，则模型的拟合能力就越强，因为决策树可以考虑更多的特征，从而更好地适应训练数据。但是如果树最大深度过深，其可能会导致模型对训练数据过度拟合，而在测试数据上的表现较差；同时，如果树最大深度过小，也会导致模型欠拟合，进而导致模型的准确性较低。

经过综合权衡并调整，随机森林分类器的最优参数为 20 个基本决策树个数，树最大深度为 3。

### 3.6.2 支持向量机超参数

在我们使用的 SVM 模型中，正则化参数为默认值 1，Gamma 调整为 0.5，kernel 使用 RBF 核函数。正则化参数越大，分类间隔越硬，容易过拟合；而正则化参数越小，分类间隔越软，容易欠拟合。而 Gamma 作为核函数的系数，其控制了支持向量的影响范围。Gamma 越小，则支持向量的影响范围越大，分类界面越平滑；Gamma 越大，支持向量的影响范围越小，分类界面越复杂。最后，我们使用了径向基函数（RBF）核函数。

# 第四章 实验效果

综合上述分析，我们使用的特征维度包括了：1.x 坐标，2.y 坐标，3.z 坐标，4.速率 v，5.偏航角 heading，6.俯仰角 pitch，7.滚转角 bank，8.x 方向速度 vx，9.y 方向速度 xy，10.z 方向速度，11.加速度（速率变化率）dv，12.group 值。我们通过使用由随机森林和支持向量机集成的投票分类器，并将随机森林超参数调整为 20 个基本决策树个数，树最大深度为 3；支持向量机超参数调整为使用正则化参数 1，Gamma 调整为 0.5，kernel 使用 RBF 核函数。

我们按照上述的方法和超参数设置，按照 8:2 在原始训练集上进行训练集、验证集划分，求数据集上的准确率和 F1 Score 情况如表 4.1 所示：

表 4.1 实验结果

| 数据集类型 | 准确率 | F1 Score |
|---|---|---|
| 训练集 | 1.0000 | 1.0000 |
| 验证集 | 0.9982 | 0.9970 |
| **测试集** | **0.9943** | **0.9936** |

可以看到，我们的意图识别模型相比于 baseline 有了准确率 13%以上的提升，且准确率十分接近于 1.0，所以我们的模型效果非常好。

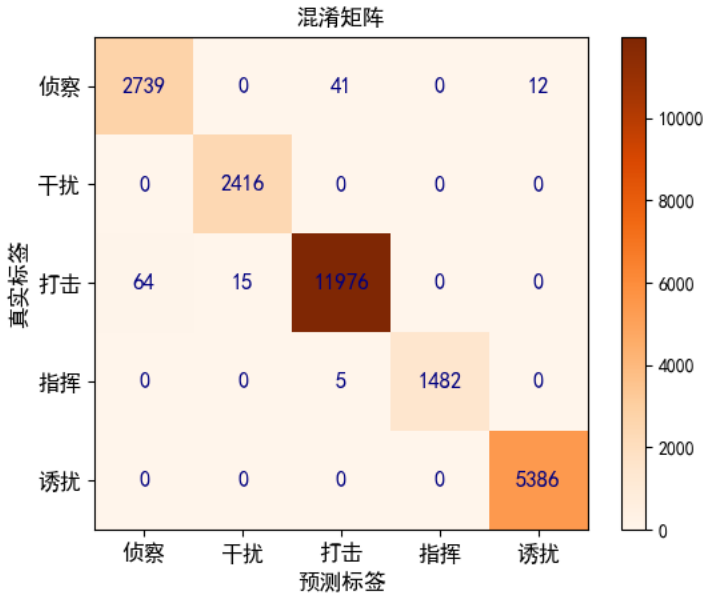


图 4.1 测试集预测结果混淆矩阵

如图 4.1 是本次预测结果的混淆矩阵热力图。通过此图可以看出，模型能够将绝大

多数的样本分类正确。此外，容易出现将侦察误分类为打击、诱扰；将打击误分类为侦察、干扰；将指挥误分类为打击。这些被分类错误的情况可以被认作是训练数据不够大等因素导致的。

# 第五章 总结与展望

## 5.1 优势

首先，我们小组通过绘制和分析不同意图的飞机的典型轨迹图、飞机意图关于不同维度的分布等特征的统计结果，能够很好地找到合适的特征来完成我们的意图识别任务。

其次，我们小组还在随机森林的基础上添加 SVM 并组合成一个投票分类器，通过分析超参数对效果的影响规律，可以得到比较合适的超参数组合。

最后，我们的意图识别模型的结果非常良好，在测试集上达到了 99.43%的准确率。

## 5.2 不足

在数据特征分析和利用方面，尽管我们已经分析了原始数据和分类标签的分布规律，但仍然处于低级特征挖掘的阶段，没能从真实的战争场景中去分析和总结各个意图的规律，进而组合生成更具有物理意义和高级语义的特征。

在训练效率方面，我们使用了 SVM，这种算法在训练时效率不高，运行代码时间很长。

## 5.3 展望

第一，我们采样 5 个点时使用的是连续的 5 个点，然而如转弯等操作是需要长时间才能实现的，所以 5 个较长时间间隔的点的数据也许会更有用。

第二，尽管我们之前已经根据 interfere_flag=1 来指定干扰标签，通过 formation=“横队”来指定诱扰标签，但我们通过混淆矩阵可以看到，仍会有其他类型的真实标签的飞机被识别为“干扰”、“诱扰”两种标签，所以我们之后还可以考虑去除“干扰”、“诱扰”这两种标签的预测可能性，让样本只在剩下的 3 个可能的 label 中进行选择，从而进一步提高分类准确率。

# 第六章 收获、体会及建议

我们通过意图识别建模任务，学习到了分析原始数据特征的重要性；此外，我们还能够学会多方面尝试模型设计和超参数设置，以不断提高模型的准确率和泛化性。

我们小组成员通过共同分析、提出不同的见解，共同提升了模型分类效果，这锻炼了我们团队合作沟通的能力，也让我们对机器学习知识和技巧更加熟悉。

建议方面，我们希望以后可以将这个项目更紧密地与数据挖掘课程上的知识点相关联，让我们能够在课程结束后及时巩固知识。