doi: 10.7690/bgzdh.2020.04.015

基于多种预测算法的飞机故障预测效果研究

朱兴动[1]，章思宇[2]，宋建华[3]
（1. 海军航空大学，山东 烟台 264001；2. 海军航空大学青岛校区研究生队，山东 青岛 266041；
3. 海军航空大学青岛校区，山东 青岛 266041）

摘要：为大幅提高飞机的维修故障预测精度，在充分研究 Fisher 判别法、逻辑回归、随机森林和支持向量机 4 种算法的基础上，使用某型飞机故障维修记录数据作为基础数据集，在 R 平台上实现这 4 种算法，以分析比较 4 种算法在故障预测上的效果差异。结果表明，支持向量机的预测效果最好。

关键词：故障预测；Fisher 判别法；逻辑回归；随机森林；支持向量机；Kappa 系数
中图分类号：TJ07　　**文献标志码：**A

Research on Aircraft Fault Prediction Effect Based on Various Prediction Algorithms

Zhu Xingdong[1], Zhang Siyu[2], Song Jianhua[3]
(1. *Navy Aviation University*, *Yantai* 264001, *China*;
2. *Brigade of Graduate*, *Qingdao Branch*, *Navy Aviation University*, *Qingdao* 266041, *China*;
3. *Qingdao Branch*, *Navy Aviation University*, *Qingdao* 266041, *China*)

Abstract: In order to greatly improve the accuracy of aircraft maintenance fault prediction, after sufficient studying on 4 algorithms, which include Fisher discriminant, logistic regression, random forest and support vector machine, using fault maintenance record data of certain type aircraft as basic dataset, and choosing R platform to implement these 4 algorithms. Compare the effect difference of 4 methods in fault prediction. Finally, the results show that support vector machine has the best prediction effect.

Keywords: fault prediction; Fisher discriminant method; logistic regression; random forest; support vector machine; Kappa coefficientm

0　引言

飞机的故障维修记录数据是研究飞机各个系统故障特性的重要数据基础，如果能够基于这些故障记录数据做出相关故障预测，对于飞机的维护工作将会有极大的帮助。现行的预测与分类算法种类繁多，适用场景也各不相同，针对某一预测问题选择最佳的算法，能够大幅度提高预测的精度。笔者在收集第一手数据资料的基础上，分析研究 Fisher 判别法、逻辑回归、随机森林、支持向量机 4 种比较常用的预测算法，并编写程序进行预测效果分析择优。

1　数据的预先处理

飞机的故障维修记录数据大都以自然语言叙述，且包含大量冗余数据序列，难以做到直接使用算法构建模型，需要对数据进行处理，删去一些不必要的数据列，并进行标准化操作。

删除冗余数据序列后的数据如表 1 所示。

数据标准化的方法采用的是将一列数据中，各个项目的出现频数按照从大到小排列，并从 1 开始赋值，而对于工作时间这类数值较大的数据，为防止数量级较大的数据在计算时覆盖其他数据的影响，需对其进行数量级的调整，调整的方式为：

$$C=\begin{cases}0, x=0\\ 1, 0<x\leqslant 500\\ 2, 500<x\leqslant 1\,000\\ 3, 1\,000<x\leqslant 1\,500\\ 4, 1\,500<x\leqslant 2\,000\\ 5, x>2\,000\end{cases}。\tag{1}$$

经过标准化处理后的数据如表 2 所示。

收稿日期：2019-12-05；**修回日期：**2020-01-11
作者简介：朱兴动(1967—)，男，海南人，博士，教授，从事装备保障信息化研究。E-mail: 1091252435@qq.com。

表 1　删除冗余数据序列后的数据

所属系统	故障部位	故障件翻修次数	故障件修后时次/h	专业	发动机修后工作时间/h	发现时机	故障件工作时次
操纵系统	其他	1	97.91	电气	0	飞行中	0
环控系统	座舱	0	0	飞机	0	定期检修	0
动力系统	发动机舱	0	150	特设	171.33	机械日	0
起落装置	起落架舱	0	421	特设	0	机械日	0

表 2　标准化处理后的数据

所属系统	故障部位	故障件翻修次数	故障件修后时次	专业	发动机修后工作时间	发现时机	故障件工作时次
操纵系统	5	1	1	10	0	7	0
环控系统	12	0	0	13	0	13	0
动力系统	9	0	1	15	1	14	0
起落装置	8	0	1	15	0	14	0

2　分类预测法研究

2.1　CART 算法与随机森林

分类与回归树（classification and regression tree，CART）算法是一种常用的数据挖掘分类预测算法，同时适用于离散型变量和连续型变量。若目标变量是离散的，则由算法生成的树就是分类树；若目标变量是连续的，则生成的树就是回归树。

CART 算法首先将数据升序排列，从小到大以相邻数字的中间值将样本分为 2 组，通过 Gini 系数计算 2 组样本中输出变量取值的异质性[1]：

$$G(t)=1-\sum_{j=1}^{K}p^2(j|t)。\tag{2}$$

其中：t 为结点；K 为输出变量的类别数；$P(j|t)$为结点 t 输出变量取 j 的概率，输出变量取值差异性越大，Gini 系数也越大。

CART 算法通过 Gini 系数的减少来描述输出异质性的下降

$$\Delta G(t)=G(t)-\frac{N_{\mathrm{r}}}{N}G(t_{\mathrm{r}})-\frac{N_{\mathrm{l}}}{N}G(t_{\mathrm{l}})。\tag{3}$$

其中：$G(t)$和 N 分别为分组前输出变量的 Gini 系数和样本量；$G(t_{\mathrm{r}})$、N_{r} 分别为分类后右子树的 Gini 系数和样本量；$G(t_{\mathrm{l}})$、N_{l} 分别为分类后左子树的 Gini 系数和样本量。通过这种方式，反复计算可以使异质性下降到最大的分割点，即 $\Delta G(t)$ 达到最小值为当前最佳分割点。

随机森林是一种组合分类方法[2]。它基于 Bootstrap 抽样原理，从原始数据集中随机抽取若干个训练样本子集。以无剪枝决策树为基分类器，对每个 Bootstrap 样本进行建模，并将生成的多个决策树集成，再利用所构建的决策树群对数据进行分类投票，最后由投票结果决定样本的最终分类或预测结果。

2.2　Fisher 判别法

Fisher 判别法是一种经典的分组判别法，基本思想是将高维的数据点投影到低维空间，使得数据点更加聚集。当分组数为 k 时，指标为 p 个，借助方差分析构造出 k 个判别函数，函数的通式为

$$F_i=\sum_{1}^{p}c_i x_i+C。\tag{4}$$

其中，确定参数 c_i 的原则是使得组间差距最大，组内差距最小。对于一个未分类的样本数据，将 p 个指标分别代入求出 F_i 值后，最大值对应的分组即为该样本所在组[3]。

2.3　逻辑回归

逻辑回归也被称为对数几率回归，采用回归的方式解决分类的问题。

逻辑回归用条件概率分布的形式表示 $P(Y|X)$，随机变量 $\boldsymbol{x}=\left(x^{(1)},x^{(2)},\cdots,x^{(n)}\right)$，是一个 n 维向量，Y 的取值为 0 或 1，即[4]：

$$P(Y=1|x)=\left(\exp(w\bullet x+b)\right)\Big/\left(1+\exp(w\bullet x+b)\right)；\tag{5}$$

$$P(Y=0|x)=1\Big/\left(1+\exp(w\bullet x+b)\right)。\tag{6}$$

或

$$\phi(x)=1\Big/\left(1+\mathrm{e}^{-w^{\mathrm{T}}x-b}\right)。\tag{7}$$

设一个二分类问题，输出为 $y=\{0,1\}$，线性回归模型 $z=w^{\mathrm{T}}x+b$ 的值为实数，为了便于分类，引入 Sigmoid 函数：

$$g(z)=1\Big/\left(1+\mathrm{e}^{-z}\right)。\tag{8}$$

使用 Sigmoid 函数将 z 的值转化至区间[0,1]内，由于其取值范围为[0,1]，就可以将计算所得结果认为是该点分为 1 类的概率。于是，就可以非常自然地约定将 Sigmoid 函数计算所得值大于等于 0.5 归

为类别 1，小于 0.5 归为类别 0。

2.4 支持向量机

支持向量机是一种有监督的建模方法，适用于小样本、非线性和高维数据。其本身是定义在特征空间上间隔最大化的线性二分类器，对于原本线性不可分的数据样本，通过引入核函数的方式，将原始样本空间映射到希尔伯特空间，从而使数据样本在特征空间中变得线性可分[5]。文中支持向量机模型使用的是径向积核函数。

对于线性可分的数据集，通过求解等价的凸二次规划问题，得到的分离超平面[6]为：

$$\boldsymbol{w}^{\mathrm{T}}x+b=0。\tag{9}$$

其中：x 为样本数据；w 为法向量；b 为截距。

相应的分类决策函数为：

$$f(x)=\mathrm{sgn}\left(\boldsymbol{w}^{\mathrm{T}}x+b\right)。\tag{10}$$

对于线性不可分的数据集，需要在每一个样本点引入一个松弛变量 $\xi_i\geqslant 0$，则线性不可分问题可变换为如下的优化问题：

$$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i \quad \text{s.t. } y_i\left(wx_i+b\right)\geqslant 1-\xi_i,$$

$$\xi_i\geqslant 0,\ i=1,2,\cdots,n。$$

其中：C 为正参数，取值为 C>0，一般由实际情况决定；$y_i\in\{+1,-1\}$，是 x_i 的类标记。

最后通过核函数与软间隔最大化，线性不可分的支持向量机分类决策问题可表示为：

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K\left(x,x_i\right)+b\right)。\tag{11}$$

其中：a_i 是拉格朗日乘子；$K\left(x,x_i\right)$ 为核函数。

3 多算法的飞机故障预测效果评估

3.1 数学模型建立

笔者使用 R 语言进行算法的数学建模。R 语言是一种专门为数据挖掘设计的语言，含有大量的功能扩展包，使算法编写的过程透明化，使用者只需要知道要用的数学模型位于哪一个程辑包中，使用过程中载入相应的程辑包并调用其中函数即可。

1) Fisher 判别法。

Fisher 判别法的建模需要在 R 中载入程辑包“MASS”，调用该程辑包中 lda 函数，核心代码如下：

```
lirary(MASS)
data<-read.csv("\data.csv")      \\读取数据
Fisher<−lda(System~.,data)
```

2) 逻辑回归。

在 R 中，有多种方式可以建立逻辑回归模型，文中使用“nnet”包中的 multinom 函数建立模型，核心代码如下：

```
library(nnet)
mandata<- data<-read.csv("\data.csv")
Logistic<-multinom(System~.,data=mandata,maxit=1 000)   \\设定逻辑回归的最大迭代次数为 1 000
```

3) 随机森林。

随机森林算法需要在 R 中加载程辑包“randomForest”，调用包中的 randomForest 函数。该函数的重要参数及意义如表 3 所示。

表 3 randomForest 函数参数及其意义

参数	意义
formula	确定自变量和因变量
data	使用的数据集
ntree	在森林中树的个数，默认为 500
mtry	每棵树使用的特征数
importance	是否计算变量的特征重要性，默认 FALSE
proximity	是否计算各个挂车之间的相似性

ntree 和 mtry 2 个参数需要经过实验来获得，实验的结果如表 4 所示。

表 4 模型参数拟合过程

ntree	mtry	模型拟合度	ntree	mtry	模型拟合度
2	500	0.787	3	500	0.824
2	1 000	0.788	3	1 000	0.826
2	1 500	0.791	3	1 500	0.827
2	2 000	0.791	3	2 000	0.826
4	500	0.826	5	500	0.816
4	1 000	0.825	5	1 000	0.816
4	1 500	0.826	5	1 500	0.818
4	2 000	0.826	5	2 000	0.816
6	500	0.812	7	500	0.807
6	1 000	0.810	7	1 000	0.809
6	1 500	0.811	7	1 500	0.809
6	2 000	0.811	7	2 000	0.809

经过实验确定 ntree 与 mtry 2 个参数分别为 3 和 1 500 时，模型的训练效果最好。

```
library(randomForest)
data<-read.csv("\data.csv")
RF<-randomForest(System~.,data,mtry=3,ntree=1 500)
```

4) 支持向量机。

支持向量机算法需要在 R 中加载程辑包“e1071”，并调用该包中的 svm 函数。由于支持向量机本身是一个二分类器，而实现多分类有 2 种方式：“一对多”和“一对一”方式[7]。

“一对多”方式：假设有 K 个类别，可以创建 K 个分类器。对于每个分类器，需要尝试将一个特定的类从其他类中区分出来。在选择用于判定的最佳类型时，则需要选择能使观测数据离分割超平面

尽可能远(与其他类距离最远)的类别。

“一对一”方式：首先为所有可能输出的类分别创建一个分类器，再使用这些分类对观测数据进行分类，并对每次胜出的类别进行累计，然后选取票数多的类别。在 e1071 包中的 svm 函数已经封装了多分类中的“一对一”方式，在数学建模过程中直接使用即可。

svm 函数的重要参数如表 5 所示。

表 5　svm 函数参数及其意义

参数	意义
formula	确定自变量和因变量
data	使用的数据集
type	分类：C-classification(default)/nu-classification；文本分类：one-classification；回归：eps-regression(default)/nu-regression
kernel	inear/polynomial/radial/sigmoid
cost	误差预算
gamma	控制相似度计算的局部性

对于核函数，建模中使用“radial”，cost 与 gamma 参数的确定需要通过实验来确定，测试的结果如表 6 所示。

表 6　模型参数拟合过程

cost	gamma	模型拟合度	cost	gamma	模型拟合度
0.01	0.01	0.399	0.1	0.01	0.447
0.01	0.05	0.419	0.1	0.05	0.502
0.01	0.10	0.435	0.1	0.10	0.568
0.01	0.50	0.441	0.1	0.50	0.634
0.01	1.00	0.424	0.1	1.00	0.674
1.00	0.01	0.487	10.0	0.01	0.608
1.00	0.05	0.611	10.0	0.05	0.753
1.00	0.10	0.694	10.0	0.10	0.791
1.00	0.50	0.813	10.0	0.50	0.879
1.00	1.00	0.850	10.0	1.00	0.912
100.00	0.01	0.704	100.0	10.00	0.954
100.00	0.05	0.808	100.0	50.00	0.960
100.00	0.10	0.844	100.0	100.00	0.962
100.00	0.50	0.919	100.0	500.00	0.964
100.00	1.00	0.931	100.0	1 000.00	0.964

从以上实验发现：随着 cost 与 gamma 2 个参数不断增加，模型的拟合度也不断增加并趋于稳定。进一步测试发现，再次增加 cost 与 gamma 后，最后拟合度稳定于 0.964，因此，可以选取合适的 cost=100，gamma=500，核心代码如下所示：

```
library(e1071)
data<-read.csv(“\data.csv”)
SVM<-svm(System~., data, kernel="radial", cost=100, gamma=500)
```

3.2　故障预测效果评估

分类器性能评估方法有多种，如对于二分类器常用的 ROC 曲线等[8]，但是对于多分类器的性能评估来说，ROC 曲线这一直观的方法就变得不再适用。对于上述 4 个算法模型的分类性能评价，笔者选择计算 Kappa 系数来进行评价。

Kappa 系数是比较 2 个观测者对同一事物或者同一名观察者对同一事物的 2 次观察结果是否一致，用由于偶然因素造成的异质性和实际观测的一致性之间差别大小作为评价基础所提出的统计指标[9]。

Kappa 系数计算是基于混淆矩阵的，其计算方法为：

$$k=(p_0-p_e)/1-p_e。\tag{12}$$

其中，p_0 表示总分类准确度，p_e 表示为

$$p_e=(a_1\times b_1+a_2\times b_2+\cdots+a_C\times b_C)/(n\times n)。\tag{13}$$

其中：a_i 表示第 i 类真实样本个数；b_i 代表第 i 类预测结果样本个数。Kappa 系数计算结果的取值为[−1,1]，一般情况下 Kappa 系数取值在 0～1，Kappa 值越接近 1，一致性越高，分类器的分类效果越好。

文中所用数据集是一个多分类数据集，类总数为 52。首先，使用训练所得的模型进行预测，并用预测结果与真实样本计算混淆矩阵，之后，使用 R 语言编写的 Kappa 值计算函数得出 4 个模型的 Kappa 值，计算结果如表 7 所示。

表 7　Kappa 值计算结果

参数	Fisher 判别法	随机森林	逻辑回归	支持向量机
p_0	0.170	0.527	0.191	0.664
p_e	0.079	0.063	0.078	0.059
Kappa	0.101	0.495	0.122	0.643

从上述计算结果可以看出：4 个模型中支持向量机的 Kappa 值最高，达到了 0.643；因此，可以得到结论，支持向量机对于本数据集来说，分类效果最佳。

4　结束语

笔者分析了 4 种常用的分类预测算法在飞机故障预测上的分类预测效果，在研究算法的基础上使用 R 语言进行了模型的训练与验证，针对多分类器的特点，在分类器性能评价量化指标中选择了 Kappa 值，分析对于文中飞机故障数据集的分类效果最佳的算法。实际计算验证发现：在 Fisher 判别法、随机森林、逻辑回归和支持向量机 4 种分类预测算法中，支持向量机的预测效果最好，综合预测正确率达到了 66.4%。由于飞机自身系统复杂，系统数目极多，若想提高预测的正确率，需要增加训练数据量。此外，预测算法数量众多，支持向量机是否为故障预测的最佳算法，或者还存在比支持向量机更加适合的算法，还需进一步的实验论证。

（下转第 78 页）

4 结论

根据仿真结果可以看出：经过优化后的缓冲器有效地降低了航炮后坐、前冲力和后坐前冲位移，使最大后坐、前冲行程之和小于 25 mm，最大后坐阻力小于 25 kN，达到了技术要求。连续射击 0.15 s 后，航炮的前冲运动被抵消掉，达到了浮动射击的效果。环簧—液压式缓冲器比环簧式缓冲器的缓冲效率高，能够更多地吸收后坐动能，缓冲效果更好。

在实际设计工作中，可以根据实际需要调整各部分结构尺寸。该思路可以为航炮缓冲器的设计提供一种新方法。

参考文献：

[1] 王永存, 周霖, 袁穗新, 等. 一种降低后坐力的火炮发射方法[J]. 火炮发射与控制学报, 2007(4): 10−12.

[2] 蒲玉成, 王惠源, 解志坚. 转管武器总体技术的若干问题[J]. 火炮发射与控制学报, 2005(1): 9−16.

[3] 郝秀萍, 蒲玉成, 王惠源. 超高射速自动机缓冲装置参数优化设计方法[J]. 火炮发射与控制学报, 2009(4): 32−33.

[4] 齐晓林. 航空自动武器[M]. 北京: 国防工业出版社, 2008: 63−75, 205−216.

[5] 余驰, 张钢峰, 杨超. 航炮射击炮振响应抑制特性分析[J]. 兵工自动化, 2019, 38(4): 20−23.

[6] 蒲玉成, 王惠源, 李强. 自动机结构设计[M]. 北京: 国防工业出版社, 2009: 312−329.

[7] 王月梅. 理论力学[M]. 北京: 兵器工业出版社, 1996: 44−48.

[8] 高跃飞. 火炮反后坐装置设计[M]. 北京: 国防工业出版社, 2010: 305−320.

[9] 张相炎. 火炮设计理论[M]. 北京: 北京理工大学出版社, 2005: 98−101.

[10] 姚养无. 火炮与自动武器动力学[M]. 北京: 兵器工业出版社, 2000: 71−77, 131−136.

[11] 郭竞尧, 刘彦, 李勇, 等. 某液压弹簧式浮动机仿真及优化[J]. 液压与气动, 2014(2): 85−87.

****

（上接第 49 页）

[2] 吴俊, 黄晓明, 林晨, 等. 基于业务流程管理的智能变电站配置文件全寿命周期管控工程化应用[J]. 电工技术, 2016(5): 81−82.

[3] 李翌辉, 史亚斌, 胡进寿, 等. 基于改进型遗传算法的复杂产品生产车间布局优化方法[J]. 兵工自动化, 2018, 37(5): 67−72.

[4] 郑世华, 张艳丽. 工作流程管理系统-BPM 的特点与应用分析[J]. 科技视界, 2015(1): 254−262.

[5] 王履华, 高权忠, 陈海华. GIS-BPM 引擎的设计与实现[J]. 测绘与空间地理信息, 2015, 38(1): 137−140.

[6] 陈立云, 罗均丽. 跟我们学建流程体系[M]. 北京: 中华工商联合出版社, 2014: 193−207.

****

（上接第 65 页）

参考文献：

[1] 史逸民, 史达伟, 郝玲, 等. 基于数据挖掘 CART 算法的区域夏季降水日数分类与预测模型研究[J]. 南京信息工程大学学报(自然科学版), 2018, 10(6): 760−765.

[2] 马海荣, 程新文. 一种处理非平衡数据集的优化随机森林分类方法[J]. 微电子学与计算机, 2018, 35(11): 28−32.

[3] 赵丽娜. Fisher 判别法的研究及应用[D]. 哈尔滨: 东北林业大学, 2013: 30−32.

[4] 陈春玲, 吴凡, 余瀚. 基于逻辑斯蒂回归的恶意请求分类识别模型[J]. 计算机技术与发展, 2019, 29(2): 124−128.

[5] 周苏, 胡哲, 文泽军. 基于 K 均值和支持向量机的燃料电池在线自适应故障诊断[J]. 同济大学学报(自然科学版), 2019, 47(2): 255−260.

[6] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法综述[J]. 电子科技大学学报, 2011, 40(1): 2−10.

[7] FORTE R M. 预测分析: R 语言实现[M]. 北京: 机械工业出版社, 2016: 122−123.

[8] 郝知远. 基于改进的支持向量机的股票预测方法[J]. 江苏科技大学学报(自然科学版), 2017, 31(3): 339−343.

[9] 郭轶斌, 郭威, 秦宇辰, 等. 基于 Kappa 系数的异质性检验及其软件实现[J]. 中国卫生统计, 2016, 33(1): 169−171.