第 15 卷第 4 期
2009 年 4 月

计算机集成制造系统
Computer Integrated Manufacturing Systems

Vol. 15 No. 4
Apr. 2009

文章编号:1006 - 5911(2009)04 - 0777 - 09

基于改进随机森林的故障诊断方法研究

庄进发,罗　键,彭彦卿,黄春庆,吴长庆
(厦门大学 自动化系,福建　厦门　361005)

摘　要:为解决不可识别故障诊断中无法有效定位的问题,提出一种基于改进随机森林的故障诊断方法。该方法通过改进决策树的 bagging 方式,采用条件概率指数进行决策树的无偏节点分裂,并以权重投票法综合决策树的分类结果。在此基础上,利用变量重要性测量来获取辅助故障定位的故障原型指数,从而较好地弥补了随机森林和传统机器学习在故障诊断中的不足和局限性。最后在一个标准数据集和田纳西 - 伊斯曼故障诊断的问题上进行验证,结果证明了该方法的有效性与可行性。

关键词:故障诊断;随机森林;故障原型;田纳西 - 伊斯曼
中图分类号:TP277　　**文献标识码**:A

Fault diagnosis method based on modified random forests

ZHUANG Jin-fa, LUO Jian, PENG Yan-qing, HUANG Chun-qing, WU Chang-qing
(Department of Automation, Xiamen University, Xiamen 361005, China)

Abstract: To solve the problem of inefficient determining fault location in unidentified fault diagnosis of traditional machine-learning technologies, a fault diagnosis method based on modified random forests was proposed. Firstly, random decision trees were created via modified algorithm of bagging and unbiased split selection based on conditional probability index so as to construct random forests. Secondly, weighted voting was applied to combine the prediction of the decision trees. Then, fault prototypes were computed through the measurement of variable-importance in random forests, which assisted in determining the fault location. Finally, the proposed method was illustrated and documented thoroughly in an application of standard dataset and Tennessee Eastman Process (TEP) fault diagnosis. The results verified the presented approach's feasibility and effectiveness.

Key words: fault diagnosis; random forests; fault prototypes; Tennessee Eastman

0　引言

故障是系统的非正常状态,即在正常条件下,系统实际功能输出或附加输出超越规定界限的现象,故障造成系统或某些部件在一定程度上的破坏[1-2]。故障从确定性角度可以分为可识别故障与不可识别故障两类。可识别故障指故障发生的所在位置或引起故障发生的原因,能够从故障类别标签号信息中直接确定,例如先前已发生过并得到确诊的故障,其发生位置或原因与故障类别标签号信息之间存在着一一对应关系,即根据故障类别标签号信息可找出故障发生的原因;不可识别故障指故障发生所在位置或引起故障发生的原因,暂且无法从故障类别标签号信息中直接确定,例如当有新的故障被检测出来并且存在一定的样本 n 时,通过专家知识或模式聚类算法,可以把样本 n 聚为 c 类,此时新故障 i 发

收稿日期:2008-04-29;**修订日期**:2008-07-30。Received 29 Apr. 2008; accepted 30 July 2008.
基金项目:国家自然科学基金资助项目(60704043)。**Foundation item**: Project supported by the National Natural Science Foundation, China (No. 60704043).
作者简介:庄进发(1980 -),男,福建厦门人,厦门大学自动化系博士研究生,主要从事智能控制、机器学习等的研究。
E-mail: jinfazhuang@xmu.edu.cn。

生的所在位置或原因无法从故障的类别标签号信息 $m_i(i=1,2,\cdots,c)$ 中得到确定,只有在工程人员通过专家知识或者其他算法找到故障发生的所在位置或者原因,并建立故障类别与故障发生原因之间的对应关系之后,故障发生原因才可通过故障类别来确定,由此不可识别故障也就转化为可识别故障,这对于在线故障诊断研究是非常重要的。

故障诊断指在一定的检测策略的指导下,对被诊断系统实施自动检测,它主要包括三个方面:①故障检测,即确定系统是否有故障发生[3];②故障识别,若故障发生则显示与故障最相关的监控变量;③故障位置的确定,即确定故障发生的所在位置[4-5]。故障诊断在确保工业系统正常运行和安全生产等方面起着重要的作用,其理论和方法也经历了不同的发展阶段。迄今为止,多个领域的研究者已提出了多种故障诊断方法,如基于数学模型故障诊断(参数估计方法、观测器方法、对偶关系方法等)[4]、基于人工智能的方法(包括基于机器学习、基于进化计算、基于模糊数学等)[6],这些故障诊断方法在流程工业和电力系统安全检测等多个领域都得到了广泛的应用。其中基于模型的故障诊断是在已知系统数学模型的情况下,通过比较实际系统输出和数学模型的输出来产生残余误差,当系统没有故障发生时,残余误差为零,否则残余误差不为零。与其他的故障诊断方法相比,基于数学模型的故障诊断效果较好,但是随着现代工业系统复杂程度的增加和系统规模的扩大,其数学模型已经很难获取,使得基于模型故障诊断的应用受到一定限制。基于机器学习方法是智能诊断中一类有效的方法,与基于模型方法相比,该方法不需要系统准确的模型,而是通过历史的监控输入输出数据来建立系统的黑箱故障诊断模型,这类方法同时兼有故障检查、识别与定位功能,成为该领域的一个研究热点。

在现有基于机器学习的故障诊断研究中,如基于多层前馈网络(Back Propagation, BP)、支持向量机(Support Vector Machine, SVM)等故障诊断,很多学者把重点集中在对可识别故障诊断的研究上,并取得了一定的成果[7-9]。随机森林(random forests)是 Leo Breiman 于 2001 年提出的一种组合分离器算法,具有较好的泛化性和准确性[10];文献[11]将随机森林应用于航空发动机的可识别故障诊断中,获得了很好的性能,但是当有不可识别故障产生时,这类基于机器学习方法的故障诊断仍然无法很好地辅助工程人员确定故障发生的位置,降低了故障诊断的效率。为解决上述问题,本文提出了一种基于改进随机森林故障诊断方法。该方法通过改进决策树的 bagging 方式,采用条件概率指数进行决策树的无偏节点分裂,并以权重投票法则综合决策树的分类结果。在此基础上,利用变量重要性测量来获取辅助故障定位的故障原型指数,最后在一个标准数据集和田纳西 - 伊斯曼(Tennessee Eastman Problem, TEP)故障诊断的问题上进行了验证。

1 随机森林

定义 1 随机森林。随机森林是由多个决策树 $\{h(x,\Theta_k), k=1,2,\cdots,n\}$ 组成的分类器,其中 Θ_k 是相互独立且同分布的随机向量,最终由所有决策树综合投票决定输出结果。

1.1 随机决策树的生成与投票法则

随机决策树是组成随机森林的最小决策单元,它的生成可以概括为两个"随机"特征:①通过 Bagging[12]方法生成每棵决策树的随机训练样本;②随机选择训练样本中的特征来进行随机决策树的节点分裂。

在组合分类器算法中,综合各个单分类器的决策结果得出组合分类器结果的方法有很多,对于随机森林,大多数投票法则已经证明其有效性[13]。

1.2 变量重要性计算

随机森林的一个重要特性是给出变量重要性测量,其计算过程如下:①对已生成的随机森林,用 oob[10]数据测试其性能,得到一个 oob 准确率 e;②随机地改变 oob 数据中某个特征 v 的值(即给特征 v 人为地加入噪声干扰)得 oob_d,再用 oob_d 数据测试随机森林的性能,得到一个新的 oob 准确率 $\hat{e}$;③$\hat{e}-e$ 作为相应特征 v 的重要性度量值。由上述可知,随机森林变量重要性计算是描述所有变量对所有类别的一个整体计算。

1.3 随机森林的收敛性

给定 k 个分类器集合 $\{h_1(x), h_2(x), \cdots, h_k(x)\}$、输入向量 x 和输出向量 y,定义间隔函数

$$mg(x,y) = a_{vk}I(h_k(x)=y) - \max_{j\neq y} avkI(hk(x)=j)。 \quad (1)$$

式中:$I(\cdot)$ 为指示函数,$avk(\cdot)$ 为取平均值。分类器的泛化误差

$$PE^{*} = P_{x,y}(mg(x,y) < 0)。 \quad (2)$$

将上面的结论推广到随机森林 $h_k(X) = H(X, \theta_k)$。如果森林中树的数目较大,可以用大数定律和树的结构得到如下定理。

定理 1　随着树的数目的增加,对于所有随机向量,$\theta, \cdots, PE^{*}$趋向于

$$P_{x,y}(p\theta(h(x,\theta) = y) - \max_{j \neq y} p\theta(h(x,\theta) = j) < 0)。 \quad (3)$$

定理 1 的证明在文献[10]中已经给出,并且表明随机森林不会出现过拟合。这是随机森林的一个重要特点,随着树的增加,泛化误差 PE^{*} 将趋向某一上界。

定义 2　随机森林的边缘函数

$$mr(x,y) = p\theta(h(x,\theta) = y) - \max_{j \neq y} p\theta(h(x,\theta) = j)。 \quad (4)$$

S 为分类器$\{h(X,\theta)\}$的强度

$$s = E_{x,y}mr(x,y)。 \quad (5)$$

假设 $s \geq 0$,根据切比雪夫不等式,由式(4)和式(5)可以得到

$$PE^{*} \leq \mathrm{var}(mr)/s^{2}。 \quad (6)$$

不等式(7)要求的 $var(mr)$ 具有以下形式:

$$\begin{cases} \mathrm{var}(mr) = \bar{\rho}(E\theta sd(\theta))^{2} \\ \mathrm{var}(mr) \leq \bar{\rho}E\theta\mathrm{var}(\theta) \end{cases}。 \quad (7)$$

而

$$\begin{cases} E\theta\mathrm{var}(\theta) \leq E\theta(E_{x,y}mg(\theta,x,y))^{2} - s^{2} \\ E\theta Var(\theta) \leq 1 - s^{2} \end{cases}。 \quad (8)$$

由式(6)～式(8)得到以下结论:

定理 2　随机森林的泛化误差上界的定义为

$$PE^{*} \leq \bar{\rho}(1 - s^{2})/s^{2}, \quad (9)$$

其中 $\bar{\rho}$ 是相关系数的均值,s 是树的分类强度。

2　基于改进无偏随机森林的故障诊断

定义 3　故障原型。设故障样本包含有 n 类,m 个特征,其中特征 $v_i\ (i = 1,2,\cdots,m)$ 与类别 $C_j\ (j = 1,2,\cdots,n)$ 之间的相关性值为 r_{ij},则定义故障原型为

$$P = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \cdots & & \cdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix}。 \quad (10)$$

由定义 3 可得,故障原型是衡量特征 v_i 与故障 C_j 的相关性程度,它可以辅助工程人员定位故障发生的位置。

本文首先采用 Leo Breiman 的随机森林算法对 TEP 故障诊断问题进行研究和仿真,结果表明随机森林能够较好地识别故障类别,但是应用其变量重要性测量计算故障原型时,无法得出所需结果。因此,本文对随机森林改进算法的可行性和有效性开展了研究,使其更好地应用于故障诊断。

文献[14]从试验的角度指出,用于构建随机决策树的随机样本(约有 63.2%的样本没有重复的[10])采用放回抽样的 bagging 方式,这种产生方式从统计推理的角度上可能会影响决策树的有效性,甚至失效;文献[15]指出,随机决策树在运用 Gini 指数对决策树的节点进行分裂时是有偏的,即偏向那些具有离散性质的属性特征、不平衡类数据样本中样本较多的属性特征和具有缺失值的属性特征。

另外在变量重要性测量计算过程中,采用 *oob* 的准确率计算其测量值。文献[16]指出,这种计算方法存在一定不足,即无法充分衡量变量的重要性。同时,变量重要性测量值只给出整体变量对整体类别的重要性测量,没有详细地给出单个变量对单个类别的重要性测量,在一定程度上无法有效辅助工程人员定位故障位置。

2.1　随机决策树的改进算法

针对上述不足,本文分析了相关的改进策略。由式(9)可知,随机森林泛化误差上界与随机决策树之间的相关度 $\bar{\rho}$ 和随机决策树分类强度 s 有关,即

$$\bar{\rho} = E(h(x,\Theta_k),s,v)。 \quad (11)$$

其中:Θ_k, Θ'_k 为建树的随机样本;s, s 为决策方式(节点分裂方式,节点分裂变量选取);v, v 为投票方式。

由式(9)和式(11)可知,在分类强度 s 不变单的情况下,降低相关度 $\bar{\rho}$ 可以降低泛化误差 PE^{*}。由于随机森林的决策树采用的是不修剪的增长方式,通过式(11)的三个指数(Θ_k, s, v)降低相关度的同时,决策树的分类强度不受影响[10],甚至可以忽略。基于上述分析,本文提出以下改进策略:

(1)用于构建随机决策树的随机样本采取动态 bagging 方式,即每棵决策树在构建时随机选取训练样本的比例为 50%～63.2%,并且采取的是不放回抽样。文献[17]指出,bagging 样本的比例选择在 50%～63.2%之间是合理的。经过这样的动态 bagging 选择,每棵树的样本差异度比例明显增加,

即将原来的 0 %～26. 4 %范围扩展到 0 %～50 %。由式(11)可知,决策树之间的相关度 $\bar{\rho}$ 将降低。

(2)随机决策树节点的分裂上,本文采用区别于 Gini 值的无偏分裂方式,它是基于条件概率的一种无偏分裂[18],

$$Mr(A_i) = \frac{(\sum_{j=1}^{m_i} p(v_{i,j})^2 \sum_{i=1}^{c} p(La_i \mid v_{i,j})^2)}{(\sum_{i=1}^{c} p(v_{i,j})^2 (1 - \sum_{i=1}^{c} p(La_i)^2))} - \frac{(\sum_{j=1}^{m_i} p(v_{i,j})^2 \sum_{i=1}^{c} p(La_i)^2)}{(\sum_{i=1}^{c} p(v_{i,j})^2 (1 - \sum_{i=1}^{c} p(La_i)^2))}。 \quad (12)$$

式中:m_i 为属性 A_i 的值;$p(v_{i,j})$ 为属性 A_i 是 v_j 值的概率;c 为训练样本中类别的个数;$p(La_j \mid v_{i,j})$ 为类标 La_j 在属性 A_i 条件下值为 v_j 的概率。

(3)随机决策树对于每个样本的分类性能是不一样的,如何更合理地综合各个决策树的分类结果,是提高随机森林性能的一个策略。本文提出了基于样本相似度的权投票法则。相似度[10]指对于一棵决策树,一个样本经其分类,最后肯定落在决策树的叶子点上,若两个样本落在相同的叶子节点上,则两样本的相似度加 1,对于所有的样本和所有的决策树,样本相似度为

$$\mathrm{pro} = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \cdots & & \cdots \\ p_{1n} & \cdots & p_{mn} \end{bmatrix}。 \quad (13)$$

式中,p_{ij} $(i = 1,2,\cdots,n, j = 1,2,\cdots,n)$ 为第 i 个样本与第 j 个样本对于所有决策树落在同一个节点的次数。对于 oob 样本同样也存在一个相似度矩阵 pro_{oob},选取 $p_{ij} \geq t$ 的一类样本,测试所有决策树 K 的间隔函数 $mg(x, y)$。若有决策树 K_z 的 $mg_z(x, y) \leq 0$,则 K_z 的权重值为一个较小的值 W_o;若 $mg_c(x, y) \geq 0$,则 K_c 的权重值为 $W_c = avg(\sum_{i=1}^{k} mg_c(x, y))$。

2.2 变量重要性测量改进算法

文献[18]指出,随机森林在变量重要性测量过程中存在有偏性,同时无法给出故障原型。上一节提出无偏性及性能的改进策略,以下研究主要针对随机森林输出故障原型的改进策略。随机森林在输出故障原型的不足表现如下:

$$M1_h = \max\{0; \hat{e}_h - \hat{e}\}, h = 1,2,\cdots,m。 \quad (14)$$

式中:$M1_h$ 为第 h 个特征变量对于所有故障类别的重要性测量值;$\hat{e}$ 为利用 oob 样本对随机森林测试后获得的一个准确率;$\hat{e}_h$ 为在第 h 个变量加入噪声后,oob 样本对随机森林测试后获得的一个准确率;m 为特征个数。最后随机森林的变量重要性值输出为

$$\mathrm{M1} = [M1_1 \quad \cdots \quad M1_m]^{\mathrm{T}}。 \quad (15)$$

由式(15)可知,随机森林的变量性测量是一个所有特征对所有类别的测量,它无法满足故障定位需求,同时其测量变量重要性只是应用了准确率一个方面,存在一定的不足,因此本文做以下改进:

(1)将 oob 样本根据故障样本的类别 k $(k = 1, \cdots, c)$ 分为 c 个子样本 $oob_1, \cdots, oob_k, \cdots, oob_c$。

(2)将噪声 d 加入到 oob_c 第 h 变量中,获得噪声样本 oob_k^h $(h = 1,2, \cdots, m, k = 1,2, \cdots, c)$,分别将 oob_c 与 oob_k^h 对随机森林做测试,得到第 h 个变量对于第 c 类故障类别的准确率 $\hat{e}_c^h$ 和 $\hat{e}_c^{h\prime}$,将 $\hat{e}_c^h$ 与 $\hat{e}_c^{h\prime}$ 代入式(14)得

$$M1_c^h = \max\{0; \hat{e}_c^{h\prime} - \hat{e}_c^h\}, h = 1,2, \cdots, m。 \quad (16)$$

其中 $M1_c^h$ 为第 h 个特征对第 c 类故障类别的一种基于准确率 e 的重要性测度值。

(3)将 oob_c 与 oob_k^h 代入式(17)得

$$M2_k^h = \max\{0; avs[mg(y, x) - mg_h(y, x)]\}。 \quad (17)$$

其中:$mg(y, x)$ 为间隔函数,$M2_k^h$ 为第 h 个特征对第 c 类故障类别的一种基于间隔函数的重要性测度值。

(4)将 $M1^{*h}_k = M1_k^h + M2_k^h$ 作为最后变量重要性的输出,即特征 v_i $(i = 1,2, \cdots, m)$ 与类别 C_j $(j = 1,2, \cdots, n)$ 之间的相关性测量值为 $r_{ij} = M1^{*h}_k$。所以故障原型为

$$\mathrm{P} = \mathrm{M1}^* = \begin{bmatrix} M1^{*1}_1 & \cdots & M1^{*1}_k \\ \cdots & & \cdots \\ M1^{*h}_1 & \cdots & M1^{*h}_k \end{bmatrix}。 \quad (18)$$

式中:k 为样本的类别数,$k = 1, \cdots, c$;h 为特征数,$h = 1,2, \cdots, m$;$\mathrm{M1}^* = \mathrm{P}$,P 为故障原型。

2.3 算法伪码

本文改进随机森林算法的伪码如下:

```
Algorithm.
Begin
    ntree = m, nselvar = k, data = [v1, v2, ···, vp], [d1, d2] = sizeof(data)
    For each tree i ∈[1, m] do
```

take a bootstrap sample $data_i$ without replacement for construct random decision tree

build a random decision tree t_i

For each split point $j \in [1, 2d1+1]$ of tree t_i do

Select k variables as splitting variables randomly

Compute $Mr(A_i)$ to decide which variables will be selected to split

Split the point

End

Get out-out-bag sample oob_i

Get out-of-bag estimates oob_test_i

Class oob_i into $oob_1, \cdots, oob_k, \cdots, oob_c$

Compute $M1_c^h, M2_k^h$

End;

$oob\_test = [oob\_test_1, oob\_test_2, \cdots, oob\_test_m]$

Weight voting oob_test for getting misclassification rate

Weigh voting $M1^{*h}_k$ for getting fault prototypes $M1^*$

End

其中 Ntree 为森林中的决策树个数,nselvar 为变量分裂选择个数,$data_i$ 为 bagging 的随机样本,$Mr(A_i)$ (见式(12))为决策树节点的分裂值,$M1^{*h}_k$ (见式(18))为故障原型。

3　实验与分析

3.1　实验数据

为了验证本文改进算法的可行性,采用两个标准数据集,即能量测试集(http://www.stat.wisc.edu/～loh/)和 TEP(http://brahms.scs.uiuc.edu)故障诊断问题。

数据集 1 包含 Null Case(NC)和 Power Case(PC)两种类型的数据(如表 1)。

表 1　测试集 1 描述

数据类型	含特征变量个数	分类数	与类别相关的变量
Null Case	5	1	无
Power Case	5	1	X_2

TEP 故障诊断问题(数据集 2)由美国伊斯曼化学公司创建,其目的是为评价过程控制和监控方法提供一个现实的工业过程。TEP 作为比较各种方法的数据源,已在过程监控领域得到了广泛的应用[19]。TEP 包含有 22 类故障(17 类为可识别故障,5 类为不可识别故障)、55 个监控变量、960 个样本数(每 3 min 采集一次,总共运行 48 h)。根据文献[20]的研究建议,本文只选取 24 h～48 h 时间段的 480 个数据(包含可识别故障 1,4,5;不可识别故障 17,18,20),如表 2 所示。

表 2　过程故障

故障类别	故障位置	故障原因	故障类型
IDV(1)	A/C 进料比率,B 成分不变	阶跃	可识别故障
IDV(4)	反应器冷却水的入口温度	阶跃	可识别故障
IDV(5)	冷凝器冷却水的入口温度	阶跃	可识别故障
IDV(17)	未知	未知	不可识别故障
IDV(18)	未知	未知	不可识别故障
IDV(20)	未知	未知	不可识别故障

3.2　验证变量重要性计算的无偏性

为了验证改进随机森林在变量重要性测量的无偏性,本文在测试集中应用了 Leo Breiman 的随机森林算法与基于改进思想的随机森林,实验结果如图 1 和图 2 所示。

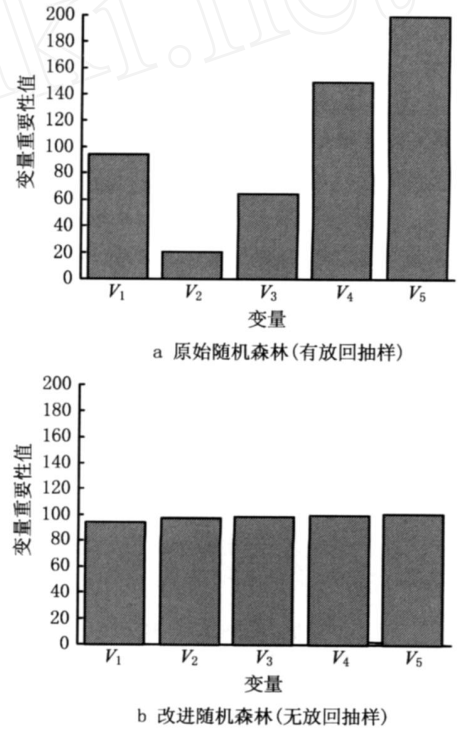
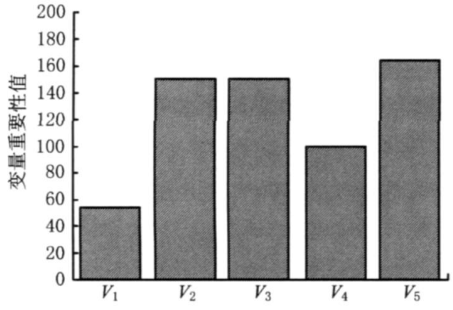
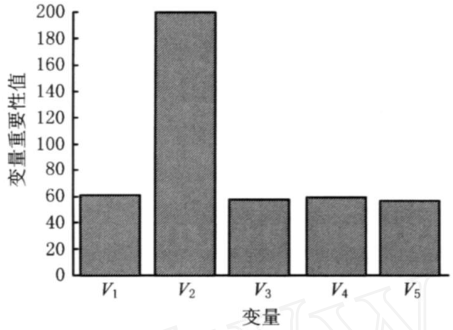


图1　基于原始随机森林与改进随机森林Null Case的应用

从图 1 可以看出,在 Null Case 中,$V_1, \cdots, V_5$ 这五个变量与类别是没有相关的,即这五个变量的重要性值应该均等,但是图 1a 显示变量 V_5 与类别的相关性最大,与实际不符;图 1b 显示五个变量的重要性值基本一致,与实际符合。为了再次证明这一结论,本文同样把原始随机森林和改进后的随机森林应用于 Power Case,其结果如图 2 所示。在 Power Case 数据集中,只有 V_2 与类别相关,图 2a 显示,变量 V_2,V_5 与类别比较相关,不完全符合实



a 原始随机森林有放回抽样



b 改变随机森林无放回抽样

图2 基于原始随机森林与改进随机森林Power Case的应用

际情况;图 2b 显示,变量 V_2 与类别最相关,其他四个变量的相关程度基本一致,符合实际情况。

3.3 改进随机森林的准确率分析

本节应用改进随机森林于 TEP 的故障诊断中,并与其他机器学习方法的准确率进行比较。实验结果如表 3 所示,其中 BP 采用三层网络结构,输入层 55 节点,隐层激活函数为 *tansig*,节点为 23,输出激活函数 *purelin*,节点为 6;SVM1 采用高斯核(RBF),σ= 1;SVM2 采用多项式核(polynomial),*gamma* = 2.3。

表 3 机器学习准确率比较

机器学习方法	训练数据准确/ %	测试数据准确率/ %
改进随机森林	98.74	97.81
BP-net	93.35	91.78
原始随机森林	96.44	95.21
SVM1(RBF)	95.17	94.56
SVM2(polynomial)	85.32	80.63

表 3 显示,原始的随机森林相比 BP-net,SVM 准确率都有一点提高;同时显示,组合学习分类器算法比单一的分类器有所提高,改进随机森林的准确率比原始随机森林又有一定的提高,从而说明了本文算法的有效性和可行性。

3.4 原随机森林诊断故障发生位置的不可行性

本节将原随机森林算法应用于 TEP 故障位置诊断中,其结果如表 4 所示(只取变量重要性值最大的前 10 条)。

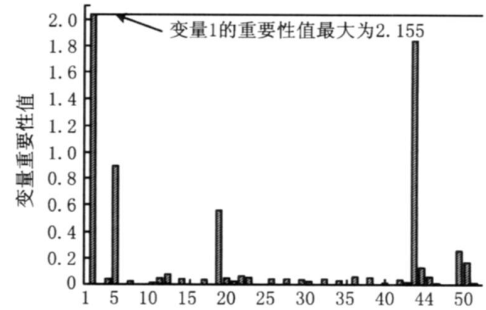
表 4 变量重要性

特征变量序号	原始变量重要性值	Z 转换重要性值	显著性
51	15.407	45.683	0.003
52	7.420	29.450	0.003
9	2.431	28.831	0.005
46	6.552	24.177	0.008
50	3.887	17.939	0.011
19	3.980	16.927	0.012
20	3.294	16.569	0.013
44	6.656	16.179	0.017
18	5.394	15.296	0.019
4	5.981	14.817	0.020

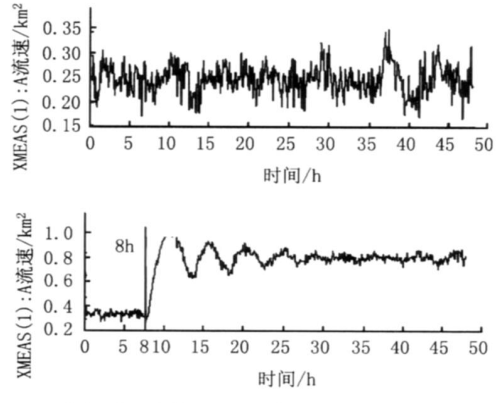
表 4 显示,随机森林只给出整体变量对整体故障类别的一个整体性排列,无法给出故障原型。

3.5 改进随机森林诊断故障发生位置的可行性

本节将改进随机森林应用于可识别故障的诊断中。在该实验中,首先假定故障的发生位置未知。其实验结果如图 3 所示。



a 故障IDV(1)的故障原型



b 变量1在正常情况下和故障IDV(1)发生下的状态

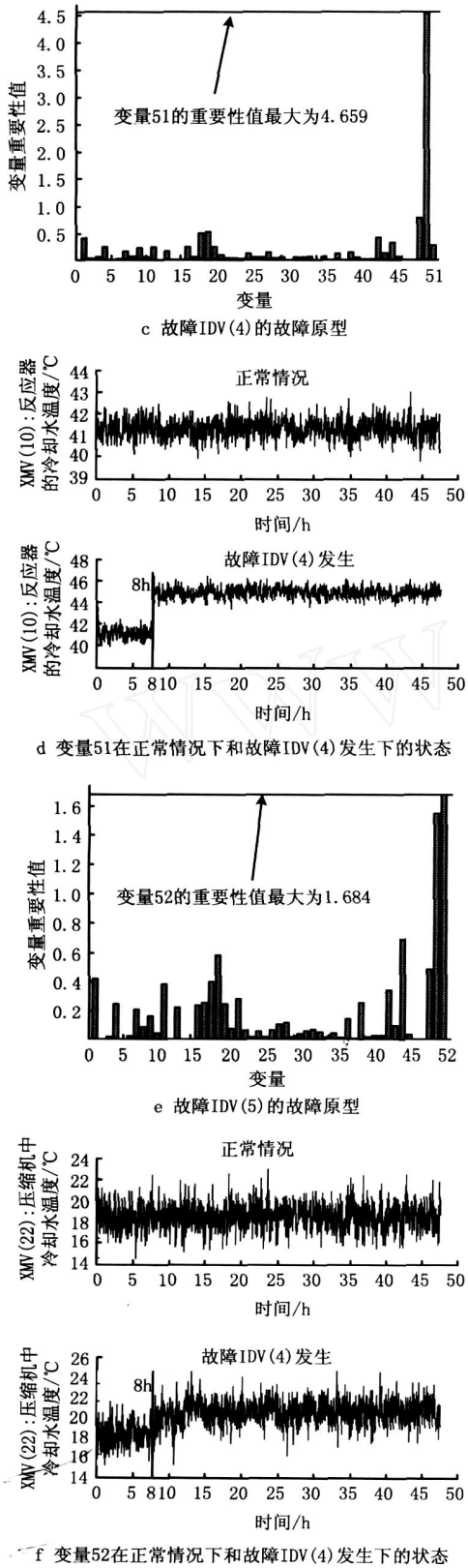


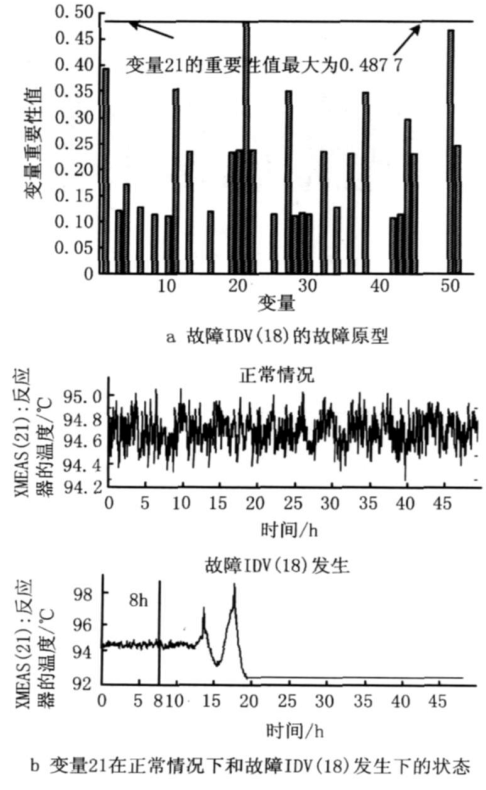
图3　可识别故障(1,4,5)的故障原型及其在不同条件下48 h运行状态

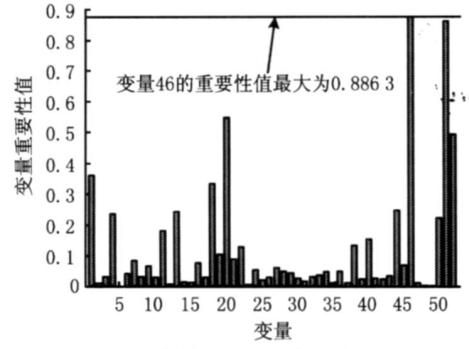
图 3a 显示,在故障 IDV(1)发生的情况下,变量 1,44 的重要值最突出,即故障 IDV(1)发生位置或者原因与变量 1,44 最相关。为进一步验证,本文画出了变量 1 在正常情况下和在故障 IDV(1)发生情况下的运行状态(如图 3b,其中故障 IDV(1)是在系统运行 8 h 后加入的),从图上可以看出,变量 1 在故障 IDV(1)被加入后相对于正常状态下发生了明显变化。同样分析,故障 IDV(4)的故障发生位置与监控变量 51 位置最相关,故障 IDV(5)的故障发生位置与监控变量 52 位置最相关。为了验证本仿真实验与实际的相符程度,本文将结果与文献[21](如表 5)做了比较。从表 5 可以看出,图 3 的结果与实际的情况较为符合。

表 5　故障与监控变量的相关性

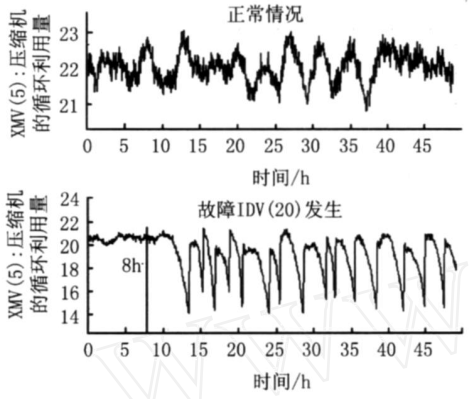
故障类别	故障位置	变量描述
IDV(1)	X_1, X_{44}	A/C 进料比率,B 成分不变
IDV(4)	X_{51}	反应器冷却水流
IDV(5)	X_{52}	成分 H

从以上实验与分析中可以看出,改进随机森林在可识别故障的诊断中能够较好地与实际情况相符。依据同样的步骤,本文将改进随机森林应用于不可识别故障的诊断仿真实验中,其结果如图 4 所示。

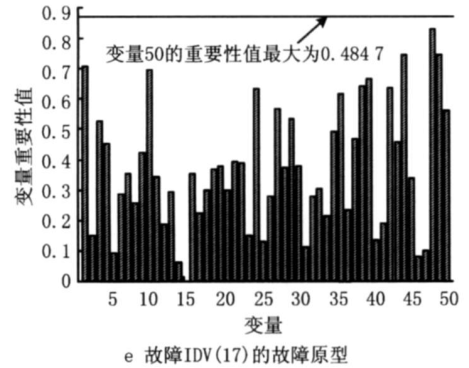




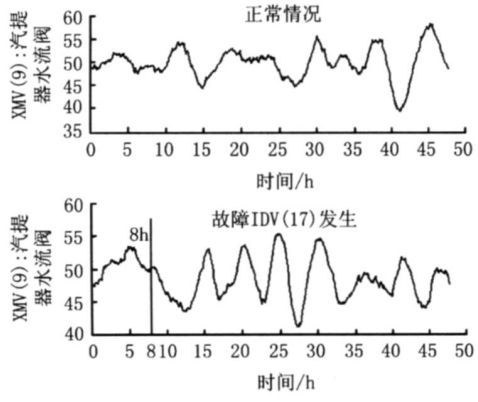
c 故障IDV(20)的故障原型



d 变量46在正常情况下和故障IDV(20)发生下的状态



e 故障IDV(17)的故障原型



f 变量50在正常情况下和故障IDV(17)发生下的状态

图4 不可识别故障(18, 20)的故障原型及其在不同条件下48 h运行状态

同分析图 3 的方法一样,可以看出与故障 IDV(17), IDV(18)和 IDV(20)最相关的变量为监控变量 50, 21 和 46。根据文献[20]的研究成果和 TEP 仿真器构建原理[20],得出本文的改进随机森林在不可识别故障的诊断可行且有效。

4 结束语

本文提出一种基于改进随机森林的故障诊断方法。该随机森林通过动态 bagging 方式形成构建决策树的随机样本,采用条件概率指数代替 Gini 指数,进行决策树的无偏节点分裂。以基于样本的相似度的权重投票法则综合决策树的分类结果,并细化随机森林变量重要测量。在此基础上得到辅助定位故障发生位置的故障原型指数。最后,以一个标准数据集和 TEP 故障诊断问题为对象,将本文的改进算法应用于可识别故障与不可识别故障的诊断中,实验结果验证了本文算法的可行性和有效性。

参考文献:

[1] WANG Daoping, ZHANG Yizhong. Theory and application of fault diagnosis[M]. Beijing: China Machine Press, 2001: 16-30 (in Chinese). [王道平,张义忠. 故障智能诊断系统的理论与方法[M]. 北京:机械工业出版社, 2001: 16-30.]

[2] SVETNIK A T, LIAW V, CULBERSON C. QSAR modelling using random forest, an ensemble learning tool for regression and classification[J]. Journal of Chemical Information and Computer Sciences, 2003, 43(3): 947-958.

[3] RAICH A, CINAR A. Statistical process monitoring and disturbance diagnosis in multivariable continuous processes[J]. Aiche Journal, 1996, 42(4): 995-1009.

[4] VENKATSUBRAMANIAN V, RENGASWAMY R, YIN K, et al. A review of process fault detection and diagnosis Part I: quantitative model-based methods[J]. Computers and Chemical Engineering, 2003, 27(3): 293-311.

[5] PALADE V, BOCANIALA C D, JAIN L C. Computational intelligence in fault diagnosis[M]. Berlin, Germany: Springer, 2006: 1-10.

[6] VENKATASUBRAMANIAN V, RENGASWAMY R, KAVURI S N. A review of process fault detection and diagnosis Part III: Process history based methods[J]. Computers and Chemical Engineering, 2003, 27(3): 327-346.

[7] TZAFESTAS S G, DALIANIS P J. Fault diagnosis in complex systems using artificial neural networks[C]// Proceedings of the 3rd IEEE Conference on Control Applications. Washington, D.C., USA: IEEE, 1994: 877-882.

[8] GE M, DU R, ZHANG G, et al. Fault diagnosis using sup-

port vector machine with an application in sheet metal stamping operations[J]. Mechanical Systems and Signal Processing, 2004, 18(1):143-159.

[9] HE Yigang, TAN Yanghong, SUN Yichuang. A neural network approach for fault diagnosis of large-scale analogue circuits[J]. IEEE International Symposium on Circuits and Systems, 2002, 1:153-156.

[10] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1):5-32.

[11] YAN Weizhong. Application of random forest to aircraft engine fault diagnosis[C]// Proceedings of IMACS Multiconference on Computational Engineering in Systems Applications. Washington, D.C., USA:IEEE, 2006:468-475.

[12] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140.

[13] WANG Yu, ZHOU Zhihua, ZHOU Aoying. Machine learning and its application[M]. Beijing: Tsinghua University Press, 2006:1-40(in Chinese). [王　钰, 周志华, 周傲英. 机器学习及其应用[M]. 北京:清华大学出版社, 2006:1-40.]

[14] POLITIS D N, ROMAMO J P, WOLF M. Subsampling[M]. New York, N.Y., USA:Springer, 1999.

[15] STROBL C, BOULESTEIX A L, ZEILEIS A. Bias in random forest variable importance measures: illustrations, sources and a solution[EB/OL]. [2008-04-03]. http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper490.pdf.

[16] AUSTIN P C, TU J V. Bootstrap methods for developing predictive models[J]. The American Statistician, 2004, 58(2):131-137.

[17] FRIEDMAN J H, HALL P. On bagging and nonlinear estimation[J]. Journal of Statistical Planning and Inference, 2007, 137(3):669-683.

[18] ROBNIK-ŠIKONJA M. Improving random forests[J]. Lecture Notes in Computer Science, 2004, 3201:359-370.

[19] CHEN G, MCAVOY T J. Predictive on-line monitoring of continuous process[J]. Journal of Process Control, 1999, 8(5):409-420.

[20] DOWNS J J, VOGEL E F. A plant-wide industrial-process control problem[J]. Computers & Chemical Engineering, 1993, 17(3):245-255.

[21] CHIANGL H, RUSSELL E L, BRAATZ R D. Fault detection and diagnosis in industrial systems[M]. Berlin, Germany:Springer, 2001:141-142.

(上接第 771 页)

参考文献:

[1] NI Xiaodan, GONG Jinke, LIU Jinwu, et al. Modeling and analyzing the spring-back of the thin auto-closed ring with cut after grinding process[J]. Transactions of the Chinese Society of Agricultural Machinery, 2007, 38(4):158-162, 176(in Chinese). [倪小丹, 龚金科, 刘金武, 等. 薄壁封闭环磨削回弹的建模与数值研究[J]. 农业机械学报, 2007, 38(4):158-162, 176.]

[2] LIAO Mingfu, LU Yajuan, ZHOU Xiaolong. Diagnosis of the thermal bow of a shaft in a three stage centrifugal compressor [J]. International Journal of Plant Engineering and Management, 2003, 8(1):1-8.

[3] TIAN Shujun, LIU Wanhui, ZHANG Hong. Research of VR-based 3D NC machining environment and key technology[J]. Journal of System Simulation, 2007, 19(16):3727-3730(in Chinese). [田树军, 刘万辉, 张　宏. 基于 VR 的三维数控加工环境及其关键技术研究[J]. 系统仿真学报, 2007, 19(16):3727-3730.]

[4] LIU Jinwu, GONG Jinke, GUO Jianxin, et al. Descriptive geometry methods for determining the track of wire incision to embed cut of ring[J]. Journal of Engineering Graphics, 2007, 28(3):123-127(in Chinese). [刘金武, 龚金科, 郭建新, 等. 确定卡环嵌入式接口线切割轨迹的画法几何方法研究[J]. 工程图学学报, 2007, 28(3):123-127.]

[5] LIU Jinwu, GONG Jinke, GUO Jianxin, et al. Research on mathematical layout technique for track of WEDM in processing the three dimension cut[J]. Electromachining & Mould, 2006(1):19-21, 24(in Chinese). [刘金武, 龚金科, 郭建新, 等. 三维接口电火花线切割轨迹的数学放样技术研究[J]. 电加工与模具, 2006(1):19-21, 24.]

[6] LIU Jinwu, GONG Jinke, GUO Jianxin, et al. Virtual process technology on wire incision to embed type cut of ring[J]. Journal of Hunan Institute of Engineering: Natural Science Edition, 2006, 16(3):21-24(in Chinese). [刘金武, 龚金科, 郭建新, 等. 卡环嵌入式接口虚拟线切割工艺研究[J]. 湖南工程学院学报:自然科学版, 2006, 16(3):21-24.]

[7] LIU Xiaonian. Machanical graphic[M]. Beijing: China Machin Press, 1999:33-380(in Chinese). [刘小年. 机械制图[M]. 北京:机械工业出版社, 1999:33-380.]

[8] LIU Zijian, HUANG Hongwu. ZONG Zian. Principle and application technique[M]. Changsha: Hunan University Press, 2000:78-142(in Chinese). [刘子建, 黄红武, 宗子安. 计算机辅助设计(CAD)原理与应用技术[M]. 长沙:湖南大学出版社, 2000:78-142.]