控制与决策
Control and Decision

基于迹距离划分决策树的高炉故障诊断方法

刘亚雪, 张敬川, 王显鹏

引用本文:
刘亚雪，张敬川，王显鹏. 基于迹距离划分决策树的高炉故障诊断方法[J]. 控制与决策，2025，40(5)：1533-1540.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.0843

基于迹距离划分决策树的高炉故障诊断方法

刘亚雪[1,2]，张敬川[1,2]，王显鹏[2,3†]

(1. 东北大学 工业智能与系统优化国家级前沿科学中心，沈阳 110819；
2. 东北大学 智能工业数据解析与优化教育部重点实验室，沈阳 110819；
3. 辽宁省智能工业数据解析与优化工程实验室，沈阳 110819)

摘　要: 随着工业自动化和智能化的发展, 决策树模型在高炉故障诊断领域得到了广泛应用, 但对于炼铁过程中存在高维度、非线性和强耦合的特点, 传统决策树模型的构建容易陷入局部最优解, 效率较低且复杂度较高. 针对这些问题, 首先引入迹距离函数, 并证明在迹距离函数中任何局部最优解也是全局最优解的性质; 接着针对决策树的节点分裂过程, 提出一种基于迹距离划分的决策树模型, 记作 TraceTree. 此模型一方面可以更快速地评价一个节点的划分效果, 有效降低决策树模型的复杂度; 另一方面能够识别出对故障诊断最有贡献的特征参数并获得更高的诊断精度. 与其他改进模型的对比实验结果表明, 所提出的模型在更短的训练时间内能取得最优的高炉故障诊断效果, 及时地对高炉炉况进行监测与诊断.

关键词: 高炉炼铁过程；故障诊断；数据挖掘；决策树；迹距离函数；节点分裂

中图分类号: TP273　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.0843

引用格式: 刘亚雪, 张敬川, 王显鹏. 基于迹距离划分决策树的高炉故障诊断方法 [J]. 控制与决策, 2025, 40(5): 1533-1540.

Blast furnace fault diagnosis method based on trace distance partitioning decision tree

LIU Ya-xue[1,2]，*ZHANG Jing-chuan*[1,2]，*WANG Xian-peng*[2,3†]

(1. Frontier Science Center for Industrial Intelligence and Systems Optimization，Northeastern University，Shenyang 110819，China；2. Key Laboratory of Data Analytics and Optimization for Smart Industry，Northeastern University，Shenyang 110819，China；3. Liaoning Engineering Laboratory of Data Analytics and Optimization for Smart Industry，Shenyang 110819，China)

Abstract: With the development of industrial automation and intelligence, the decision tree model has been widely applied in the field of blast furnace(BF) fault diagnosis. However, for the characteristics of high dimensionality, nonlinearity, and strong coupling in the ironmaking process, the construction of the traditional decision tree model is easy to fall into the local optimal solution, with low efficiency and high complexity. To tackle the above issues, this paper firstly introduces the trace distance function and proves that any local optimal solution is also a global optimal solution in the trace distance function. It then proposes a decision tree model based on trace distance partitioning for the node splitting process in decision trees, referred to as TraceTree. On one hand, this model evaluates the division effect of a node more quickly and reduces the complexity of the decision tree model effectively. On the other hand, it can identify the features that contribute the most to fault diagnosis and obtain higher diagnosis accuracy. Finally, the comparison with other improved models shows that the model can achieve optimal diagnosis of BF faults with less training time, and monitor and diagnose the BF conditions in a timely manner.

Keywords: blast furnace ironmaking process；fault diagnosis；data mining；decision tree；trace distance function；node splitting

0　引　言

随着工业自动化和智能化水平的提高, 工业生产设备的复杂性和可靠性要求也在不断增加. 在这种背景下, 钢铁工业故障诊断系统的准确性和效率

收稿日期: 2024-07-16；**录用日期:** 2024-09-20.
基金项目: 国家自然科学基金项目 (62473086)；辽宁省科技计划联合计划项目 (2023JH2/101800047).
责任编委: 刘向杰.
†通信作者. E-mail: wangxianpeng@ise.neu.edu.cn.

对于保障钢铁生产安全和提高经济效益至关重要[1]. 作为钢铁生产的第 1 道工序, 高炉炼铁是一个连续且复杂的生产过程. 铁矿石、焦炭等原料从高炉顶部通过布料系统不断地注入高炉内部, 同高温热风中的氧气燃烧生成含有一氧化碳和氢气的高炉煤气, 高炉煤气在上升的过程中会与下降的铁矿石等炉料产生还原反应生成铁水[2], 而铁矿石中未还原的杂质、焦炭及煤中的灰分与石灰石等熔剂结合产生炉渣. 现代化高炉由于采用了精料技术, 吨铁渣量大幅度降低, 因此往往采用渣铁混出技术将熔融铁水和炉渣从高炉的出铁口一同排出. 与其他工业生产过程不同, 高炉本身的生产机理模型十分复杂, 涉及气体、固体和液体的共存[3], 其监测环境条件恶劣, 需要在密闭和高尘的环境下进行操作. 在高炉生产过程中, 原燃料成分的变化和高炉操作制度的改变等因素都会引起炉况的不稳定甚至引发高炉故障, 例如煤气流分布异常、崩料、悬料、管道等, 这会导致高炉产量下降、能耗增加, 甚至发生重大安全事故. 由于高炉炼铁过程涉及大量的监测参数, 具有高维度、非线性和强耦合的特点. 开发一种简单、高效的高炉故障诊断方法, 对于维持高炉炼铁过程的稳定顺行和钢铁工业的可持续发展具有重要意义.

随着工业自动化和智能化的发展, 利用机器学习技术对高炉故障进行诊断变得越来越重要[4-6]. 机器学习模型在高炉故障诊断模型的训练过程中首先要学到故障数据, 但高炉运行过程绝大多数时间均处于正常工况, 所以高炉生产过程的故障诊断数据集具有小样本问题以及样本间不平衡的问题. 目前国内外学者使用改进的机器学习算法在高炉故障诊断问题上取得了较好的效果, Lou 等[7] 提出了一种称为深度平稳核学习支持向量机 (DSKL-SVM) 的高炉故障诊断算法. Zhou 等[8] 设计了一种新的多输出最小二乘支持向量回归模型和评价方法, 提高了高炉故障识别的速度. Luo 等[9] 提出了基于 AdaBoost 的加权支持向量机集成预测器进行高炉内部热状态预测. Gao 等[10] 提出了一种工作点识别方法结合主成分分析 (PCA) 和谱聚类的算法, 该算法可以有效地降低高炉异常虚警率. Zhou 等[11] 提出一种基于核偏最小二乘 (KPLS) 的故障识别方法对高炉运行状态和铁水质量进行实时监测. Zhu 等[12] 提出了一种基于高斯混合模型对高炉故障监测的 MWPCA 算法. Shang 等[13] 提出了将变量增量代替绝对测量的统计分析方法 (RTCSA) 用于炼铁过程监测, 以实现高炉炼铁过程的早期异常检测.

决策树模型作为一种强大的数据挖掘工具, 能够通过对历史故障数据的学习有效地捕捉数据中隐藏的模式和规律, 自动构建出故障诊断规则, 为故障诊断提供了一种直观且易于理解的方法并快速分析故障原因, 提高了故障诊断的准确性和效率, 从而在故障诊断领域得到了广泛的应用[14-16]. 而构建决策树模型的核心关键是最优特征属性的划分策略, 一个好的划分策略不仅会使决策树的建立更高效, 而且还能降低模型的复杂度, 但对于炼铁过程中存在高维度、非线性和强耦合的特点, 传统的决策树模型在寻找最优特征属性时效率较低、复杂度较高且容易陷入局部最优解.

针对以上这些特性, 本文首先引入迹距离函数, 并证明在迹距离函数中任何局部最优解也是全局最优解的性质, 接着针对决策树的节点分裂过程, 提出一种基于迹距离划分的决策树模型 TraceTree, 并应用于高炉诊断模型中. 所提出的高炉故障诊断建模流程如图 1 所示. 首先, 从高炉生产过程数据采集系

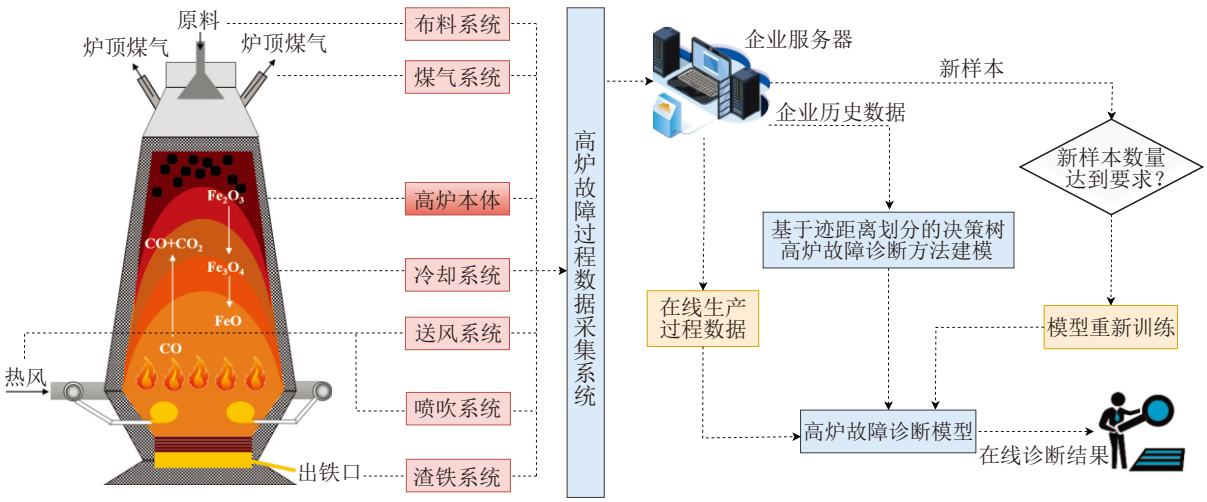

图1 基于迹距离划分决策树的高炉故障诊断示意图

统中获取所需数据集, 数据来源包括高炉的布料系统、煤气系统、冷却系统、送风系统、喷吹系统、渣铁系统, 以及高炉本体上遍布的各传感器采集得到的信息, 这些生产过程数据会与现场工况记录的高炉故障标签数据一起被储存在企业服务器中; 其次, 从企业服务器中获取历史生产过程数据, 并用该数据建立基于迹距离划分的决策树高炉故障诊断模型; 然后, 从高炉生产过程数据采集系统中获取各采样点上的实时采样值, 得到高炉在线生产过程数据, 将在线数据传递给高炉故障诊断模型, 从而向现场操作人员输出当前高炉炉况的在线诊断结果; 此外, 高炉生产过程数据采集系统会将所有新采样的样本储存在企业服务器中, 当新样本的数量满足要求时, 会结合历史数据和新数据来重新训练高炉故障诊断模型, 保证模型能够跟踪诊断高炉的最新工况.

1　背景介绍

1) ID3 算法. ID3 算法[17] 的核心是在决策树各个结点上使用信息增益作为划分结点的标准选择特征, 特征 A 对集合 D 的信息增益定义为

$$\mathrm{Gain}(D,A)=\mathrm{Ent}(D)-\mathrm{Ent}(D|A), \tag{1}$$

信息熵定义为

$$\mathrm{Ent}(D)=-\sum_{k=1} p_k\log_2 p_k, \tag{2}$$

其中p_k表示在样本集合 D 中第 k 类样本所占比例.

信息增益表示特征信息能够使得类信息不确定性减少的程度, 由此可见, 信息增益越大的特征具有更强的分类能力.

2) CART 算法. 基尼指数 (Gini index) 是 CART 算法[18] 中用于选择最佳分割的特征和分割点的一种度量, 其定义为

$$\mathrm{Gini}(D)=1-\sum_{k=1}\left(\frac{|C_k|}{|D|}\right)^2, \tag{3}$$

其中C_k是 D 中属于第 k 类的样本子集. $\mathrm{Gini}(D)$反映了从数据集 D 中随机抽取两个样本, 其类别标记不一致的概率. 基尼指数的值越小, 意味着 D 的纯度越高.

3) 基于决策树的高炉故障诊断方法. 决策树模型作为一种强大的数据挖掘工具, 能够通过对高炉运行状况相关的各类数据 (如炉壁温度、铁水产量、送风风量等历史故障数据) 的学习, 自动构建出故障诊断规则, 而这些规则可以通过树状结构更加清晰地展示故障特征与故障类型之间的逻辑关系. 此外, 模型在训练过程中, 会根据信息增益 (式 (1)) 或基尼系数 (式 (3)) 等度量, 选择最能区分故障类别的特征作为决策节点, 识别出对故障诊断最有贡献的特征参数, 将待诊断的高炉运行数据输入到训练好的决策树模型中, 模型会根据数据在决策树中进行递归判断, 最终给出对应的故障类别预测, 保证高炉故障诊断模型的快速离线迭代和实时在线预测.

2　所提出算法

本节首先引入迹距离函数, 并证明在迹距离函数中任何局部最优解也是全局最优解的性质; 然后给出基于迹距离函数的决策树节点分裂过程; 最后讨论 TraceTree 模型的整体框架.

2.1　迹距离函数

定义 1　对于任意的矩阵 A, B, 两矩阵之间的迹距离定义为

$$f=1-\mathrm{Tr}(\sqrt{A}\sqrt{B}), \tag{4}$$

其中$\mathrm{Tr}(\cdot)$表示矩阵的迹.

引理 1　对于优化问题$\min\limits_{x\in\chi} g(x)$, 如果 g 是凸函数且χ是凸集, 则任何局部最优解也是全局最优解. 进一步, 最优点集χ_{opt}也是凸的[19].

定理 1　迹距离函数的任何局部最优解也是全局最优解.

证明　对于任意的$\lambda\in[0,1]$, 有

$$f(\lambda A_1+(1-\lambda)A_2)\geqslant\lambda f(A_1)+(1-\lambda)f(A_2), \tag{5}$$

即迹距离f是凸函数.

对于矩阵集合$\Omega=\{A|\text{ 矩阵}A\text{的元素和为}1\}$, $\mu\in[0,1]$, 任意的$A_1, A_2\in\Omega$, 有

$$\mu A_1+(1-\mu)A_2\in\Omega, \tag{6}$$

即集合Ω是凸集.

由上述可知, 迹距离函数f是凸函数且集合Ω是凸集. 根据引理 1 可知, 迹距离函数的任何局部最优解也是全局最优的. □

2.2　基于迹距离函数的划分标准

1) 获得每类标签的样本完全分类正确的情形下, 在当前数据集样本个数中所占比例的矩阵 Q.

首先, 根据数据集已有的标签, 计算矩阵

$$A=\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k}\\ A_{21} & A_{22} & \cdots & A_{2k}\\ \vdots & \vdots & \ddots & \vdots\\ A_{k1} & A_{k2} & \cdots & A_{kk}\end{bmatrix}. \tag{7}$$

其中: k 表示当前训练集的故障标签类别数, 对于矩阵 A, 当$i\neq j$时, $A_{ij}=0$, $i,j=1,2,\ldots,k$, 即非对角线元素为 0, A_{ii}表示第 i 类故障标签的样本个数, 且$\sum\limits_i A_{ii}=n$, n表示样本个数.

进一步计算矩阵

$$Q=\begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1k} \\ Q_{21} & Q_{22} & \cdots & Q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{k1} & Q_{k2} & \cdots & Q_{kk} \end{bmatrix}. \tag{8}$$

其中: $Q_{ij}=A_{ij}/n$, $i=1,2,\ldots,k$, Q_{ii}表示每类标签的样本在当前数据集样本个数中所占的比例, 则概率和为 1, 即对角线元素和为 1.

2) 获得每类标签的样本不完全分类正确的情形下, 在当前数据集样本个数中所占比例的矩阵 P.

首先, 根据样本集 D 中的样本特征对当前数据集进行分类, 把 D 中的第 t 个特征X^t的第 s 个元素记作x_s^t. 其中: $X^t=(x_1^t,\ldots,x_s^t,\ldots,x_n^t)$, $t=1,2,\ldots,m$,$s=1,2,\ldots,n$, n表示训练集的样本个数, m表示样本维数. 按照x_s^t的大小将当前数据集划分为两个子数据集, 即将X^t中大于x_s^t的元素所对应的样本划分到左节点中, 将X^t中小于等于x_s^t的元素所对应的样本划分到右节点中, 以子节点下的数据集中出现频率最高的标签作为该子节点的诊断标签, 将划分后两个子节点的故障诊断标签与真实故障标签进行对比, 计算矩阵

$$B=\begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1k} \\ B_{21} & B_{22} & \cdots & B_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ B_{k1} & B_{k2} & \cdots & B_{kk} \end{bmatrix}, \tag{9}$$

其中B_{ij}表示把第 i 类故障标签的样本分到第 j 类故障标签中的个数, 且$\sum\limits_{ij} B_{ij}=n$, $i,j=1,2,\ldots,k$.

进一步计算矩阵

$$P=\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix}. \tag{10}$$

其中: $P_{ij}=B_{ij}/n$, $i=1,2,\ldots,k$, P_{ij}表示把第 i 类故障标签的样本分到第 j 类故障标签中的个数在当前样本集中所占的比例, 则概率和为 1, 即矩阵元素和为 1.

3) 计算两个矩阵之间的迹距离.

将矩阵P,Q代入式 (4), 当分类完全正确时, 即当$P=Q$时, $f=1-\mathrm{Tr}(\sqrt{P}\sqrt{Q})=1-\mathrm{Tr}(\sqrt{Q}\sqrt{Q})=1-\mathrm{Tr}(Q)=1-\sum\limits_i Q_{ii}=0$; 当分类不完全正确, 即$P\neq Q$时, 有$0<f=1-\mathrm{Tr}(\sqrt{P}\sqrt{Q})<1$, 这就意味着$P\to Q$时, $f\to 0$ (P 不断地趋近于 Q 时, 迹距离f的值趋近于 0).

当进行决策树的节点分裂时, 与传统的 CART 算法选择最小的基尼指数所对应的特征属性不同, TraceTree 算法选取最小的迹距离f值所对应的特征属性作为最优特征进行节点分裂, 意味着该特征属性不仅是在当前数据集中最具区分性的特征, 而且还是在故障诊断过程中最具有贡献的特征属性.

2.3 算法框架

算法 1 基于迹距离划分的分类决策树TraceTree(D,C,δ,ε).

输入: n行m列的样本集合D, 特征集C, 阈值δ和ε;

输出: 决策树TraceTree.

1) 创建根结点N;

2) if D中所有目标属性值相同或$C=\varnothing$ **then**

3) 将N记为叶结点, 其类别标记为D中样本数最多的类;

4) else

5) $t\leftarrow \mathrm{argmin}\{f(X^1),\ldots,f(X^m)\}$;

6) N.特征 $\leftarrow X^t$;

7) N.分割点 $\leftarrow s$;

8) $D_{\text{left}}\leftarrow D[D[X^t]<s]$;

9) $D_{\text{right}}\leftarrow D[D[X^t]>=s]$;

10) **if** $\mathrm{len}(D_{\text{left}}),\mathrm{len}(D_{\text{right}})<\delta$ 或 $f(X^t)<\varepsilon$ **then**

11) N.左子树 $\leftarrow$ 创建叶结点(D_{left});

12) N.右子树 $\leftarrow$ 创建叶结点(D_{right});

13) **else**

14) N.左子树 $\leftarrow$ 构建分类决策树TraceTree(D_{left}, $C-\{X^t\},\delta,\varepsilon$);

15) N.右子树 $\leftarrow$ 构建分类决策树TraceTree(D_{right}, $C-\{X^t\},\delta,\varepsilon$);

16) return TraceTree;

如算法 1 所示, 具体实现步骤如下:

step 1: 获取数据集并拆分为训练集与测试集, 将训练集作为当前样本集合 D, 所有特征属性的集合 C, 样本个数的阈值δ和阈值ε;

step 2: 对于每个特征, 计算所有可能的分割点, 对当前样本集合中的第 t 个特征$X^t=(x_1^t,x_2^t,\ldots,x_n^t)$进行遍历, 其中$t=1,2,\ldots,m$, m表示样本维数, n表示当前数据集的样本个数, 计算在特征X^t不同元素下的迹距离f值: $f_1^t,f_2^t,\ldots,f_n^t$(算法 1 第 5 行);

step 3: 选择迹距离f值最小的分割点作为当前节点的最佳分割点, 对当前样本集合中的每个特征进行遍历, 得到在所有特征下的最小迹距离f值, 即$f=\min\left(f^1,f^2,\ldots,f^m\right)$, 选择该迹距离$f$值下对应的特征作为最优特征, 其对应的分割点作为最优分割点 (算法 1 第 6、第 7 行);

step 4: 根据上一步获得的分割点将样本集合 D 分为两部分 D_{left} 和 D_{right}, 其中 D_{left} 是 D 的特征 X^t 中小于 s 的样本集合, D_{left} 是 D 的特征 X^t 中大于等于 s 的样本集合 (算法 1 第 8 、第 9 行);

step 5: 对于每个子集, 重复上述步骤, 直到满足停止条件, 停止条件是当前结点中的样本个数小于预定的阈值 δ, 或当前结点下的数据集的迹距离 f 值小于预定的阈值 ε(算法 1 第 10 ～ 第 15 行);

step 6: 生成基于迹距离划分的决策树模型 TraceTree.

3　实验与结果分析

3.1　评价指标

本文使用准确率、精确率、召回率和 F1-Score 这 4 个评估分类模型性能的常用指标, 它们不仅反映了模型的预测能力, 还能指导理解模型在实际应用中的可靠性和有效性.

1) 准确率 (accuracy)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{11}$$

其中: TP(true positive) 是真正例, TN(true negative) 是真负例, FP(false positive) 是假正例, FN(false negative) 是假负例.

2) 精确率 (precision)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{12}$$

3) 召回率 (Recall)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{13}$$

4) F1-Score 是精确率和召回率的调和平均值, 用于精确率与召回率之间的权衡, 有

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{14}$$

3.2　基准测试问题

为了评估所提出故障诊断算法的通用性, 首先将其在基准测试问题上进行测试. 选取 Wisconsin Breast Cancer Dataset 作为基准数据集, 该数据集在乳腺癌诊断领域具有广泛的应用, 是一个具有代表性的数据集. 将所提出的 TraceTree 算法与其他 3 种模型分别在该数据集上进行测试, 并使用准确率、精确率、召回率和 F1-Score 等指标来评估各算法的诊断效果.

3.2.1　对比算法

为了验证基于迹距离的决策树模型 (TraceTree) 的有效性和优越性, 引入 3 个对比算法: 决策树 (DecisionTree)、支持向量机 (SVM)、深度神经网络 (DNN). 其中: SVM 采用径向基函数作为核函数, DNN 为 4 层网络结构 (包括 1 个输入层、1 个输出层、2 个隐含层), 每层的节点数依次为 68-70-35-2.

3.2.2　实验结果分析

如表 1 所示, TraceTree 算法在所有关键性能指标上均表现出色, 其准确率达到了 93.71 %, 召回率为92 %, 这些指标的高水平表明 TraceTree 算法在正确分类样本方面具有极高的能力, 无论是在识别恶性肿瘤 (召回率) 还是在整体分类准确性 (准确率) 上. TraceTree 算法较高的 F1-Score 得分进一步证实了其在平衡精确率和召回率方面的卓越性能, 以及处理复杂分类问题时的高效性和适应性. 此外, 因为恶性肿瘤样本通常较少, TraceTree 在处理类别不平衡问题时表现出了优势, 这在乳腺癌数据集中尤为重要, 在实际的临床应用中具有较高的诊断准确性和可靠性. 由图 2 所示的诊断结果可以看出, 在所有的诊断效果中, TraceTree 中预测错误的点最少.

表1　各算法的实验结果对比　　%

模型	准确率	精确率	召回率	F1-Score
TraceTree	93.71	91.23	92.86	92.04
DecisionTree	92.31	89.47	91.07	90.27
SVM	90.91	100	76.79	86.87
DNN	91.61	89.29	89.29	89.29

结果分析显示, 该算法在区分良性和恶性肿瘤方面具有较高的准确性和可靠性. 其余 3 种算法虽然略低于 TraceTree, 但仍然显示出了可接受的分类能力. TraceTree 算法的优越性可能归因于其独特的树状结构以及具有全局最优解的迹距离的划分标准, 能够有效地捕捉数据中的复杂模式和非线性关系, 追踪到数据中的决策路径, 使得其具有较好的特征选择能力, 在处理数据时能够保持较高的准确性, 并提高模型的整体性能.

3.3　高炉生产过程故障诊断

3.3.1　数据采集

采集中国某钢厂高炉 2019-12-31 ～ 2023-10-14 的实际生产数据, 包括高炉运行过程中的 68 个过程参数 (如炉身各处温度、静压力等, 即模型的输入特征), 以及现场工况记录的高炉故障标签数据 (炉况正常或故障, 即模型的目标变量). 故障标签数据是由现场专家基于当日高炉运行状态与各项重要指标 (如铁水产量) 对高炉炉况平稳性的记录, 包括炉况正常和故障两种状态. 在采集的数据中, 不同类别参数的采集时间粒度不同, 温度、静压力等参数的采集间隔为 1 h, 铁水和渣的检化指标等参数的采集间隔为 1 个班次 (8 h), 故障标签数据的记录间隔为 1 天.

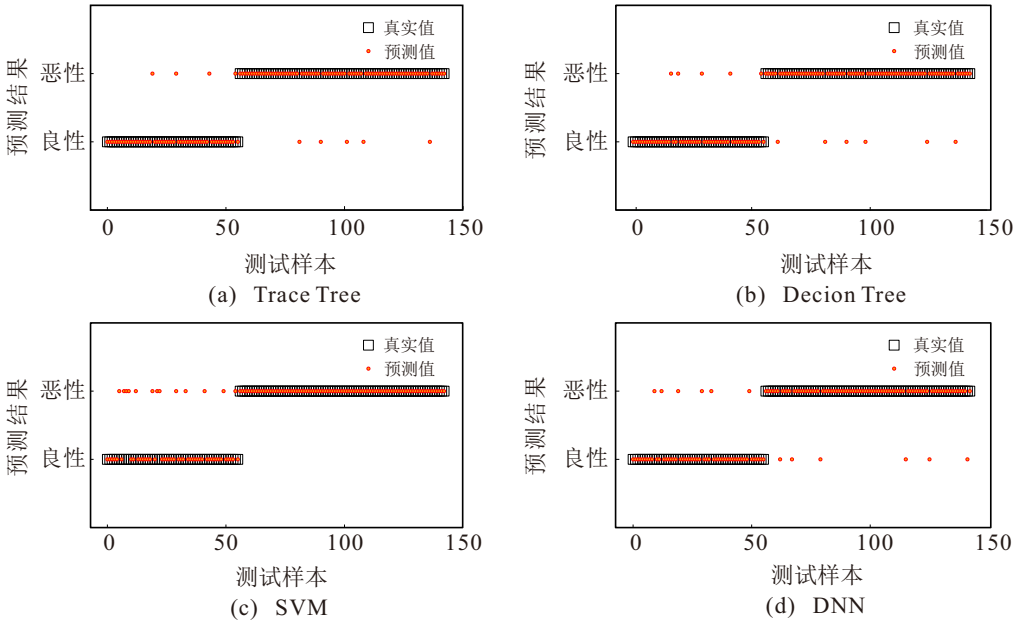

图2 诊断结果

3.3.2 数据预处理

由于高炉各采集项之间存在着采样时间上的不匹配问题, 首先需要将数据集的时间粒度对齐. 将每1 天中各采集项的均值作为当天的样本值, 与当天的高炉故障标签一起构成 1 条训练样本, 即得到所有参数的天粒度数据, 构成本文的高炉故障诊断数据集. 接着, 采用均值插值的方式填补数据集中的缺失值, 使用箱线图法剔除数据集中的异常值, 经过这两步预处理之后, 得到样本数为 1 380、特征数为 68 的数据集. 将该数据集的 75 % 作为训练样本, 25 % 作为测试样本.

3.3.3 对比算法

为了验证所提出模型 TraceTree 在高炉故障诊断问题上的优越性, 引入 3 个对比算法, 分别是以基尼指数选择划分属性的 CART 模型, 基于距离权值的 C4.5 组合决策树模型[20] 和基于 Pearson 相关系数的决策树模型 (PCCTree)[21].

3.3.4 实验结果分析

1) 将 TraceTree 模型分别与 PCCTree 模型、改进权值 C4.5 在高炉生产过程故障诊断数据集上进行对比, 各算法的实验结果整理如表 2 所示. 由结果可知, TraceTree 算法准确率、精确率和 F1-Score 指标在所有对比算法中均为最优, 表明所提出算法在高炉故障诊断上的优越性. 在构建过程中, TraceTree 算法能自动选择最具区分性的特征进行节点分裂, 更有效地识别出对故障诊断最有贡献的特征参数, 这种特征选择策略使得 TraceTree 算法成为提取故障特征的有效工具. PCCTree 模型和基于距离权值的 C4.5 组合决策树模型虽然略低于 TraceTree, 但仍然在高炉生产过程中显示出了可接受的故障诊断能力.

表2 各算法的实验结果对比 %

模型	准确率	精确率	召回率	F1-Score
TraceTree	96.23	96.21	98.83	97.50
PCCTree	93.71	96.08	87.5	91.59
距离权值C4.5	90.91	86.44	91.07	88.7

除此之外, TraceTree 算法优异的性能在 ROC 曲线上得到了显著的体现, 在 ROC 曲线下的 AUC 值为 0.94, 这意味着所提出算法在高炉故障诊断方面具有较高的准确性和可靠性. PCCTree 模型紧随其后, 其 AUC 值为 0.93, 基于距离权值的 C4.5 模型 AUC 值为 0.91, 这两种基于传统决策树模型的改进算法也在故障诊断任务中表现出良好的性能. 如图 3 所示, TraceTree 算法能够在保持较低假阳率的同时, 实现较高的真阳率, 这意味着它能够找出在故障诊断任务中最为关键的特征, 有效地识别出高炉故障情况, 同时减少了误报的可能性.

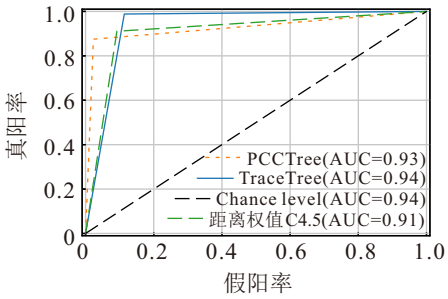

图3 ROC 曲线

2) 与传统的 CART 算法的模型训练时间做了对比实验. 首先, 对两种方法在不同数量的特征组成的

数据集下的模型训练时间做了对比实验, 设置特征数从 10 逐渐增加至 68, 对比两种方法的训练时间变化曲线, 如图 4 所示. 相比 CART 算法, 基于迹距离函数的决策树模型的训练时间在所有特征数量下都更少. 这表明 TraceTree 算法在处理特征维度较高的数据集时, 具有更低的时间复杂度, 表明该算法在计算特征间的相似度时更加高效, 从而减少了算法在构建树结构时的计算量. 此外, TraceTree 模型训练时间的稳定性可能意味着它在面对特征数量的增加时, 能够更好地控制模型的复杂度, 避免过拟合的同时保持了较快的训练速度, 这对于需要处理大规模数据集的应用场景尤为重要, 模型训练的时间效率直接影响到整个数据分析流程的效率.

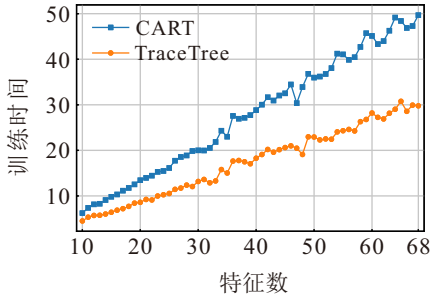

图4　不同特征数量下 TraceTree 与 CART 算法的训练时间对比

3) 将 TraceTree 算法与 CART 算法均在高炉故障诊断数据集独立运行 30 次以消除实验误差的影响, 并记录 30 次实验中两种方法的训练时间, 将实验结果用箱线图进行对比. 如图 5 所示, 基于迹距离函数的决策树模型在高炉数据集上的训练时间显著更少, 相对于 CART 算法具有明显的优势, 这种优势不仅体现在算法的执行速度上, 还体现在模型的泛化能力和预测准确性上. 这表明 TraceTree 算法能够通过对历史故障数据的学习, 自动构建出故障诊断规则, 更高效地学习到特征与故障类型之间的关系, 更快速地构建决策树故障诊断模型. 因此, TraceTree 算法在需要快速处理和响应的工业应用中, 即高炉生产监控与故障诊断中可以达到较好的效果, 保证高炉故障诊断模型的快速离线迭代和实时在线预测.

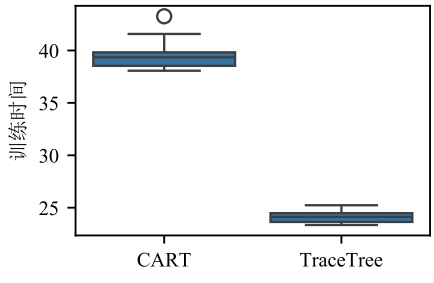

图5　算法 30 次重复试验运行时间

4) 在决策树模型中, 特征的重要性可以通过评估每个特征在模型中的决策贡献计算得到. 使用 TraceTree 算法建立故障诊断模型后, 查看重要性得分最高的 3 个特征: 入炉焦比、炉顶压力和送风风量. 入炉焦比是高炉炼铁的重要技术经济指标, 它反映了高炉操作水平的好坏; 炉顶压力是一个重要的操作参数, 它影响着高炉的稳定顺行和铁水产量; 送风风量是现场专家调节最为频繁的操作参数之一, 它很大程度上决定了高炉的产量. 可见, TraceTree 算法识别出来的 3 个重要特征与实际高炉生产过程中的重要参数是互相对应的, 表明所提出的 TraceTree 算法具有较好的实际应用前景.

4　结　论

针对高炉生产工艺的复杂性以及高炉生产过程故障诊断数据的特性, 本文改进了一种新的划分规则作为筛选的条件, 当对决策树的节点分裂时, 采用基于迹距离的划分标准来替代传统 CART 算法中基于基尼指数的划分标准, 提出一种新的基于迹距离决策树模型 (TraceTree) 并应用于高炉故障诊断领域. 此外, 实验结果验证了所提出算法在故障诊断方面的有效性, 在实际生产过程中可以帮助现场操作人员提前预知可能发生的高炉故障, 及时对高炉生产过程进行调节, 从而提高钢铁企业的生产效率.

参考文献 (References)

[1] Tang L X, Meng Y. Data analytics and optimization for smart industry[J]. Frontiers of Engineering Management, 2021, 8(2): 157-171.

[2] Yan F, Zhang X M, Yang C J, et al. Data-driven modelling methods in sintering process: Current research status and perspectives[J]. The Canadian Journal of Chemical Engineering, 2023, 101(8): 4506-4522.

[3] Zhou P, Li W P, Wang H, et al. Robust online sequential RVFLNs for data modeling of dynamic time-varying systems with application of an ironmaking blast furnace[J]. IEEE Transactions on Cybernetics, 2020, 50(11): 4783-4795.

[4] 毛文涛, 施华东, 张艳娜, 等. 轴承在线早期故障检测的无监督张量深度迁移学习方法[J]. 控制与决策, 2024, 39(3): 867-876.
(Mao W T, Shi H D, Zhang Y N, et al. Unsupervised tensor depth transfer learning method for online early fault detection of bearings[J]. Control and Decision, 2024, 39(3): 867-876.)

[5] 谢国民, 蔺晓雨. 基于改进 SSA 优化 MDS-SVM 的变压器故障诊断方法[J]. 控制与决策, 2023, 38(2): 459-467.
(Xie G M, Lin X Y. Transformer fault diagnosis method

based on improved SSA optimized MDS-SVM[J]. Control and Decision, 2023, 38(2): 459-467.)

[6] 王进花, 岳亮辉, 曹洁, 等. 基于随机变分推理贝叶斯神经网络的发电机轴承故障诊断[J]. 控制与决策, 2023, 38(4): 1015-1021.
(Wang J H, Yue L H, Cao J, et al. Fault diagnosis of generator bearing based on stochastic variational inference Bayesian neural network[J]. Control and Decision, 2023, 38(4): 1015-1021.)

[7] Lou S W, Yang C J, Wu P, et al. Fault diagnosis of blast furnace iron-making process with a novel deep stationary kernel learning support vector machine approach[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-13.

[8] Zhou P, Guo D W, Wang H, et al. Data-driven robust M-LS-SVR-based NARX modeling for estimation and control of molten iron quality indices in blast furnace ironmaking[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(9): 4007-4021.

[9] Luo S H, Dai Z A, Chen T X, et al. A weighted SVM ensemble predictor based on AdaBoost for blast furnace Ironmaking process[J]. Applied Intelligence, 2020, 50(7): 1997-2008.

[10] Gao L H, Chen H P. Abnormal detection of blast furnace condition using PCA similarity and spectral clustering[C]. The 13th IEEE Conference on Industrial Electronics and Applications. Wuhan. 2018: 2198-2203.

[11] Zhou P, Zhang R Y, Liang M Y, et al. Fault identification for quality monitoring of molten iron in blast furnace ironmaking based on KPLS with improved contribution rate[J]. Control Engineering Practice, 2020, 97: 104354.

[12] Zhu X, Zhang H, Yang C. MWPCA Blast furnace anomaly monitoring algorithm based on Gaussian mixture model[J]. CIESC Journal, 2021, 72(3): 1539-1548.

[13] Shang J, Chen M Y, Zhang H W, et al. Increment-based recursive transformed component statistical analysis for monitoring blast furnace iron-making processes: An index-switching scheme[J]. Control Engineering Practice, 2018, 77: 190-200.

[14] 王莹莹, 陈宏举, 杨旭光, 等. 基于决策树的水下控制模块液压系统故障诊断方法[J]. 船舶工程, 2022, 44(2): 154-164.
(Wang Y Y, Chen H J, Yang X G, et al. Fault diagnosis method of subsea control module hydraulic system based on decision tree[J]. Ship Engineering, 2022, 44(2): 154-164.)

[15] 张炎亮, 颜健勇. 基于 G-DPSO 算法的决策树轴承故障诊断方法[J]. 工业工程, 2021, 24(6): 41-47.
(Zhang Y L, Yan J Y. Fault diagnosis of bearing based on G-DPSO and decision tree[J]. Industrial Engineering Journal, 2021, 24(6): 41-47.)

[16] 庞梦洋, 索中英, 郑万泽, 等. 基于 RS-CART 决策树的航空发动机小样本故障诊断[J]. 航空动力学报, 2020, 35(7): 1559-1568.
(Pang M Y, Suo Z Y, Zheng W Z, et al. Small sample fault diagnosis of aeroengine based on RS-CART decision tree[J]. Journal of Aerospace Power, 2020, 35(7): 1559-1568.)

[17] Maddeh M, Ayouni S, Alyahya S, et al. Decision tree-based design defects detection[J]. IEEE Access, 2021, 9: 71606-71614.

[18] Gordon A D, Breiman L, Friedman J H, et al. Classification and regression trees[J]. Biometrics, 1984, 40(3): 874.

[19] Bertsekas D P. Convex optimization theory[M]. Raleigh: Athena Scientific, 2009.

[20] 杜景林, 严蔚岚. 基于距离权值的 C4.5 组合决策树算法[J]. 计算机工程与设计, 2018, 39(1): 96-102.
(Du J L, Yan W L. Multiple classifiers of C4.5 decision tree based on distance weight[J]. Computer Engineering and Design, 2018, 39(1): 96-102.)

[21] Mu Y S, Liu X D, Wang L D. A Pearson's correlation coefficient based decision tree and its parallel implementation[J]. Information Sciences, 2018, 435: 40-58.

作者简介

刘亚雪 (1995−), 女, 博士生, 主要研究方向为机器学习与多目标优化算法, E-mail: liuyaxue@stumail.neu.edu.cn;

张敬川 (1998−), 男, 博士生, 主要研究方向为多目标优化算法的改进与应用, E-mail: zhangjc@stumail.neu.edu.cn;

王显鹏 (1980−), 男, 教授, 博士生导师, 主要研究方向为基于机器学习的工业生产过程建模、操作优化, E-mail: wangxianpeng@ise.neu.edu.cn.