计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):239−246 [doi: 10.15888/j.cnki.csa.008585]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 K-means Bayes 和 AdaBoost-SVM 的故障分类①

黄子扬, 周凌柯

(南京理工大学 自动化学院, 南京 210014)
通信作者: 周凌柯, E-mail: lingke_zhou@163.com

摘　要: 传统的故障分类方法大多假设不同类别的数据样本量是相似或相等的. 然而在实际的工业过程中采集到的数据多数是正常数据, 少部分是故障数据, 这就造成了数据的不平衡. 针对不平衡数据问题, 本文提出了一种 K-means Bayes 与 AdaBoost-SVM 相结合的故障分类方法, 通过设计两种独立的分类器, 并利用 D-S 证据理论对分类结果融合, 以弥补各自对某些类别分类能力较弱的缺陷. 实验证明, 本文提出的故障分类方法与单一 Bayes 或 SVM 比较, 具有更高的分类准确率.

关键词: 故障分类; 不平衡数据; K-means Bayes; AdaBoost-SVM; 证据融合; 机器学习

引用格式: 黄子扬,周凌柯.基于 K-means Bayes 和 AdaBoost-SVM 的故障分类.计算机系统应用,2022,31(7):239–246. http://www.c-s-a.org.cn/1003-3254/8585.html

Fault Classification Based on K-means Bayes and AdaBoost-SVM

HUANG Zi-Yang, ZHOU Ling-Ke

(School of Automation, Nanjing University of Science and Technology, Nanjing 210014, China)

Abstract: Traditional fault classification methods mostly assume similar or equal sample sizes for different types of data. However, the bulk of data collected in the actual industrial process is normal with a minority belonging to fault data, which causes data imbalance. Aiming at the imbalanced data, this study proposes the fault classification method combining K-means Bayes with AdaBoost-SVM. Two independent classifiers are designed with the D-S evidence theory to merge the classification results, so as to make up for their weak classification capabilities for certain categories. Experiments show that the fault classification method proposed in this study has higher classification accuracy than single Bayes or SVM.

Key words: fault classification; imbalanced data; K-means Bayes; AdaBoost-SVM; evidence fusion; machine learning

随着社会的发展和科技的进步, 化工过程和设备变得更加复杂与多样化, 故障诊断也成为了当代过程监控中的一项重要任务[1,2]. 近年来, 由于信息技术的快速发展, 使得工业过程中的数据可以大量采集并保存, 因此基于数据驱动的故障诊断相关方法得到了学者们的广泛关注[3,4]. 除了基于数据驱动的方法外, 基于数学模型的方法及基于知识的方法也都是常用方法. 然而基于数学模型的方法存在着模型建立难、诊断的结果直接受模型准确性影响等问题, 因此在复杂系统中该类方法的使用受到限制; 基于知识的方法则需要大量的经验和专家知识来建立知识库, 通用性差. 而基于数据驱动的方法直接对数据进行分析, 能够规避上述两类方法存在的问题, 因此目前在复杂系统中使用更多[5–7].

故障分类技术在数据挖掘和机器学习领域得到快速发展, 但是大多数的分类方法均假设其训练集中的

① 基金项目: 国家重点研发计划 (51405-01B02)
收稿时间: 2021-10-14; 修改时间: 2021-11-08; 采用时间: 2021-11-26; csa 在线出版时间: 2022-05-30

各类样本数相似或相等, 当训练集呈现数据不平衡特征时, 分类性能通常并不令人满意. 张剑飞等[8] 指出不平衡数据的分类问题广泛出现在疾病诊断、垃圾邮件处理、信用卡检测等领域, 但传统的机器学习算法在数据不平衡比过大时, 分类效果会急剧下降. Japkowicz 等[9] 指出利用决策树处理不平衡数据时, 训练过程会被多数类的样本主导, 导致对少数类的样本识别率低. 为了提高对不平衡数据的分类性能, 目前常采用的方法主要包括两个方面: 数据层面和算法层面[10,11]. 在数据层面, 为降低数据的不平衡比, 通常会采取欠采样或过采样的方法改变训练样本的数量. Liu 等[12] 提出了一种 EasyEnsemble 的欠采样方法, 该方法通过从多数类样本中有放回的随机采样出 n 个每个子集, 使子集的样本数与少数类近似相等, 并将各子集分别与少数类样本合并进行训练, 从而达到保留多数类样本信息的目的. Chawla 等[13] 提出了一种 SMOTE (synthetic minority oversampling technique) 的过采样方法, 通过对少数类中的某个样本及其邻近样本进行叠加, 产生人造样本来降低样本间的不平衡度. 虽然上述方法对部分场景下的不平衡数据问题具有一定的效果, 但是它们仍有一些不足之处需要改进. 由此, 更复杂的重采样技术被提出. 张天翼等[14] 提出了一种改进 SMOTE 的重采样方法, 通过将合成样本从一维空间扩展至更高维空间, 使新样本更加多样化. 李忠智等[15] 结合卷积神经网络和生成对抗网络, 利用卷积神经网络从故障样本中提取训练特征后输入至对抗网络, 并由解码器网络来生成新的故障样本. Chen 等[16] 提出了一种 K-means Bayes 方法, 利用 T 阈值 K-means 在既不减少多数类样本也不增加少数类样本的前提下, 提高对少数类故障的识别能力. 在算法层面, 通常是对现有的分类算法进行修改, 以增强对少数类的学习能力. 如代价敏感学习、集成算法等.

Bayes 和 SVM 是故障诊断领域常用的两种方法. Lemnaru 等[17] 指出 Bayes 和 SVM 使用的前提是不同类型样本的数据量近似相等, 当数据不平衡严重时, 这两种方法通常会表现出较差的分类性能. Zhang 等[18] 将 D-S 证据理论应用于多分类器实现故障监控, 有效提高了分类性能. 本文主要研究数据不平衡的算法层面, 同时考虑到单一方法在这种数据不平衡条件下的局限性, 提出了一种基于多分类器融合的故障分类方法. 选择 K-means Bayes 作为分类模型 1, AdaBoost-SVM 作为分类模型 2, 并利用 D-S 证据理论将二者的分类结果进行融合, 进一步提升分类性能. 将该方法运用在 Tennessee Eastman (TE) 数据集上, 经仿真和实验证明了所提方法的有效性及可行性.

1 K-means Bayes 算法

1.1 Naive Bayes

Naive Bayes 是一种基于贝叶斯定理和条件独立性假设的分类方法[19]. 给定一组训练数据 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$, 其中, N 为训练样本总数, $x_i=\{x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\}^{\mathrm{T}}$ 为每一个样本数据, $y_i=\{C_1,C_2,\cdots,C_k\}$ 为各样本数据对应的标签, 对于一个测试样本 x, Bayes 分类器将后验概率 $P(Y=C_k|X=x)$ 最大的类作为 x 的类输出:

$$P(Y=C_k|X=x)=\frac{P(X=x|Y=C_k)P(Y=C_k)}{P(X=x)} \tag{1}$$

依据条件独立性假设, 且每个连续变量 $x^{(j)}$ 均服从高斯分布:

$$P(X^{(j)}=x^{(j)}|Y=C_k)\sim N(\mu_{C_k,j},\sigma^2_{C_k,j}) \tag{2}$$

对于样本 x 的分类结果为:

$$y=\arg\max_{C_k} P(Y=C_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=C_k) \tag{3}$$

1.2 K-means Bayes

K-means Bayes 算法[16] 的思想为: 在不改变原始数据集信息的情况下降低数据不平衡性对故障分类带来的影响. 算法步骤如算法 1.

算法 1. K-means 对多数类均分

1) 给定多数类样本 $X=[x_1,x_2,\cdots,x_m]$, 标准化得到 X^*;
2) 指定 k 个聚类子集: 从 X^* 中随机选择 k 个样本点作为初始聚类中心 μ, 并计算各子集的样本数 $T=m/k$;
3) 设置 2 个集合 $U=\{U_1,\cdots,U_k\}$ 和 $U^*=\{U_1^*,\cdots,U_k^*\}$, 其中 U_i 存放原始样本数据, U_i^* 存放标准化后的数据;
4) 对 X^* 的各样本 x^*, 计算与各聚类中心的距离, 找出距离最近的 μ_i 及对应的 U_i^*;
5) 判断 U_i^* 的样本数 n_i 是否 $\leqslant T$. 若 $n_i\leqslant T$, 则将 x^* 和对应的 x 分别分配至 U_i^* 和 U_i, 并转至 4) 对下一个样本进行计算; 若 $n_i\geqslant T$, 则该样本将不再分至 U_i, 并转至 4) 重新计算;
6) 当所有样本完成分类后, 计算各子集 U_i^* 的样本均值, 并作为新的聚类中心 μ_i', 并判断各 μ_i 与 μ_i' 是否相等; 若 $\mu_i=\mu_i'$ 则算法终止输出结果 U_i, 否则返回 3) 进行下一轮计算.

对于测试样本 x, 可利用式 (3) 预测其类别, 再利用

式 (4) 转换成实际类别:

$$y_{\text{real}} = \begin{cases} 1, & y \in [1,k] \\ y-k+1, & y \in [k+1,k+c] \end{cases} \tag{4}$$

2 AdaBoost-SVM 算法

2.1 SVM

SVM (support vector machine)[20] 是一种有监督的二分类模型, 其思想是寻找到一个分离超平面, 此超平面不仅能正确划分开正负实例点, 还能使离超平面最近的点 (支持向量) 离超平面尽可能的远. 给定一组线性可分的训练数据集 $T=\{(x_1,y_1),\cdots,(x_N,y_N)\}$, 其中 $x_i \in X \in R^n, y_i \in Y \in \{+1,-1\}$, 则分离超平面为:

$$wx+b=0 \tag{5}$$

分类决策函数为:

$$f(x)=\text{sign}(wx+b) \tag{6}$$

为解决多分类问题, 通常将多分类问题进行拆分并利用投票机制进行分类. 常用的拆分策略包括“一对多”和“一对一”, 本文使用“一对一”策略, 这样由少数类构成的子分类器的正类和负类可看成是平衡的, 有利于提高分类性能. 由于可能存在子分类器对某些类的分类能力较差, 影响最终的投票结果, 因此引入 AdaBoost 分类器替换[21,22].

2.2 AdaBoost

AdaBoost (adapative boost)[23] 是提升学习 Boosting 里的一种, 其思想是通过反复学习得到一系列的弱分类器, 并将这些弱分类器进行加权组合得到一个强分类器.

算法 2. AdaBoost-SVM

1) 设置每个子分类器的最低分类正确率 $Acc_{\min}$;
2) 根据类别总数 n 构建 SVM 子分类器, 其中子分类器总数为 $N=C_n^2$;
3) 对各 SVM 子分类器进行训练, 其中核函数选择径向基核函数 (RBF), 并利用网格法和三折交叉验证法进行参数寻优, 选择合适的惩罚参数 C 及核参数 g;
4) 对训练好的分类器在相应的测试集上进行分类, 计算各子分类器的实际正确率 Acc_{real};
5) 若某个分类器的 $Acc_{\text{real}}<Acc_{\min}$, 则使用 AdaBoost 分类器对其进行替换. 初始化 AdaBoost 的弱分类器个数为 10 个, 各样本的初始化权重 $w=1/Count_{\text{样本}}$;
6) 使用相同的测试集计算 AdaBoost 分类器的分类正确率 Acc^*_{real}, 若 $Acc^*_{\text{real}} \geqslant Acc_{\text{real}}$, 则选择替换, 否则不替换.

3 决策融合算法

3.1 故障分类框架

如果能设计出一种在任何情况下都具有良好泛化性能的分类器, 那么单一分类器就已经能够满足需要. 然而实际采集到的数据存在着噪声点、异常等问题, 使得上述的单一分类器难以实现. 因此考虑创建一个故障分类系统, 该系统中存在两个及以上的分类器, 并希望其分类性能优于其中任意单一分类器. 当某个分类器在识别时发生错误, 其他分类器可以纠正该分类器的错误.

图 1 展示了一种基于 D-S 证据理论的故障分类框架. 该框架主要分为两个部分: 利用训练数据进行离线建模、利用建立好的模型对数据进行在线分类. 具体来说, 本系统的实施主要包括 3 个步骤: 1) 分类器构建; 2) 计算各分类器的融合矩阵; 3) 利用 D-S 证据理论进行决策融合.

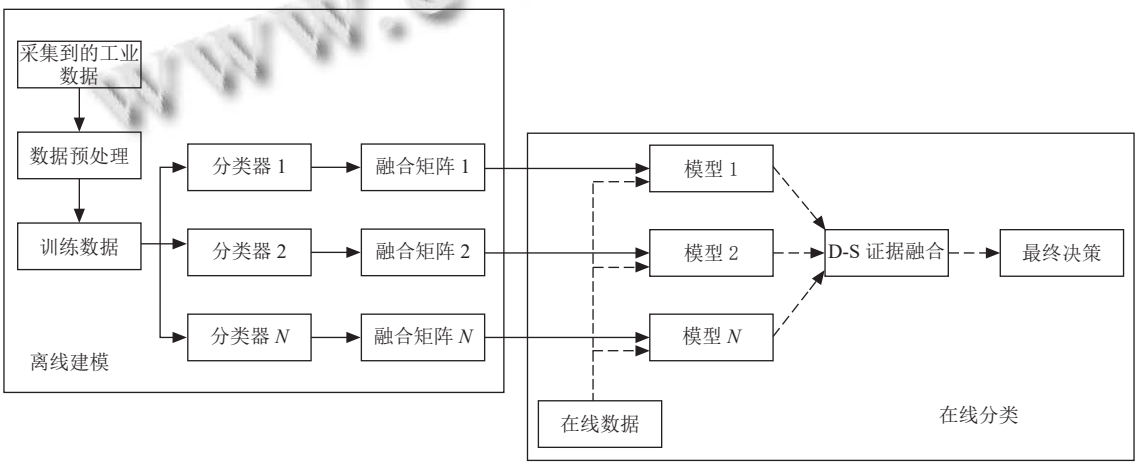


图 1 基于 D-S 证据理论的故障分类框架

3.2 计算各分类器的融合矩阵

为了进行 D-S 证据理论融合, 应计算出各分类器的融合矩阵[18]. 假设样本类别集合$T=\{F_1,F_2,\cdots,F_n\}$, 其中第i个类别称为F_i, $i=1,2,\cdots,n$, n为所有类别的总数. 分类器总个数为N, 其中第k个分类器的融合矩阵表示为FM^k, $k=1,2,\cdots,N$. 则FM^k可表示如式 (7). 其中FM^k的行代表真实类别$F_1,F_2,\cdots,F_n$, 列代表由该分类器预测出的类别$F_1,F_2,\cdots,F_n$, 元素N_{ij}^k表示由该分类器预测类别为F_j而真实类别为F_i的样本数之和. 因此对于每个分类器的融合矩阵FM^k而言, 矩阵的每列之和为定值 1.

$$FM^k=\begin{bmatrix} \dfrac{N_{11}^k}{\sum_{i=1}^{n}N_{i1}^k} & \dfrac{N_{12}^k}{\sum_{i=1}^{n}N_{i2}^k} & \cdots & \dfrac{N_{1n}^k}{\sum_{i=1}^{n}N_{in}^k} \\ \dfrac{N_{21}^k}{\sum_{i=1}^{n}N_{i1}^k} & \dfrac{N_{22}^k}{\sum_{i=1}^{n}N_{i2}^k} & \cdots & \dfrac{N_{2n}^k}{\sum_{i=1}^{n}N_{in}^k} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{N_{c1}^k}{\sum_{i=1}^{n}N_{i1}^k} & \dfrac{N_{c2}^k}{\sum_{i=1}^{n}N_{i2}^k} & \cdots & \dfrac{N_{nn}^k}{\sum_{i=1}^{n}N_{in}^k} \end{bmatrix} \tag{7}$$

3.3 基于 D-S 证据理论的决策融合

1) 计算分类器对某个样本x的预测类别为F_j时对应F_i的基本概率分配 (BPA):

$$m_k(F_i)=\frac{N_{ij}^k}{\sum_{i=1}^{n}N_{ij}^k} \tag{8}$$

2) 依据 D-S 融合规则计算联合 BPA 值:

$$\begin{cases} BPA(A)=\dfrac{1}{K}\displaystyle\sum_{A_1\cap\cdots\cap A_n=A} m_1(A_1)\cdots m_n(A_n) \\ K=\displaystyle\sum_{A_1\cap\cdots\cap A_N\neq\Phi} m_1(A_1)m_2(A_2)\cdots m_n(A_n) \end{cases} \tag{9}$$

3) 选择最大的联合 BPA 值所对应的类别F_i作为最终决策:

$$Final_{DS}=\arg\max_{i\in[1,n]}[m_{1,2,\cdots,N}(F_i)] \tag{10}$$

4 仿真实验

4.1 TE 过程

本文所提到算法均以 TE 过程数据为基础. TE 过程由伊斯曼化学公司所创建, 该仿真模型在真实化工过程基础上构建[24]. 其工艺流图如图 2. 该过程通过 4 种气态反应物 (A、C、D、E) 和惰性成分 B 生成产品 G、H 及副产品 F. TE 数据可由开源的 Simulink 代码生成, 数据集共包括 41 个测量变量和 11 个控制变量, 数据除正常类型外还包括 21 种不同类型的故障, 本文所使用的故障类型如表 1. 部分仿真结果如图 3.

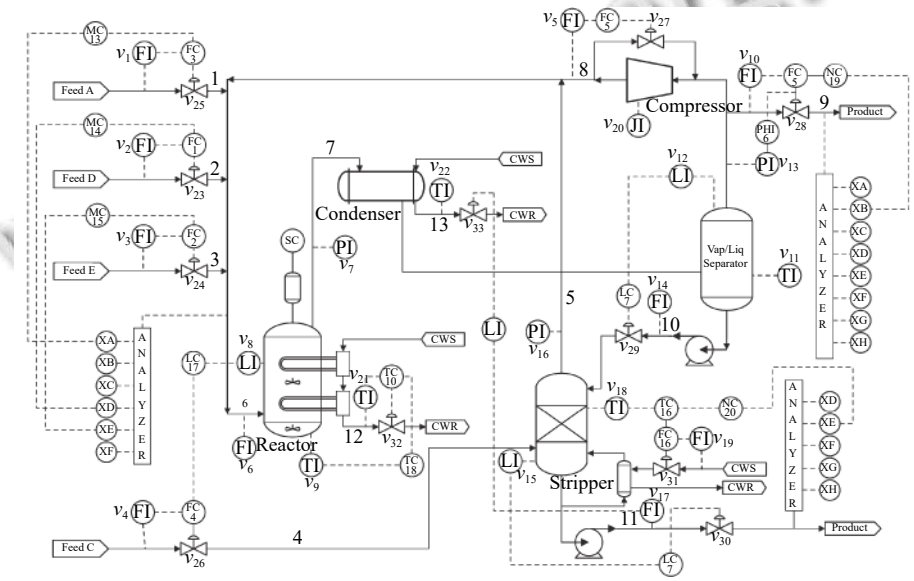
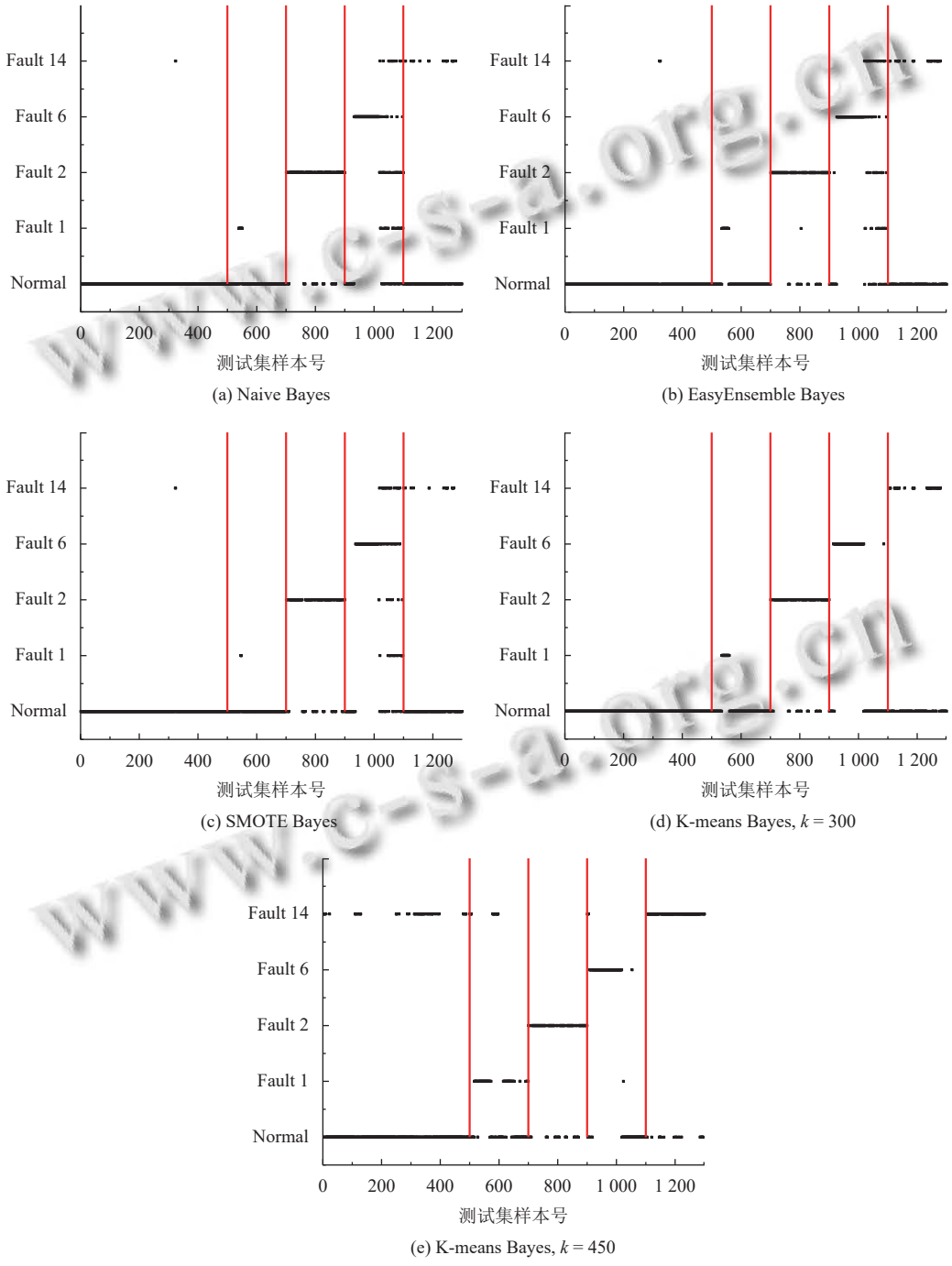


图 2 TE 过程流程图

表 1 TE 过程故障

故障序号	故障描述	故障类型	训练样本	测试样本
0	正常状态	无	900	500
1	A/C进料率变化	阶跃	10	200
2	B进料变化	阶跃	10	200
6	A进料损失	阶跃	10	200
14	反应器冷却水阀门	失灵	10	200



(a) Naive Bayes

(b) EasyEnsemble Bayes

(c) SMOTE Bayes

(d) K-means Bayes, k = 300

(e) K-means Bayes, k = 450

图 3 Bayes 相关算法仿真结果

4.2 仿真结果及分析

图 3(a) 为利用 Naive Bayes 进行故障分类的结果, 测试集的分类正确率为 62.1%, 可以看出该方法在这种不平衡数据下的分类能力较差; 图 3(b) 为利用 EasyEnsemble Bayes 的分类结果, 通过 EasyEnsemble将正常类样本有放回的抽取 15 组, 每组 10 个样本并分别与故障样本组合, 最终的分类正确率为 65.4%, 与 Naive Bayes 相比略有提升, 但故障 1 和故障 14 的识别率依然较低, 说明 EasyEnsemble 方法并不能完全解决数据稀缺性带来的问题; 图 3(c) 为利用 SMOTE Bayes 的分类结果, 通过 SMOTE 为每个少数类增加 10 个合成样本, 在一定程度上弥补了少数类样本的稀缺性. 但利用这种方式对测试集的分类结果不理想, 甚至低于 Naive Bayes, 且经实验仿真发现利用 SMOTE分别为每个少数类依次增加 20 个、30 个、40 个样本时, 预测的准确率也几乎没有提升, 其原因可能在于所使用的部分训练样本本身处于所在样本集的分布边缘, 则由此及其相邻样本产生的人造样本也会处于这个边缘, 且会越来越边缘化, 从而使分类更加的困难; 图 4(d)、图 4(e) 为利用 K-means Bayes 的分类结果, 当分类子集数 k=450 时, 与前几种方法相比正确率得到显著提升, 预测的准确率达到了 76.0%, 但对故障 1、6 的分类能力仍存在一定的缺陷.

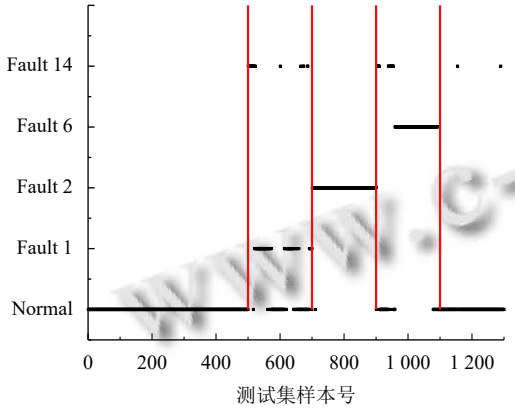


图 4 SVM 仿真结果

SVM 是一种经典的二分类学习算法, 在使用 SVM 之前, 为解决多分类问题, 本文选择“一对一”策略对多分类问题进行拆分, 如表 2 所示. 通过构建 10 个子分类器, 对测试集样本进行分类预测, 正确率为 70%, 如图 4 所示, 且对故障 1 和故障 14 的分类能力差. 通过分析各 SVM 子分类器的分类性能, 发现 C1、C4、C6、C7 及 C10 这 5 个子分类器对对应测试样本的分类能力差.

表 2 SVM 子分类器

正类	负类	序号	正类	负类	序号
正常	故障1	C1	故障1	故障6	C6
正常	故障2	C2	故障1	故障14	C7
正常	故障6	C3	故障2	故障6	C8
正常	故障14	C4	故障2	故障14	C9
故障1	故障2	C5	故障6	故障14	C10

为克服上述 5 个子分类器分类能力差的问题, 本文利用 AdaBoost 算法构建 5 个强分类器, 分别替代原 SVM 的 C1、C4、C6、C7 及 C10 再重新进行预测, 如图 5 所示. AdaBoost-SVM 的最终预测正确率为 80.7%, 除故障 14 外其余类型样本预测均比较准确, 其原因为利用 AdaBoost 构建的 C4 分类器依然无法正确识别故障 14 类型的样本.

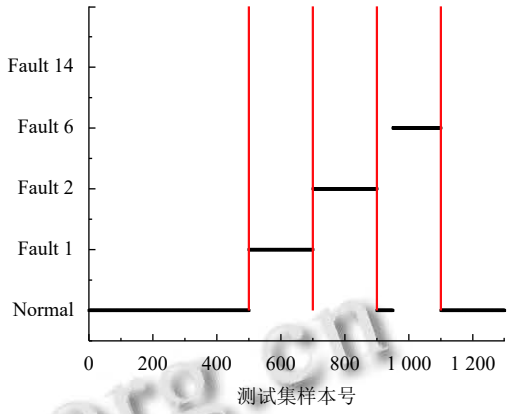


图 5 AdaBoost-SVM 仿真结果

为进一步提高分类性能, 使用本文提出的决策融合算法, 选择 K-means Bayes 和 AdaBoost-SVM 的预测结果作为证据体计算融合矩阵, 并进行 D-S 融合, 分类正确率达到 93.1%, 结果如图 6、图 7 所示.

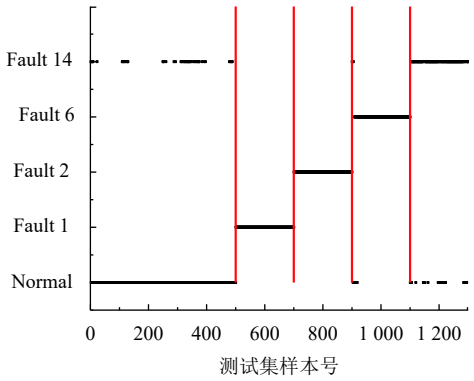


图 6 D-S 融合仿真结果

	0	1	2	6	14
0	0.656	0	0	0	0.209
1	0.126	0.990	0	0	0.080
2	0.051	0	1	0	0
6	0.127	0.010	0	1	0.016
14	0.040	0	0	0	0.695

(a) K-means Bayes

	0	1	2	6
0	0.667	0	0	0
1	0	1	0	0
2	0	0	1	0
6	0.067	0	0	1
14	0.266	0	0	0

(b) AdaBoost-SVM

图 7 两种方法的融合矩阵

在上述 2 个融合矩阵中, 矩阵的行表示数据的真实标签分类, 列表示利用模型的预测分类, 矩阵的每一列元素和为 1. 由于 AdaBoost-SVM 对所有测试样本均未分类到故障 14, 故融合矩阵中缺少预测为故障 14 的列. 表 3 展示了故障分类的部分信息融合过程, 根据各分类器对某个测试样本的预测结果, 选择对应的融合矩阵中的数据实现数据融合.

表 3 故障分类部分信息融合

真实类别	K-means Bayes	AdaBoost-SVM	联合BPA	决策
正常	正常 $m(0)=0.656$ $m(1)=0.126$ $m(2)=0.051$ $m(6)=0.127$ $m(14)=0.040$	正常 $m(0)=0.667$ $m(6)=0.067$ $m(14)=0.266$	$m(0)=0.958$ $m(6)=0.019$ $m(14)=0.023$	正常
故障1	正常 $m(0)=0.656$ $m(1)=0.126$ $m(2)=0.051$ $m(6)=0.127$ $m(14)=0.040$	故障1 $m(1)=1$	$m(1)=1$	故障1
故障14	故障14 $m(0)=0.209$ $m(1)=0.080$ $m(6)=0.016$ $m(14)=0.695$	正常 $m(0)=0.667$ $m(6)=0.067$ $m(14)=0.266$	$m(0)=0.428$ $m(6)=0.004$ $m(14)=0.568$	故障14

5 结论与展望

本文针对不平衡数据的故障分类问题, 分别提出了 K-means Bayes 和 AdaBoost-SVM 的分类策略. 利用 K-means 对多数类的样本划分为 K 个子集, 在不丢失多数类样本信息的前提下降低了不平衡度, 提高了 Bayes 的分类准确率; 利用 AdaBoost 对分类能力较差的 SVM 子分类器进行替换, 提高了 SVM 的分类准确率; 再利用 D-S 证据理论对二者的预测结果进行融合, 得到更好的分类结果. 由基于 TE 过程的仿真结果可知, 本文提出的决策融合算法与单一的传统算法相比具有更好的故障分类性能.

参考文献

1 Han YM, Ding N, Geng ZQ, *et al*. An optimized long short-term memory network based fault diagnosis model for chemical processes. Journal of Process Control, 2020, 92: 161–168. [doi: 10.1016/j.jprocont.2020.06.005]

2 Van Impe J, Gins G. An extensive reference dataset for fault detection and identification in batch processes. Chemometrics and Intelligent Laboratory Systems, 2015, 148: 20–31. [doi: 10.1016/j.chemolab.2015.08.019]

3 Thomas MC, Zhu WB, Romagnoli JA. Data mining and clustering in chemical process databases for monitoring and knowledge discovery. Journal of Process Control, 2018, 67: 160–175. [doi: 10.1016/j.jprocont.2017.02.006]

4 张妮, 车立志, 吴小进. 基于数据驱动的故障诊断技术研究现状及展望. 计算机科学, 2017, 44(6A): 37–42. [doi: 10.11896/j.issn.1002-137X.2017.6A.008]

5 Zhu JL, Yao Y, Li DW, *et al*. Monitoring big process data of industrial plants with multiple operating modes based on Hadoop. Journal of the Taiwan Institute of Chemical Engineers, 2018, 91: 10–21. [doi: 10.1016/j.jtice.2018.05.020]

6 Corona F, Mulas M, Baratti R, *et al*. On the topological modeling and analysis of industrial process data using the SOM. Computers & Chemical Engineering, 2010, 34(12): 2022–2032. [doi: 10.1016/j.compchemeng.2010.07.002]

7 周东华, 胡艳艳. 动态系统的故障诊断技术. 自动化学报, 2009, 35(6): 748–758. [doi: 10.3724/SP.J.1004.2009.00748]

8 张剑飞, 王真, 崔文升, 等. 一种基于 SVM 的不平衡数据分类方法研究. 东北师大学报 (自然科学版), 2020, 52(3): 96–104. [doi: 10.16163/j.cnki.22-1123/n.2020.03.014]

9 Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent Data Analysis, 2002, 6(5): 429–449. [doi: 10.3233/ida-2002-6504]

10 Zhuo Y, Ge ZQ. Gaussian discriminative analysis aided GAN for imbalanced big data augmentation and fault classification. Journal of Process Control, 2020, 92: 271–287.

[doi: 10.1016/j.jprocont.2020.06.014]

11 董宏成, 文志云, 万玉辉, 等. 基于 DPC 聚类重采样结合 ELM 的不平衡数据分类算法. 计算机工程与科学, 2021, 43(10): 1856–1863. [doi: 10.3969/j.issn.1007-130X.2021.10.020]

12 Liu XY, Wu JX, Zhou ZH. Exploratory under-sampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 539–550. [doi: 10.1109/TSMCB.2008.2007853]

13 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357. [doi: 10.1613/jair.953]

14 张天翼, 丁立新. 一种基于 SMOTE 的不平衡数据集重采样方法. 计算机应用与软件, 2021, 38(9): 273–279. [doi: 10.3969/j.issn.1000-386x.2021.09.043]

15 李忠智, 尹航, 左剑凯, 等. 不平衡训练数据下的基于生成对抗网络的轴承故障诊断. 小型微型计算机系统, 2021, 42(1): 46–51. [doi: 10.3969/j.issn.1000-1220.2021.01.009]

16 Chen GC, Liu Y, Ge ZQ. K-means Bayes algorithm for imbalanced fault classification and big data application. Journal of Process Control, 2019, 81: 54–64. [doi: 10.1016/j.jprocont.2019.06.011]

17 Lemnaru C, Potolea R. Imbalanced classification problems: Systematic study, issues and best practices. Proceedings of the 13th International Conference on Enterprise Information Systems. Beijing: Springer, 2012. 35–50. [doi: 10.1007/978-3-642-29958-2_3]

18 Zhang FY, Ge ZQ. Decision fusion systems for fault detection and identification in industrial processes. Journal of Process Control, 2015, 31: 45–54. [doi: 10.1016/j.jprocont.2015.04.004]

19 张新华. 基于 ICA 独立成分和加权依赖贝叶斯的传感器节点故障诊断. 重庆师范大学学报 (自然科学版), 2015, 32(2): 138–142. [doi: 10.11721/cqnuj20150231]

20 张志政, 王冬捷, 张勇亮. 基于 PSO 改进 KPCA-SVM 的故障监测和诊断方法研究. 现代制造工程, 2020, (9): 101–107. [doi: 10.16731/j.cnki.1671-3133.2020.09.015]

21 降爱莲, 杨兴彤. 基于 AdaBoost-SVM 级联分类器的行人检测. 计算机工程与设计, 2013, 34(7): 2547–2550, 2565. [doi: 10.16208/j.issn1000-7024.2013.07.050]

22 曹惠玲, 高升, 薛鹏. 基于多分类 AdaBoost 的航空发动机故障诊断. 北京航空航天大学学报, 2018, 44(9): 1818–1825. [doi: 10.13700/j.bh.1001-5965.2017.0774]

23 Rätsch G, Onoda T, Müller KR. Soft margins for AdaBoost. Machine Learning, 2001, 42(3): 287–320. [doi: 10.1023/A:1007618119488]

24 Robertso G, Thomas MC, Romagnoli JA. Topological preservation techniques for nonlinear process monitoring. Computers & Chemical Engineering, 2015, 76: 1–16. [doi: 10.1016/j.compchemeng.2015.02.002]

(校对责编: 孙君艳)