一种基于不平衡数据的系统故障诊断方法研究

陈绍炜

(西北工业大学电子信息学院　陕西 西安 710072)

1. 针对不平衡数据进行故障诊断过程中，单一算法训练个体~~不同~~模型时会出现性能不稳定的现象，提出一种自适应不平衡分类算法（Adaptive Imbalance Classification，AIC）。该算法利用AUC值优化支持向量描述算法、随机森林算法和梯度提升树算法，生成分类子模型；通过模型融合的方法生成融合子模型；最后依据AUC值选择最优子模型，得到稳定的分类结果。结合美国故障预测与健康管理（Prognostic and Health Management，PHM）协会2015年公布的设备运行数据集进行算法验证，结果表明，在不平衡数据的故障诊断问题上，该算法相较单一的不平衡诊断算法，识别率更高，诊断效果更好，且性能更稳定。
2. 不平衡数据诊断　AUC　自适应　模型融合

中图分类号 TP3　　　　文献标识码 A 　　　　DOI:

Research on System Fault Diagnosis Method Based on Imbalanced Data

Chen Shaowei

(School of Electronics and Information, Northwestern Polytechnical University , Xi’an 710072, Shaanxi, China)

1. For the training of multiple models, the performance of single algorithm will be unstable, and an adaptive imbalance classification algorithm is proposed. The algorithm combines the AUC value to optimize the support vector description algorithm, the random forest algorithm and the gradient boosting decison tree algorithm respectively to generate the classification sub-model. Then combine the model fusion method to generate the fusion sub-model. Finally, select the excellent sub-model to get stable classification results based on the AUC value. This paper combines the equipment running data set published by the American PMM (Prognostic and Health Management) Society in 2015 to verify the algorithm. The results show that the fault recognition rate is higher than that of the single imbalance classification algorithm. The classification effect is good and the performance is stable.
2. Imbalanced classification AUC Adaptable Model fusion

# 引　言

近年来，随着我国科学技术的发展和工业化水平的提高，现代工业设备，如轨道车辆，航空航天对象等，正逐渐变得大型化、智能化以及复杂化。与此同时，设备的故障诊断和维护技术显得越来越重要。复杂设备一旦发生故障，轻则会导致停机或者停产，造成企业的经济损失，重则甚至会引发重大安全事故，造成人员的伤亡以及生态环境的破坏[1]。例如，2018年10月29日印尼狮航JT610航班坠入爪哇岛北部外海，造成189 人死亡。 官方事故报告显示飞机事发前曾发现6处故障。2017年全年，我国煤矿事故达219起，死亡人数共375人。故障诊断是保证设备安全可靠稳定运行的方法之一。

目前故障诊断系统大多需要丰富完备的历史数据， 结合机器学习的方法建立诊断模型，然而在实际的工业过程中，获取这样的数据往往需要很长的时间，尤其对复杂的工业设备而言，获取设备不同状态下的数据非常困难，常常存在故障类数据缺乏的问题。故障类数据缺乏的故障诊断问题属于不平衡数据诊断问题。

关于不平衡数据诊断问题的研究目前主要集中在两个方面[2]：一是数据层面，采用数据采样方法，调整原始数据的分布，降低不平衡数据集的失衡比例；二是算法层面，通过提出新的算法思想，如单分类学习(one class learning)、代价敏感学习(cost sensitive learning)等，或改善传统的诊断算法，解决不平衡数据的故障诊断问题。

收稿日期: xxxx-xx-xx。航空科学基金(No. 20155553039)。，硕士研究生，主研领域：基于人工智能的电子系统故障诊断。**陈绍炜**，副教授。

由前人已做的研究可知，数据层面比较常见的方法包括Chawla等人提出的人工小类采样技术[3]（Synthetic Minority Oversampling TEchnical, SMOTE）算法、Nickerson等人提出的一种有指导的上抽样方法[4]、Visa等人提出的一种集成了数据信息(如数据集分布、不平衡因子、类内和类间距离等)的上抽样方法[5]以及单边选择（One-Sided Selection， OSS）算法[6]和邻域清洗算法（Neighborhood cleaning rule,NCL）[7]等。算法层面提出的方法包括Veropoulos等人提出的基于代价敏感的SVM算法[8]、Fan等人提出的AdaCost算法[9]以及Raskutti和Kowalczyk提出的支持向量描述算法[10]（Support Vector Data Description，SVDD）和孤立森林[11]（Isolation Forest）算法等。

目前的研究方法从不同的角度解决不平衡故障诊断问题，然而它们却各有利弊，在实际工程应用中仍然存在很多问题。例如，数据层面的方法容易引入噪声和删除有用的信息；代码敏感学习会因为不合理的代价项而效果不理想等等。本文从模型融合的角度出发，提出一种自适应不平衡数据故障诊断算法（Adaptive Imbalance Classification，AIC），融合数据层面和算法层面上各个方法的优势，主动适应不平衡程度不同的数据集。

# 算法分析

## 支持向量描述算法

支持向量描述（Support Vector Data Description，SVDD）是一种单分类算法。它的基本思想是：通过非线性映射将原样本空间变换为新的高维特征空间，在高维空间寻找一个最优超球体，使之尽可能包含全部样本，并权衡最大包含样本数和最小球体半径[4]。

考虑训练数据集，以及假设存在非线性映射使数据样本从原空间变换到新的高维空间，即。SVDD算法寻找的超球体球心为、半径为，引入松弛变量的概念，因此，SVDD算法求解问题的数学描述如公式(1)所示，其中表示欧式距离，为惩罚参数。



s.t. (1)

and

构建拉格朗日函数并引入核函数，求解对偶问题后，问题可表示为公式(2)。



s.t. (2)

假设任一样本点到球心的距离函数记为，SVDD算法通过判断测试样本是否在超球体内来定义分类函数，如公式(3)所示。

 (3)

## 随机森林算法

随机森林[12]（Random Forest，RF）是Bagging方法的扩展。Bagging[13]方法的主要思想是：在训练数据集上，采用自助采样法（Bootstrap sampling），有放回的随机抽取次组成训练子集，然后在训练子集上训练子分类器模型，重复采样和训练次，得到个子分类器模型，最后采用投票法融合个子分类器模型，得到最终的分类器模型如公式(4)所示。

 (4)

扩展主要体现在两个地方，第一个是RF方法设定Bagging方法的子分类器模型是分类回归树（classification and Regression Trees，CART）；第二个是RF方法引入随机属性选择，RF方法在Bagging方法形成的训练子集上，从个属性特征中随机选取个不重复的属性形成新的训练子集。

## 梯度提升树算法

梯度提升树[14]（Gradient Boosting Decison Tree，GBDT）是Boosting方法的扩展。它采用决策树作为子分类器进行模型训练，属于提升树（Boosting Tree）模型。

假设损失函数是，提升树模型的优化问题可描述如公式(5)所示，其中表示数据集的样本数量。

 (5)

直接优化公式(1-6)所描述的问题，需要优化个决策树参数及其系数，优化问题过于复杂。提升树采用前向分步算法进行模型求解。前向分步算法从前到后，每一步只学习一个树模型及其系数，采用贪心算法思想逐步逼近优化最优解。假设第步需要求解的决策树模型参数是，模型系数是，第步的模型可表示为公式(6)。

 (6)

第步的优化问题可以描述为公式(7)：

 (7)

梯度提升树第步拟合步模型损失函数的负梯度，如公式(8)所示。第步模型损失函数的负梯度可近似等价于第步的预测结果与真值之间的残差。梯度提升树每迭代一步，都向残差减小的方向前进。

 (8)

# 自适应不平衡分类算法

自适应不平衡分类算法从模型融合的角度出发，利用不同的策略训练多个子模型，依据AUC值选择最优子模型，稳定子模型的分类结果，进而主动适应不平衡程度不同的数据集。自适应不平衡分类算法具体采用四种策略：

策略一：从新的算法思想角度，优化支持向量描述算法，生成子模型；

策略二：从改进现有算法的角度，优化随机森林算法，生成子模型；

策略三：从数据的角度，通过数据采样方法，减少健康类样本平衡不同类别样本的数量，结合梯度提升树算法生成子模型；

策略四：从模型融合的角度，利用加权平均法融合上述三种子模型生成子模型。

## 优化的SVDD

本文结合类别标签信息，利用性能评价指标AUC值，优化SVDD算法的决策边界。

AUC指标需要对概率输出值进行计算，然而，SVDD算法输出的是离散量，如公式(3)所示，AUC值无法直接计算。因此，本文采用Sigmoid函数修正SVDD算法的输出量，如公式(9)所示。修正后的输出量可以视为样本处于正常状态时的概率预测值。

 (9)

优化的SVDD算法的关键步骤如下：

步骤1：初始化SVDD算法参数；

步骤2：训练超球体模型，结合Sigmoid函数获得SVDD算法的概率输出值；

步骤3：计算模型在数据集上的AUC值；

步骤4：利用差分进化算法迭代优化，寻找最优的算法参数，使模型在数据集上的AUC值最大。

## 优化的RF

本文着重提高RF子分类器的准确性，间接提高RF算法的准确性。

首先，本文结合“分层抽样”的思想，将原始训练集按照样本类别划分为多数类数据集和少数类数据集，针对多数类数据集和少数类数据集同时有放回的重采样若干次，形成“多数类子集”和“少数类子集”。优化后的训练子集由“多数类子集”和“少数类子集”组成。本文通过“分层抽样”的方式平衡训练子集样本比例，提高了少数类样本参与子分类器训练的概率，降低了因少数类样本的规模带来的对子分类器“准确性”的影响。

其次，本文优化子分类器的整体准确性。本文通过删除“准确性”不高的子分类器，提高子分类器整体的准确性。本文选择AUC值评价子分类器的“准确性”，通过设置子分类器的AUC值阈值，删除低于此阈值的子分类器。

优化的RF算法的关键步骤如下：

步骤1：初始化RF算法参数，如树的数量、树的深度、树的最大特征数等；

步骤2：结合分层抽样方法随机生成100颗CART子树；

步骤3：计算CART子树的AUC值，利用30%分位数，确定AUC值的阈值参数；

步骤4：利用分层抽样方法生成训练CART子树，计算CART子树的AUC值，若小于阈值，则删除此CART子树；

步骤5：重复生成CART子树，直至满足停止条件；

步骤6：投票法融合多个子树模型；

步骤7：利用差分进化算法迭代优化，寻找最优的算法参数，使模型在数据集上的AUC值最大。

## 优化的GBDT

GBDT算法的优化主要集中在数据层面。本文使用数据采样方法减少多数类样本的数量迫使训练数据集不同类别的样本数量基本均衡，然后利用梯度提升树完成模型训练。

本文采用邻域清洗算法（Neighborhood cleaning rule, NCL）对多数类样本进行欠采样。NCL算法结合样本的分布和结构特性，有效地保留多数类样本的信息，比传统的下抽样法更精确。

本文定义数据集的不平衡度为多数类样本数量与少数类样本数量的比值。优化的GBDT算法的关键步骤如下：

步骤1：多次运行NCL算法，直至数据集的不平衡度满足要求；

步骤2：利用差分进化算法迭代优化，寻找梯度提升树最优模型。

## 算法流程

结合上述分析，自适应不平衡分类算法框图如图1所示。整个算法执行流程为：



图1 自适应不平衡分类算法框图

步骤1：算法结合OVR（One Vs Rest）方法将不平衡分类问题分解为个二分类问题，状态的二分类数据集表示为。

步骤2：在数据集上采用优化的SVDD算法训练子模型。

步骤3：在数据集上采用优化的RF算法训练子模型。

步骤4：在数据集上结合NCL算法和GBDT算法训练子模型。

步骤5：采用加权平均法融合子模型、和，输出融合模型。

步骤6：利用交叉验证集，确定状态的最终模型。

步骤7：遍历个设备状态的数据集，，重复步骤（2）~（6），获取个最优模型，。

步骤8：对于未知样本，遍历个最优模型，输出设备状态为，的概率值，概率值最大的设备状态为未知样本的对应状态。

# 实验结果与分析

数据集选用美国PHM协会2015年公布的设备运行数据集，数据集包含60个设备从2009年到2012年的运行数据。设备运行数据分为2大类，第一类是运行信息，主要是设备的运行状态信息；第二类是故障信息，主要是该设备运行过程中故障记录信息，包括故障开始时间、故障结束时间和故障类型。

数据集的问题是：数据集包含60组设备运行数据，其中有20组设备的故障信息是不完整的，具体表现为设备的数据集合部分故障信息丢失，数据集的目标是正确预测出这20组设备丢失的故障信息。

## 数据集分析

数据集的目标是正确预测出20组设备丢失的故障信息，主要包括故障开始时间、故障结束时间和故障类型。本文将预测故障信息的问题转化为分类问题。具体策略是判断当前时刻是否是故障类型的故障开始时刻以及故障结束时刻。由于预测故障开始时刻的问题是不平衡分类问题，且预测不同设备的不同故障类型的故障开始时刻的问题是不平衡程度不同的数据集，因此本文重点讨论故障开始时刻的预测问题。

故障的产生可能会引起设备运行信息的变化，因此，利用设备运行信息训练分类模型判断当前时刻是否发生故障是合理的。然而，故障的开始时间体现的是设备状态由健康状态向故障状态的一种变化，单一时刻的设备运行信息无法充分表达设备运行状态的变化。因此，本文引入时间窗口的概念。时间窗口是指一段时间内的设备运行数据，单一时刻的数据无法体现设备状态的变化，但是一段时间的数据却可以捕获这种变化。

时间窗口的引入使得分类器有能力捕获设备状态的变化，然而也增加了数据的维度。高维度的数据不仅会带来计算量的急剧上升，甚至会引发维度灾难，导致分类器的性能会显著下降。本文采用PCA方法进行降维。

因此，数据处理的关键步骤是：

步骤1：对原设备运行矩阵矩阵，采用时间窗口方法，扩展数据维度，变为新矩阵；

步骤2：采用PCA方法对矩阵进行降维，保留95%的数据方差，即得最终的训练矩阵。

## 算法评价指标

本文选择公式(10)作为算法性能评价指标,定义为GCS。符号意味着非故障开始时间被正确标识；符号表示故障开始时间被正确标识；符号意味着非故障开始时间被错误标识为故障开始时间；符号表示故障开始时间被错误标识为非故障开始时间。

 (10)

## 结果分析

本文在预测故障开始时间的问题上比较了本文改进的相关算法如RF算法、GBDT算法和SVDD算法，以及本文提出AIC算法之间的差异。表1是各个算法在40组设备的所有故障上，针对预测故障开始时间问题的性能差异比较。表1从GCS均值和GCS方差两个维度对比了改进的GBDT算法、改进的RF算法、改进的SVDD算法和AIC算法的差异。

表1故障开始问题上各算法性能差异比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 改进的GBDT | 改进的  RF | 改进的SVDD | AIC |
| GCS均值 | 1436.14 | 2153.43 | 1864.93 | 2343.07 |
| GCS方差 | 2533473 | 2212145 | 1897230 | 1176700 |
| TNR均值 | 54.1% | 64.9% | 60.4% | 68.3% |
| 准确率均值 | 84.1% | 91.4% | 86.7% | 89.2% |

对比分析本文提出的AIC算法和改进的算法，我们可以发现，相比其余算法，AIC算法综合性能最优，它的TNR均值和GCS均值最高，且它的GCS方差低于三种改进的算法，以及准确率均值性能较优。究其原因在于，40组设备的失衡程度差异大，单一算法性能不稳定，导致单一算法GCS方差偏高，然而AIC算法结合了三种改进算法的优点，依据AUC值选择最优子模型，稳定了模型的分类结果。因此，本文认为AIC算法可以有效的解决不平衡数据的故障分类问题，故障诊断效果更好，故障识别率更高，且性能更稳定。

# 结　语

本文提出一种稳定性能好、可以适应不同不平衡程度数据集的分类方法，即自适应不平衡分类算法。本文结合美国PHM协会2015年公布的设备运行数据集分析故障开始时间的预测问题。结果表明，本文提出的AIC算法，相较单一的不平衡分类算法，故障诊断效果更好，故障识别率更高，性能更稳定。

参 考 文 献

1. 王景, 任燕. 电气设备故障诊断方法和技术[J]. 自动化与仪器仪表, 2015(5):22-24.
2. 叶枫, 丁锋. 不平衡数据分类研究及其应用[J]. 计算机应用与软件, 2018(1):132-136.
3. 张永, 李卓然, 刘小丹. 基于主动学习SMOTE的非均衡数据分类[J]. 计算机应用与软件, 2012, 29(3):91-93.
4. Hu S , Liang Y , Ma L , et al. MSMOTE: Improving Classification Performance When Training Data is Imbalanced[C] Second International Workshop on Computer Science & Engineering. IEEE, 2010.
5. Dang X T , Tran D H . SPY : a novel resampling method for improving classification performance in imbalanced data[C] Seventh International Conference on Knowledge & Systems Engineering. IEEE, 2015.
6. Jia C, Zuo Y. S-SulfPred: A sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique[J]. Journal of Theoretical Biology, 2017, 422:84-89.
7. Abdouli N O A, Aung Z, Wei L W, et al. Tackling Class Imbalance Problem in Binary Classification using Augmented Neighborhood Cleaning Algorithm[M] Information Science and Applications. 2015.
8. 张鹏翔, 刘利民, 马志强. 基于MapReduce的层叠分组并行SVM算法研究[J]. 计算机应用与软件, 2015(3):172-176.
9. Masnadishirazi H, Vasconcelos N. Cost-Sensitive Boosting[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 33(2):294-309.
10. 花小朋, 兰少华. 一种SVDD增量学习算法及应用[J]. 计算机应用与软件, 2009, 26(9):237-239.
11. Liu F T, Kai M T, Zhou Z H. Isolation Forest[C] Eighth IEEE International Conference on Data Mining. 2009.
12. 孙雪, 韩蕾, 李昆仑,等. 基于类别特征选择与反馈学习随机森林算法的邮件过滤系统研究[J]. 计算机应用与软件, 2015(4):67-71.
13. Bifet A, Holmes G, Pfahringer B. Leveraging Bagging for Evolving Data Streams[M] Machine Learning and Knowledge Discovery in Databases. 2010.
14. Son J. Tracking-by-Segmentation with Online Gradient Boosting Decision Tree[C] IEEE International Conference on Computer Vision. 2016.

本文负责人：

联系电话：

邮箱：

作者简介：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 姓名 | 职称/学位 | 主要研究领域 | 身份证号 | 手机号 | 单位及通信地址 | 邮编 | E-mail |
|  |  |  |  |  |  |  |  |
| 陈绍炜 | 副教授 | 航空电子系统设计与PHM技术 | 610103197006033656 | 18509207961 | 西北工业大学电子信息学院 | 710129 | cgong@nwpu.edu.cn |