# 数据挖掘算法应用

## 数据集：航空发动机数据集

PCoEDataSets 平台中涡扇发动机降解模拟数据集 (Turbofan Engine Degradation Simulation Data Set,TEDS)

每个样本为26列数据，由空格分隔，每列对应为：发动机序号、 时刻 tp、 操作设置（op1-op3），传感器测量值（para1-para21）

| 发动机序号 | 时刻 | op1 | op2 | op3 | para1 | para2 | ..... | para21 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | -0.0007 | -0.0004 | 100.0 | 518.67 | 641.82 | ..... | 23.4190 |
| 1 | 2 | 0.0019 | -0.0003 | 100.0 | 518.67 | 642.15 | ..... | 23.4236 |
| 1 | 3 | -0.0043 | 0.0003 | 100.0 | 518.67 | 642.15 | ..... | 23.3422 |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... |

图一：航空发动机数据集

## 数据预处理

### 故障分类

•发动机状态共有五种——正常状态，故障1-4

•选择每个发动机的一条正常状态数据和故障数据组成新的数据集，减少数据数量

•对每条数据打上标签，生成标签数据，和新数据集一起存储在.pkl临时文件中

•读取临时文件中数据，使用数据挖掘中的多种分类方法对故障类型分类

```
共有：2832条数据
正常数据：1416条
故障一有：200条，占7.06%
故障二有：519条，占18.33%
故障三有：200条，占7.06%
故障四有：497条，占17.55%
```

```
被选择的数据的向量数和特征数为：(2832, 26)
所有故障数据的向量数和特征数为：(1416, 26)
故障类型1向量数和特征数为：(200, 26)
故障类型2向量数和特征数为：(519, 26)
故障类型3向量数和特征数为：(200, 26)
故障类型4向量数和特征数为：(497, 26)
```

图二：故障分类结果

## 数据标准化

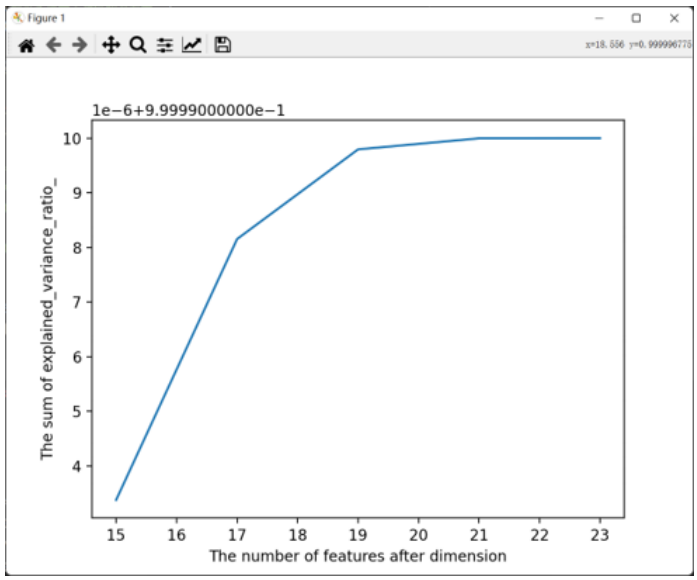
将数据按比例缩放，使之落入一个小的特定区间，实现方法：$z = (x - \mu)/s$ $s$为样本标准差，$\mu$为样本均值。

```
标准化后每个特征的均值是：
[-1.95700330e-16  1.73119522e-16  4.51616146e-16 -1.33477661e-15
  4.13981467e-16 -1.94696738e-15  1.48029737e-16  1.36990231e-15
  5.25881912e-15  2.78496623e-16  5.06813675e-16  6.04663839e-16
  3.66310874e-16 -4.91759803e-15  8.27962934e-16  1.00359143e-16
  1.85664415e-16 -1.35484844e-16 -1.37993822e-16 -4.36562274e-16
 -4.01436574e-17 -2.50897859e-16  1.70610544e-15 -1.68101565e-15]
标准化后每个特征的方差是：
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
```

图三：数据标准化结果

## 主成分分析降维

对高维度特征数据预处理方法，能够将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征（该数据集有26个维度），主成分分析方法(Principal Component Analysis，PCA)，是一种使用最广泛的数据降维算法。PCA主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。




图四：选择留下不同维度能代表原数据集的信息量

经过试验，保留19个维度的向量，能表示原数据集99.9999796%的信息量，数据集从26维降维为19维。

```
每个参数在数据中所占的信息比例为：
[8.13262180e-01 1.68634298e-01 1.03780884e-02 3.82340048e-03
 1.95140483e-03 9.52764565e-04 3.89290788e-04 2.79271758e-04
 1.86238447e-04 4.61269128e-05 3.69837982e-05 3.38937625e-05
 8.58851493e-06 6.45035480e-06 4.39630571e-06 2.78760728e-06
 1.98938296e-06 1.46410714e-06 1.77768994e-07]
降维后的数据能够代表原数据的比例为：99.99997960635348%
```

图五：数据集参数降维分析结果

## 数据集分类

•发动机数据集分为训练集01-04和测试集111-444

•在一个训练集或测试集中有多个发动机的数据，记录了每个发动机从开始运行到故障时刻的全部数据，最后一刻的数据是故障时刻的数据

•共有4种故障类型，分别存储在训练集和测试集中，同一个文件中存储的是同一种故障类型

•编号相同的训练集和测试集存储的是同一种故障类型

| | | |
|---|---|---|
| test_FD111.txt | 2022/1/26 19:40 | 文本文档 |
| test_FD222.txt | 2022/1/26 19:40 | 文本文档 |
| test_FD333.txt | 2022/1/26 19:40 | 文本文档 |
| test_FD444.txt | 2022/1/26 19:40 | 文本文档 |
| train_FD001.txt | 2022/1/26 19:41 | 文本文档 |
| train_FD002.txt | 2022/1/26 19:41 | 文本文档 |
| train_FD003.txt | 2022/1/26 19:41 | 文本文档 |
| train_FD004.txt | 2022/1/26 19:41 | 文本文档 |

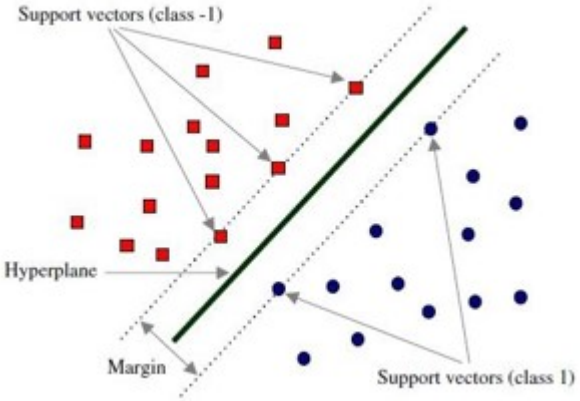
图六：数据集分成训练集与测试集

# 数据挖掘算法应用

## SVM算法

SVM算法是一种二值分类器方法，通过找到两类之间的一个线性可分的直线(或超平面)，将数据分类。

首先假设二分类目标是-1或者1，有许多条直线可以分割两类目标，但是定义分割两类目标有最大距离的直线为最佳线性分类器。




图七：SVM算法图示

### 相关参数

**·核函数（kernel）**：使非线性可分问题从原始特征空间映射到高维空间变得线性可分的非线性函数（poly核、RBF核、拉普拉斯核……）

**·度（degree）**：针对poly核的参数，默认为3，度越大则分割效果越明显，但会造成过拟合现象

**·惩罚系数（C）**：控制超平面附近错误分类的惩罚力度，C越大模型泛化能力越差

**·核函数幅宽（gamma）**：控制每个SVM的作用范围，gamma越大作用范围越小，对未知的样本分类效果越差

**·核函数常数项（coefo）**：poly核与sigmoid核使用的参数，默认为0

**·停止误差值（tol）**：控制训练程度，tol越小，精度越高

**·最大迭代次数（max_iter）**：默认-1（无限制），控制训练过程

主要调整参数为：核函数（kernel）、度（degree）、惩罚系数（C）、核函数幅宽（gamma）、核函数常数项（coefo）

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
  kernel='rbf', max_iter=-1, probability=True, random_state=None,
  shrinking=True, tol=0.001, verbose=False)
clf = SVC(probability=True)
clf.fit(trainX,trainY)
```

图七：sklearn库提供的SVC对象

## 分类结果

对正常状态发动机（标签0），和故障一状态发动机（标签1）数据进行二分类的结果。

```
训练集的向量数和特征数为：(2548, 19)
训练集标签的向量数和特征数为：(2548,)
预测准确度为： 0.8485915492957746
混淆矩阵为：
[[142   1]
 [ 42  99]]
```

图七：对正常状态发动机（标签0），和故障一状态发动机（标签1）数据进行二分类的结果

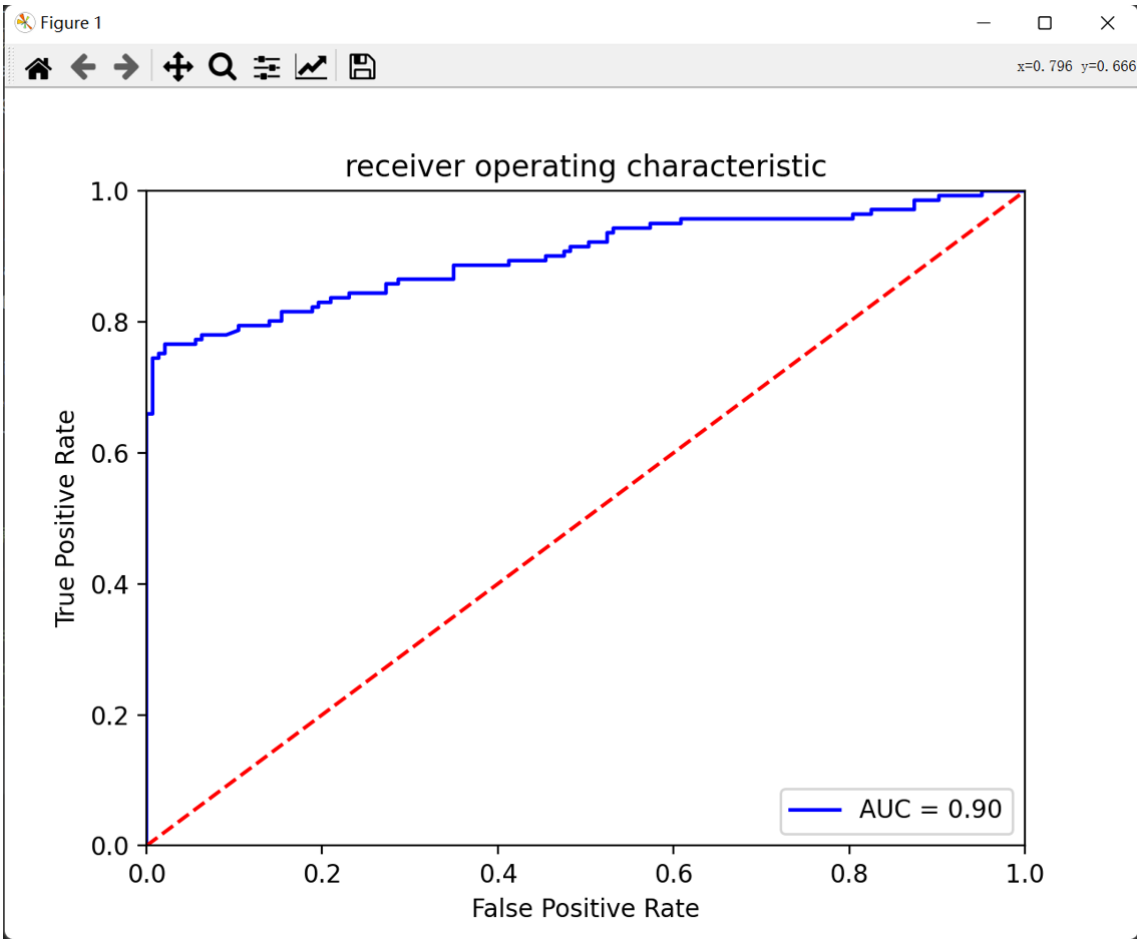
对分类结果的准确度进行进行统计：

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.77 | 0.99 | 0.87 | 143 |
| 1 | 0.99 | 0.70 | 0.82 | 141 |
| accuracy | | | 0.85 | 284 |
| macro avg | 0.88 | 0.85 | 0.85 | 284 |
| weighted avg | 0.88 | 0.85 | 0.85 | 284 |

图八：SVM分类结果准确度统计

- Precision：预测正确率
- Recall：召回率
- f1值：precision和recall的调和平均值
- Accuracy：正确预测的样本/总样本
- Macro avg：该项指标的平均值
- Weighted avg：该项指标的加权平均值

绘制ROC曲线并计算AUC，AUC 的全称是 Area under the Curve of ROC，也就是ROC曲线下方的面积。在机器学习领域，经常用 AUC 值来评价一个二分类模型的训练效果。




图九：AUC计算结果

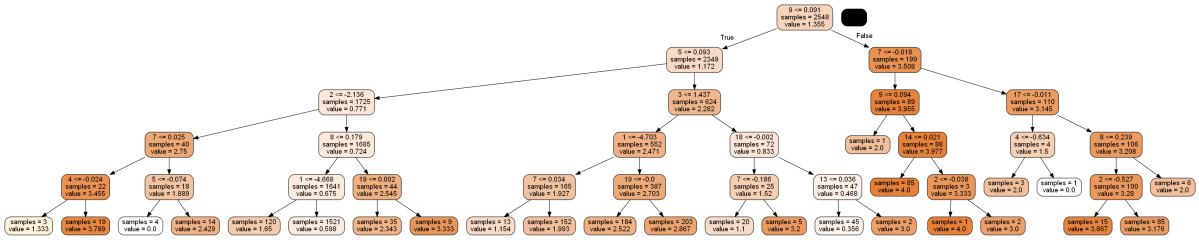
# 其他数据分类模型计算结果

## 朴素贝叶斯模型、决策树、随机森林

```
降维后的数据能够代表原数据的比例为：99.99997960635365%
训练集的向量数和特征数为：(2548, 19)
训练集标签的向量数和特征数为：(2548,)
预测的准确率为：67.61%
```

图十：朴素贝叶斯预测结果

```
降维后的数据能够代表原数据的比例为：99.99997960635356%
训练集的向量数和特征数为：(2548, 19)
训练集标签的向量数和特征数为：(2548,)
树深度为5预测的准确率为：58.06%
树深度为6预测的准确率为：56.90%
树深度为7预测的准确率为：60.22%
树深度为8预测的准确率为：57.82%
树深度为9预测的准确率为：58.00%
树深度为10预测的准确率为：56.85%
树深度为11预测的准确率为：56.04%
树深度为12预测的准确率为：47.26%
树深度为13预测的准确率为：53.82%
树深度为14预测的准确率为：48.38%
树深度为15预测的准确率为：45.02%
树深度为16预测的准确率为：47.10%
树深度为17预测的准确率为：40.56%
树深度为18预测的准确率为：31.63%
树深度为19预测的准确率为：31.89%
```



图十一：决策树预测结果

```
每个参数在数据中所占的信息比例为：
[8.13262180e-01 1.68634298e-01 1.03780884e-02 3.82340048e-03
 1.95140483e-03 9.52764565e-04 3.89290788e-04 2.79271758e-04
 1.86238447e-04 4.61269128e-05 3.69837982e-05 3.38937625e-05
 8.58851493e-06 6.45035480e-06 4.39630571e-06 2.78760728e-06
 1.98938296e-06 1.46410714e-06 1.77768994e-07]
降维后的数据能够代表原数据的比例为：99.99997960635358%
训练集的向量数和特征数为：(2548, 19)
训练集标签的向量数和特征数为：(2548,)
预测的准确率为：73.59%
```

图十二：随机森林预测结果