DOI: 10.7641/CTA.2017.60349

基于EasyEnsemble的化工过程故障诊断性能改进

夏丽莎[1†], 杨玉英[1], 方华京[2]

(1. 上海理工大学 管理学院, 上海 200093; 2. 华中科技大学 自动化学院, 湖北 武汉 430074)

摘要: 化工过程故障诊断中样本数据分布不均衡现象普遍存在. 在使用不均衡样本作为训练集建立各类故障诊断分类器时, 易出现分类器的识别率偏置于多数类样本的结果, 由此产生虽正常状态易识别, 但更受关注的故障状态却难以被诊断的现象. 针对该问题, 本文提出一种基于EasyEnsemble思想的主元分析–支持向量机(EasyEnsemble based principle component analysis–support vector machine, EEPS)故障诊断算法, 通过欠采样方法抽取多数类样本子集组建多个新的均衡数据样本集, 使用主元分析(principle component analysis, PCA)进行特征提取并使用支持向量机(support vector machine, SVM) 算法进行训练, 得到多个基于SVM的故障诊断分类器, 然后使用Adaboost算法集成最终的分类, 从而提高故障诊断准确性. 所提方法被用于TE(Tenessee Eastman)化工过程, 实验结果表明, EEPS算法能够有效提高分类器在不均衡数据集上的诊断性能和预报能力.

关键词: 化工过程; 数据不均衡; EasyEnsemble; 故障诊断

中图分类号: TP277 **文献标识码**: A

Fault diagnosis performance improvement for chemical process based on EasyEnsemble method

XIA Li-sha[1†], YANG Yu-ying[1], FANG Hua-jing[2]

(1. School of Business, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. School of Automation, Huazhong University of Science and Technology, Hubei Wuhan 430074, China)

Abstract: Imbalanced dataset is a phenomenon existing massively in the field of chemical process fault diagnosis. The recognition rate of the classifier will be biased to the majority class samples when using imbalanced dataset as the training set. As a result, the normal state is easy to identify, while the fault state people concerned are difficult to be diagnosed. In this paper, an EasyEnsemble based principle component analysis-support vector machine (EEPS) fault diagnosis algorithm is proposed. After constructing a number of balanced subsets by under-sampling from the majority class, principle component analysis (PCA) is used for feature extraction and a number of support vector machine (SVM) sub-classifiers are trained accordingly. Then an integral classifier is developed by using the Adaboost algorithm. This integral classifier can be used for fault diagnosis and prognosis. The experimental results on Tenessee Eastman (TE) chemical process show that the proposed EEPS improves the diagnosis and prognosis performance on the imbalanced dataset.

Key words: chemical process; imbalanced dataset; EasyEnsemble; fault diagnosis

1 引言(Introduction)

化工生产涉及人们衣食住行和国家工业、农业、国防等各个领域, 对国民经济和人们的日常生活有着举足轻重的作用. 由于化工过程系统结构复杂, 具有高度非线性和时变性, 故障的潜伏往往不可避免[1]. 而一旦发生故障, 可能会引发一系列连锁反应, 导致整个生产过程不能正常工作, 甚至引起财产损失、人员损失和环境污染等严重后果. 因此, 对运行状态进行准确、有效地早期预警和故障诊断, 对化工过程的安全生产起着十分重要的作用.

由于化工过程的精确数学模型存在过于复杂、难以建立和难以获得等问题, 随着计算机技术和人工智能的迅速发展, 各种数据驱动的智能故障诊断技术受到越来越多的关注, 如主元分析(principle component analysis, PCA)、K–最近邻(K–nearest neighbor, KNN)[2]、人工神经网络(artificial neural network, ANN)、粗糙集(rough set, RS)、隐马尔可夫(hidden markov model, HMM)、支持向量机(support vector

收稿日期: 2016−05−26; 录用日期: 2016−10−18.

†通信作者. E-mail: lisaxss@163.com; Tel.: +86 15821207879.

本文责任编委: 周东华.

上海高校青年教师培养资助计划(10–15–303–808)资助.

Supported by Foundation for Young Teachers in Shanghai Higher Education Institutions (10–15–303–808).

machine, SVM)、符号有向图 (signed directed graph, SDG)等[3]. 虽然这些数据驱动的方法各有特点, 但是它们都基于一个共同的出发点, 就是以采集的数据为基础, 通过各种技术挖掘数据中的隐含信息进行故障诊断[4]. 然而在数据采集时, 数据驱动的故障诊断方法往往会面临样本数据不均衡现象, 即获得的样本数据集中, 正常运行的样本数据较多, 故障样本的数量相对较少[5]. 一般认为当少数类与多数类的类分布比例低于1:2时, 数据具有分布不均衡特征. 由于在训练分类器过程中, 通常以提高整体分类准确率为目标, 因此在使用这样的不均衡样本作为训练集, 建立各类故障诊断分类器时, 易出现故障诊断分类器的识别率偏置于多数类样本(正常状态)的结果[6]. 尽管此时多数类(正常状态)样本识别正确率处于可接受的范围, 但是少数类(故障状态)样本识别正确率极低. 然而在化工过程状态监控和故障诊断中, 少数类(故障状态)才恰是人们关注的重点. 过程监控和故障诊断的核心任务之一就是及时、有效地识别这些少数类(故障状态). 因此, 在设计故障诊断分类器时, 人们希望少数类样本能被更多地识别出来. 事实上, 这也正是数据驱动的故障诊断方法在实际工程应用中的局限性所在. 如何利用这种不均衡的数据样本获得更好的诊断效果, 是当前数据驱动的故障诊断技术研究中共同面临并急需解决的问题.

样本数据不均衡现象对故障诊断带来的困难和挑战, 其中最主要的是故障诊断分类器性能的降低[7]. 为解决此问题, 本文提出一种基于EasyEnsemble思想的故障诊断算法, 通过抽取多数类样本子集组建多个新的均衡数据样本集并使用SVM算法进行训练, 得到多个基于SVM的故障诊断分类器, 然后使用Adaboost算法集成最终的分类. 考虑到特征选择可以提高分类器的性能, 因此在训练SVM故障诊断分类器前先对原始数据进行特征选择, 提出了基于EasyEnsemble思想的主元分析支持向量机(EasyEnsemble based PCA–SVM, EEPS)故障诊断算法.

2 基于EasyEnsemble思想的PCA–SVM故障诊断算法基本原理(Algorithm for EasyEnsemble based PCA–SVM fault diagnosis)

2.1 EasyEnsemble分类器(EasyEnsemble classifier)

EasyEnsemble分类器是一种下采样算法, 它独立地从多数类样本集中随机抽取多个样本子集, 随机抽取样本的数量与少数类的样本数量一致, 然后将抽取的样本子集分别和少数类样本集组成, 形成多个新的样本集, 用于训练多个分类器. 所有的分类器通过Bagging算法集成为最终的分类器[8]. 算法的流程如下:

输入: 训练数据集$S_r = \{(x, y)\}$, 欠采样次数T;

输出: 集成模型N.

1) 开始.

2) for $k = 1 : T$.

3) 通过Bootstrap算法从多数类样本集(正例样本集) S_r^+中得到一个子集S_{rk}^+, S_{rk}^+的样本数量和少数类样本集(负例样本集) S_r^-相同.

4) 在训练子集$S_{rk}^+ \cup S_r^-$上通过Adaboost算法训练个体模型N_k:

$$N_k(x) = \mathrm{sgn}\{\sum_{j=1}^{n_k} \alpha_{k,j} h_{k,j}(x) - \theta_k\},$$

其中: $h_{k,j}$为N_k第j个弱分类器, $\alpha_{k,j}$为$h_{k,j}$的权重, θ_k为训练子集$S_{rk}^+ \cup S_r^-$的实际类别集.

5) end for.

6) 集成模型N:

$$N(x) = \mathrm{sgn}\{\sum_{k=1}^{T} \sum_{j=1}^{n_k} \alpha_{k,j} h_{k,j}(x) - \sum_{k=1}^{T} \theta_k\}.$$

7) 结束.

2.2 基于EasyEnsemble的PCA–SVM故障诊断(EasyEnsemble based PCA–SVM fault diagnosis)

特征提取是指将原始特征转换为一组具有明显物理意义或者统计意义的特征. 其目的是为了从原始特征中找出最有效(同类样本的不变性、不同样本的鉴别性、对噪声的鲁棒性)的特征, 从而减少数据存储和输入数据带宽、减少数据冗余和噪声特征, 实现在低纬上提高故障诊断性能和减少计算复杂度的效果. 在化工过程中, 温度、压强和液位等各特征之间存在复杂相关性, 因此本文提出一种EEPS故障诊断算法, 详细的算法描述如下:

1) 使用PCA方法对实时监控得到的不均衡训练数据样本集进行特征提取, 通过寻找一种最能够代表原始数据的线性变换的方式, 将原始高维样本数据投影到新的正交低维子空间, 从而达到降噪和去冗余的效果;

2) 根据数据对应的化工过程状态, 将特征提取后的训练数据新子空间划分为正例样本集S_r^+和负例样本集S_r^-, 其中正例样本集对应为化工过程处于正常状态, 为数据中的多数类, 负例样本集对应为化工过程处于故障状态, 为数据中的少数类;

3) 基于EasyEnsemble思想对正例样本集S_r^+进行T次Bootstrap欠采样, 得到T个正例样本子集S_{rk}^+ ($k = 1, 2, \cdots, T$), 其中S_{rk}^+的样本数量和负例样本集S_r^-的样本数量相同. 将T个正例样本子集分别与负例样本集S_r^-组合, 得到T个新的训练子集$S_{rk}^+ \cup S_r^-$ ($k = 1, 2, \cdots, T$);

4) 通过Adaboost算法对T个新训练子集进行SVM训练, 得到多个SVM故障诊断子分类器. 进行bagging集成, 得到最终的EEPS故障诊断分类器;

5) 结束.

为评测此基于EasyEnsemble思想的PCA–SVM故障诊断分类器性能, 选择适用于不均衡数据的评价标准, 如分类准确率ACC, F–measure, AUC等进行评估. 将经过由PCA特征提取后的测试数据集用于基于EasyEnsemble思想的PCA–SVM故障诊断分类器, 得到故障诊断结果, 并进一步进行性能评估. 图1为如上所基于EasyEnsemble的PCA–SVM(EEPS)故障诊断框架.

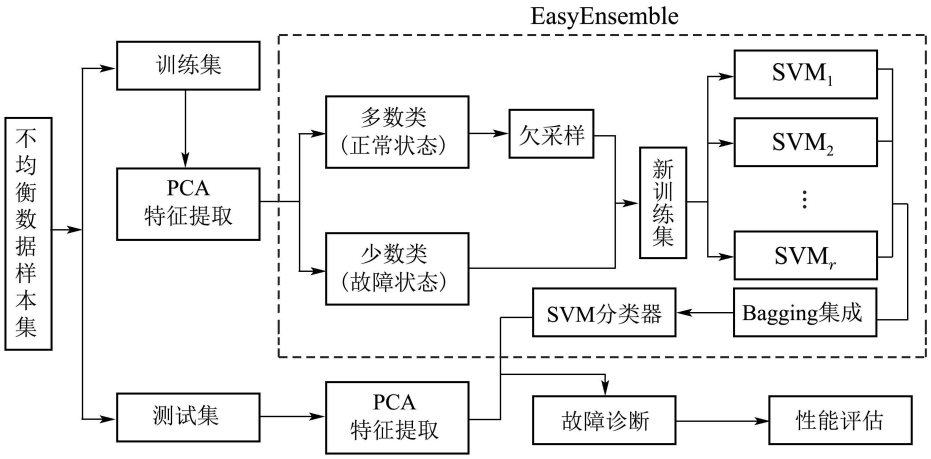

图 1 基于EasyEnsemble的PCA–SVM(EEPS)故障诊断框架

Fig. 1 A framework for the proposed EasyEnsemble based PCA–SVM (EEPS) fault diagnosis approach

3 TE化工过程故障诊断实验(Experiment on the TE chemical process)

Tenessee Eastman (TE)过程是由美国Eastman化学公司开发的一种高级仿真工业过程模型, 用于开发、研究和评价过程控制技术和监控方法, 大量文献引用它作为数据源, 来进行控制、优化、过程监控、故障诊断等研究. TE过程包含41个测量变量, 分别为22个连续测量变量和19个成分测量值, 涉及变量多、相互关联性强. 过程可以预设21个不同的故障类型, 具体参见文献[9].

3.1 数据准备和实验参数设置(Data preparation and experimental parameter setting)

针对TE过程故障存在的情况, 为对应20种不同的故障类型, 共设计20次不同的仿真实验. 设定每次仿真时间为48 h, 采样间隔为3 min, 设定于仿真32 h引入故障. 因此, 在除故障类型6之外的其余19种故障类型情况下, 每次仿真均共采集961条数据, 其中前720个为正常状态下的数据, 后241个为引入故障后的数据. 对于故障类型6, 仿真过程中发生由于汽提液液面过低, 超过预设的下限, 引起系统停工状况, 因此共采集得到864条数据, 其中前720个为正常状态下的数据, 后144个为从引入故障到系统停工期间的数据. 数据中少数类与多数类的类分布比例均低于1:2, 分布具有不均衡特征.

整个TE过程包含41个测量变量, 在使用PCA进行特征提取时, 设定使得累积贡献率达85%以上为应保留的主元个数; 设定对正例样本集S_r^+欠采样的次数为$T=7$, 因此对应每种故障类型, 通过欠抽样均可得到7个正例样本子集个数; 设定RBF核函数为训练SVM故障诊断子分类器时的核函数.

3.2 算法评价方法(Evaluation criteria for EEPS)

实验选用分类准确率ACC和F–measure(包括正类F–measure: TF–measure, TFM和负类F–measure: NF–measure, NFM)作为主要评价指标, 同时列出其他4个指标: 正查全率T–Recall(即正常状态诊断准确率)、负查全率N–Recall(即故障状态诊断准确率)、正查准率T–Precision和负查准率N–Precision. 它们的定义分别如下:

$$\text{T–Recall} = \text{TP}/\left(\text{TP}+\text{FN}\right), \tag{1}$$

$$\text{N–Recall} = \text{TN}/\left(\text{FP}+\text{TN}\right), \tag{2}$$

$$\text{T–Precision} = \text{TP}/\left(\text{TP}+\text{FP}\right), \tag{3}$$

$$\text{N–Precision} = \text{TN}/\left(\text{FN}+\text{TN}\right), \tag{4}$$

$$\text{ACC} = \left(\text{TP}+\text{TN}\right)/\left(\text{TP}+\text{TN}+\text{FP}+\text{FN}\right), \tag{5}$$

$$\text{TF–measure} = \frac{2\times\text{T–Recall}\times\text{T–Precision}}{\text{T–Recall}+\text{T–Precision}}, \tag{6}$$

$$\text{NF–measure} = \frac{2\times\text{N–Recall}\times\text{N–Precision}}{\text{N–Recall}+\text{N–Precision}}, \tag{7}$$

其中TP, FN, FP和TN为表1所示分类问题混淆矩阵的元素. 当查全率和查准率都比较高时, 说明故障诊断分类器性能比较好; F–measure的值越大, 故障诊断分类器性能越好.

表 1 混淆矩阵

Table 1 The confusion matrix

	预测为正类样本	预测为负类样本
实际为正类样本	TP	FN
实际为负类样本	FP	TN

3.3 实验结果分析(Experimental result analysis)

实验以ACC, TFM和NFM为主要评价指标, 使用本文提出的EEPS故障诊断算法, 对TE过程的20种不同故障类型分别进行故障诊断. 同时结合未基于EasyEnsemble思想的PCA–SVM(PS)故障诊断算法进行比较. 改进前后的TE过程诊断性能见表2.

由表2可知, 使用本文提出的EEPS故障诊断算法后, 20种不同故障类型的3个指标均有不同程度的提升, 改进效果较好. 以故障2为例, 进一步比较改进前后对于正常状态和故障状态的诊断准确率.

表 2 改进前后TE过程20种不同故障类型的诊断性能比较

Table 2 TE 20 fault modes diagnostic performance comparisons with PS and EEPS

故障类型	故障诊断算法	ACC	TFM	NFM	故障类型	故障诊断算法	ACC	TFM	NFM
故障 1	PS	0.7057	0.7578	0.6245	故障 11	PS	0.8134	0.8589	0.7246
	EEPS	0.7503	0.8003	0.6667		EEPS	0.8387	0.8794	0.7567
故障 2	PS	0.8133	0.8560	0.7313	故障 12	PS	0.8215	0.8664	0.7311
	EEPS	0.8772	0.9109	0.8027		EEPS	0.8314	0.8738	0.7461
故障 3	PS	0.8230	0.8672	0.7346	故障 13	PS	0.7179	0.7682	0.6393
	EEPS	0.8491	0.8882	0.7680		EEPS	0.7513	0.8010	0.6685
故障 4	PS	0.8143	0.8598	0.7248	故障 14	PS	0.8143	0.8598	0.7249
	EEPS	0.8398	0.8804	0.7571		EEPS	0.8387	0.8798	0.7551
故障 5	PS	0.8097	0.8560	0.7195	故障 15	PS	0.8032	0.8486	0.7160
	EEPS	0.8293	0.8719	0.7445		EEPS	0.8304	0.8730	0.7449
故障 6	PS	0.8480	0.9027	0.6518	故障 16	PS	0.8249	0.8688	0.7365
	EEPS	0.8819	0.9264	0.7018		EEPS	0.8439	0.8841	0.7611
故障 7	PS	0.8057	0.8527	0.7145	故障 17	PS	0.7761	0.8247	0.6871
	EEPS	0.8231	0.8664	0.7385		EEPS	0.8158	0.8605	0.7289
故障 8	PS	0.7638	0.8130	0.6761	故障 18	PS	0.8036	0.8510	0.7115
	EEPS	0.8137	0.8585	0.7275		EEPS	0.8293	0.8719	0.7445
故障 9	PS	0.8189	0.8638	0.7298	故障 19	PS	0.6908	0.7422	0.6114
	EEPS	0.8429	0.8832	0.7599		EEPS	0.7305	0.7807	0.6505
故障 10	PS	0.7642	0.8125	0.6776	故障 20	PS	0.6930	0.7435	0.6176
	EEPS	0.8231	0.8666	0.7377		EEPS	0.7263	0.7769	0.6460

由表3可知, 相较于PS故障诊断算法, 对正例样本集进行7次欠采样, 建立bagging之后的SVM故障诊断分类器之后, 7项指标体现的诊断效果整体较好. 其中对于正常状态的诊断准确率总体提升10.49%, 对于故障状态的查准率总体提升14.03%, 正类F–measure (TFM) 提升6.41%, 负类F–measure (NFM)提升9.76%, 表明改进之后诊断能力提升显著. 此外, 故障状态的诊断准确率和正常状态的查准率分别高达99.59% 和99.83%, 说明此方法能有效的应用于不均衡数据下的故障诊断.

4 结论(Conclusions)

本文针对化工过程故障诊断中普遍存在的正常状态监控数据多、故障数据获取少的样本数据不均衡现象, 提出了一种基于EasyEnsemble思想的PCA–SVM故障诊断算法, 同时使用ACC和F–measure等7个指标作为不均衡数据故障诊断的性能评价标准, 在TE过程的20种不同类型故障数据集上进行实验. 实验结果表明, 本文提出的EEPS故障诊断算法能够显著提高不均衡数据集的诊断精度, 是一种解决不均衡数据集诊断问题的有效方法.

表 3 故障类型2下PS与EEPS方法各评价指标比较

Table 3 Evaluation criteria comparison with PS and EEPS algorithm for fault mode 2

评价指标	PS 1	PS 2	PS 3	PS 4	PS 5	PS 6	PS 7	平均值	EEPS	改进率/%
T–recall	0.8431	0.8486	0.6361	0.8722	0.7264	0.7139	0.6667	0.7581	0.8375	10.47
N–recall	0.9751	0.9876	0.9751	0.9627	0.9793	0.9876	0.9793	0.9781	0.9959	1.82
T–precision	0.9902	0.9951	0.9871	0.9859	0.9905	0.9942	0.9897	0.9904	0.9983	0.80
N–precision	0.6753	0.6859	0.4728	0.7160	0.5450	0.5360	0.4958	0.5896	0.6723	14.03
ACC	0.8762	0.8835	0.7211	0.8949	0.7898	0.7825	0.7451	0.8133	0.8772	7.86
TFM	0.9107	0.9160	0.7736	0.9256	0.8381	0.8310	0.7967	0.8560	0.9109	6.41
NFM	0.7980	0.8095	0.6369	0.8212	0.7003	0.6949	0.6583	0.7313	0.8027	9.76

参考文献(References):

[1] ZHOU Zhijie, HU Changhua, ZHOU Donghua. Fault prediction techniques for dynamic systems based on non-analytical model [J]. *Information and Control*, 2006, 35(3): 608 – 613.
(周志杰, 胡昌华, 周东华. 基于非解析模型的动态系统故障预报技术 [J]. 信息与控制, 2006, 35(3): 608 – 613.)

[2] WANG Guozhu, LIU Jianchang, LI Yuan, et al. Fault diagnosis of industrial processes based on weighted k-nearest neighbor reconstruction analysis [J]. *Control Theory & Applications*, 2015, 32(7): 873 – 880.
(王国柱, 刘建昌, 李元, 等. 加权k最近邻重构分析的工业过程故障诊断 [J]. 控制理论与应用, 2015, 32(7): 873 – 880.)

[3] GAO X, HOU J. An improved SVM integrated GS–PCA fault diagnosis approach of Tennessee Eastman process [J]. *Neurocomputing, Part B*, 2016, 174(22): 906 – 911.

[4] LI Han, XIAO Deyun. Survey on data driven fault diagnosis methods [J]. *Control and Decision*, 2014, 2(1): 1 – 10.
(李晗, 萧德云. 基于数据驱动的故障诊断方法综述 [J]. 控制与决策, 2014, 2(1): 1 – 10.)

[5] MICHELA A, PIETRO D, FRANCESCO M. An experimental study on evolutionary fuzzy classifiers designed for managing imbalanced datasets [J]. *Neurocomputing*, 2016, 146(25): 125 – 136.

[6] MIKEL G, ALBERTO F, EDURNE B, et al. Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets [J]. *Information Sciences*, 2016, 354(1): 178 – 196.

[7] JOSE A S, BARTOSZ K, MICHAL W. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets [J]. *Pattern Recognition*, 2016, 57: 164 – 178.

[8] LIU Tianyu, LI Guozheng. The imbalanced data problem in the fault diagnosis of rolling bearing [J]. *Computer Engineering and Science*, 2010, 32(5): 150 – 153.
(刘天羽, 李国正. 滚动轴承故障诊断中数据不均衡问题的研究 [J]. 计算机工程与科学, 2010, 32(5): 150 – 153.)

[9] ANDREAS B, LAWRENCE R, MOHIEDDINE J. Revision of the Tennessee Eastman process model [C] //*Proceedings of the 9th IFAC Symposium on Advanced Control of Chemical Processes*. Whistler, Canada: IFAC, 2015, 48(8): 309 – 314.

作者简介:

夏丽莎　(1987–), 女, 讲师, 博士, 目前研究方向为数据驱动的故障诊断、故障预报, E-mail: lisaxss@163.com;

杨玉英　(1987–), 女, 讲师, 博士, 目前研究方向为能源经济与安全管理, E-mail: yyy876@126.com;

方华京　(1955–), 男, 教授, 博士生导师, 目前研究方向为网络化控制、故障诊断与预报, E-mail: hjfang@mail.hust.edu.cn.