(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116204825 A
(43)申请公布日 2023.06.02

(21)申请号 202310122557.1
(22)申请日 2023.02.10
(71)申请人 湖北文理学院
地址 441053 湖北省襄阳市襄城区隆中路296号
(72)发明人 王峰 杭波 熊伟 项东升 黄金洲 花俏枝
(74)专利代理机构 武汉科皓知识产权代理事务所(特殊普通合伙) 42222
专利代理师 肖明洲
(51)Int.Cl.
G06F 18/241(2023.01)
G06F 18/2411(2023.01)
G06F 18/2431(2023.01)
G06F 18/214(2023.01)
G05B 23/02(2006.01)

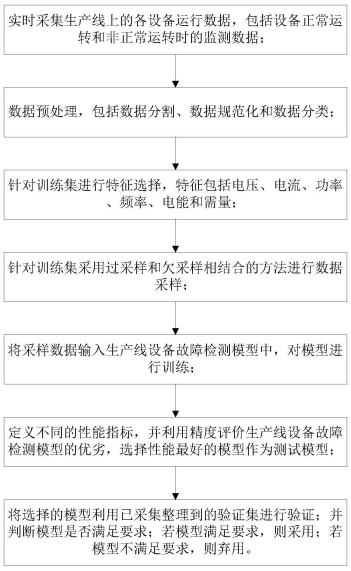
权利要求书4页 说明书13页 附图4页

(54)发明名称

一种基于数据驱动的生产线设备故障检测方法

(57)摘要

本发明公开了一种基于数据驱动的生产线设备故障检测方法,首先实时采集生产线上的各设备运行数据,包括设备正常运转和非正常运转时的监测数据;然后数据预处理;针对训练集进行特征选择;针对训练集采用过采样和欠采样相结合的方法进行数据采样;将采样数据输入生产线设备故障检测模型中,对模型进行训练;定义不同的性能指标,并利用精度评价生产线设备故障检测模型的优劣,选择性能最好的模型作为测试模型;最后将选择的模型利用已采集整理到的验证集进行验证;并判断模型是否满足要求;若模型满足要求,则采用;若模型不满足要求,则弃用。本发明通过采集电气设备中的状态数据可实现对生产线上电气设备运转状态的检测和预警。



CN 116204825 A

1.一种基于数据驱动的生产线设备故障检测方法，其特征在于，包括以下步骤：

步骤1：实时采集生产线上的各设备运行数据，包括设备正常运转和非正常运转时的监测数据；

步骤2：数据预处理，包括数据分割、数据规范化和数据分类；

所述数据分割，是将数据分为训练集、测试集和验证集；所述数据分类，是把已经知道的结果分成有设备故障和无设备故障两类数据；

步骤3：针对训练集进行特征选择，所述特征包括电压、电流、功率、频率、电能和需量；

采用方差分析特征选择ANOVA f-score的方法用于选择若干相关的特征，利用得分判断某项特征对于因变量而言是否重要，选取特征f得分最高变量的百分比作为训练特征；

步骤4：针对训练集采用过采样和欠采样相结合的方法进行数据采样；

步骤5：将采样数据输入所述生产线设备故障检测模型中，对模型进行训练；

所述生产线设备故障检测模型为：

$\Delta S_t^l = \left| S_t^l - S_0^{l*} \right|$；

其中，$S = \{U \mid U_l \mid \overline{U} \mid \overline{U_l} \mid I \mid \overline{I} \mid P \mid Q \mid S \mid PF \mid F_r \mid F_s \mid EP \mid EQ \mid W_{EP} \mid W_{EQ} \mid W_{VAh} \mid P_{xl} \mid Q_{xl} \mid S_{xl}\}$；

相电压U、线电压U_1、平均相电压$\overline{U}$、平均线电压$\overline{U_l}$；相电流I、平均相电流$\overline{I}$；有功功率P、无功功率Q、视在功率S、功率因数PF；各类电力参考值的实际频率F_r、各类传感器的采样频率F_s；有功电能EP、无功电能EQ、组合有功总电能W_{EP}、组合无功总电能W_{EQ}、视在总电能W_{VAh}；有功需量P_{xl}、无功需量Q_{xl}、视在需量S_{xl}；S的取值集合中“|”表示或者，表明S的取值可被集合中的任意值取其一替代；t＝T，T表示对应S取值的特征值采样周期，因此S_0^{l*}表示初始安装时某位置l的原始值，ΔS_t^l表示$S_t^l - S_0^{l*}$的差值变化量，若$\Delta S_t^l > \sigma^l$，σ^l为某待监测传感器可设定的预警阈值；l＝$\{L_{num}\}$，l用于标定检测传感器的位置，它用一串字符串编码集合L_{num}表示，num＝$\{f_n \sim w_n \sim pl_n \sim ep_n \sim c_n\}$，$f_n$表示工厂编号，$w_n$表示车间编号，$pl_n$表示流水线编号，$ep_n$表示设备编号，$c_n$表示元器件编号；

所述训练过程，包括输入、模型通配、数值计算、数据比较、参数优化和输出过程；首先参数输入：S_t^l, S_0^{l*}；然后进行模型通配：

$S = \{U \mid U_l \mid \overline{U} \mid \overline{U_l} \mid I \mid \overline{I} \mid P \mid Q \mid S \mid PF \mid F_r \mid F_s \mid EP \mid EQ \mid W_{EP} \mid W_{EQ} \mid W_{VAh} \mid P_{xl} \mid Q_{xl} \mid S_{xl}\}$；数值计算：$\Delta S_t^l = \left| S_t^l - S_0^{l*} \right|$，数据比较：$\Delta S_t^l (>, \leq) \sigma^l$和参数优化：$\sigma^l \rightarrow \sigma^{l*}$；最后结果输出：$W_t^l = \{1 \mid 0\}$；$W_t^l = \{1 \mid 0\}$表示输出结果从1或0中二选一，若结果为1，表明t时间l处的元器件预警，需要提醒更换；若结果为0，则表示元器件正常，不需要更换；

步骤6：定义不同的性能指标，并利用精度评价生产线设备故障检测模型的优劣，选择性能最好的模型作为测试模型；

步骤7：将选择的模型利用已采集整理到的验证集进行验证；并判断模型是否满足要求；分类准确率大于等于阈值，则认为满足模型要求；

若模型满足要求，则采用；

若模型不满足要求，则弃用。

2.根据权利要求1所述的基于数据驱动的生产线设备故障检测方法，其特征在于：步骤2中所述数据规范化，采用平均数插补法进行缺失值的计算和填充；采用最小值最大值度量标准化方法用于数据的标准化；

$$\overline{F_i^t}=\frac{F_i^t-F_{\min}^t}{F_{\max}^t-F_{\min}^t}\quad(1)；$$

其中，F_i^t表示电气设备运转状态值中在采样时间片t内的第i个特征值，$F_{\min}^t$表示在采样时间片t内第i个特征值的最小值，$F_{\max}^t$表示在采样时间片t内第i个特征值的最大值；$\overline{F_i^t}$表示采样时间片t内的第i个特征值经过上述公式标准化后的标准化值。

3.根据权利要求1所述的基于数据驱动的生产线设备故障检测方法，其特征在于：步骤2中所述数据分类，采用支持向量机执行分类任务；

所述支持向量机的线性表示形式为：

$f(x)=w^{Tr}\cdot x+b\quad(2)$；

其中，x表示输入变量，w表示权重矩阵，b表示偏差，Tr表示矩阵的转置；

所述支持向量机利用如下公式解决优化问题，

$$Min(\alpha\cdot\|w\|^2+P\sum_{i=1}^{n}(\delta_i^-+\delta_i^+))\quad(3)；$$

其中，P表示惩罚因子，δ_i^+和δ_i^-分别表示与训练数据相关的第i个特征值的正向惩罚和负向惩罚，||w||表示权重矩阵，α表示动态可调参数，它取值介于0-1之间；Min表示公式3需要满足$(w_i\cdot x_i+b)-y_i<\varphi+\delta_j^+$的可允许范围$\varphi$内的最小化取值，$\varphi$为具有取值上限的动态可调参数；$x_i$和$y_i$分别表示训练数据的第i个特征值的输入变量和输出变量；n表示特征集合总数。

4.根据权利要求1所述的基于数据驱动的生产线设备故障检测方法，其特征在于：步骤2中所述数据分类，采用多层感知神经网络N(k)；

$$N(k)=\sum_{j=1}^{m}\eta_j a_j\left(\sum_{i=1}^{n}w_i x_i(k)+w_0\right)+\eta_0\quad(4)；$$

其中，η_j表示多层感知神经网络在隐藏层中神经元j的权重系数，a_j表示对应神经元j的激活函数，n表示输入层中神经元的数目，m表示隐藏层中神经元的数目，w_i表示以神经元i为多层感知神经网络输入的输入层变量$x_i(k)$的权重，$x_i(k)$表示以神经元i为多层感知神经网络输入的输入层变量，w_0表示输入层的偏差，η_0表示输出层的偏差。

5.根据权利要求1所述的基于数据驱动的生产线设备故障检测方法，其特征在于：步骤2中所述数据分类，采用随机森林算法；

$$D(x)=\arg\max_{z}\sum_{l=1}^{L}H(d_l(x=Z))\quad(5)；$$

其中，D(x)表示随机森林作为分类模型的组合，d_l表示单棵决策树l对应的决策树分类模型，H表示指示函数，x表示输入变量；Z表示输出变量，小z表示Z的作用域，它是Z的一个集合分布，其含义为所有输出Z的范围取值作用域为小z；L表示决策树全集或者说决策树总数。

6.根据权利要求1所述的基于数据驱动的生产线设备故障检测方法，其特征在于：步骤2中所述数据分类，采用梯度提升树GDBT将一系列弱学习器转换为较强的学习器；

一棵GDBT树为：

$$\Upsilon=\sum_{\mu}\mu(z'_i,z_i)+\sum_{j}\partial(\beta_j)\quad(6)；$$

其中，若给定数据集具有n个实例和d个特征，则$\mu(z'_i,z_i)$是给定的凸型损失函数，$\{(z'_i,z_i)\}_{i=1}^{n}$，$z'_i\in\mathbb{R}^d$，$z_i\in\mathbb{R}$，即表示参数z'_i,z_i构造的参数对分别由特征集合$\mathbb{R}^d$和实例集合$\mathbb{R}$产生；$\partial(\beta_j)=\frac{1}{2}\chi_j\left\|V_j\right\|^2$，是一个正则化项，每一个$\beta_j$是一个与决策树相关的变量，$x_j$表示变量$\beta_j$对应的正则化参数，$V_j$表示变量$\beta_j$对应的叶子节点的权重，i，j分别表示随机森林中不同的节点编号。

7.根据权利要求1所述的基于数据驱动的生产线设备故障检测方法，其特征在于：步骤2中所述数据分类，采用随机欠采样提升树RUSBT将多个弱学习器组合成一个强学习器；

通过随机欠采样提升树，计算伪损失值λ_t为：

$$\lambda_t=\sum_{(i,b):b_i\neq b}E_t(i)(1-\theta_t(a_i,b_i)+\theta_t(a_i,b))\quad(7)；$$

其中，a_i表示特征空间A中的一个编号为i的点，b_i表示类标签集合B中的一个编号i为类标签，则训练数据集中的每一个实例都可以用元组(a_i,b_i)进行表示；θ_t表示弱假设$\theta_t(a_i,b_i)$和$\theta_t(a_i,b)$表示弱假设θ_t的输出；针对实例a_i，$E_t(i)$表示第i个实例迭代t次后的权重，t表示重复迭代次数；其中，类标签是做数据分类的时候给某个类别人为打的标签；

计算权重更新参数β_t：

$$\beta_t=\frac{1-\lambda_t}{1+\lambda_t}\quad(8)；$$

计算第i个实例迭代t+1次后的权重$E_{t+1}(i)$：

$$E_{t+1}(i)=E_{t+1}(i)\beta_t^{\frac{1}{2}(1+\theta_t(a_i,b_i)-\theta_t(a_i,b:b\neq b_i))}\quad(9)；$$

对第i个实例迭代t+1次后的权重$E_{t+1}(i)$进行规格标准化处理；

$$E_{t+1}(i)=\frac{E_{t+1}(i)}{\sum_i E_{t+1}(i)}\quad(10)；$$

计算最终假设$\Theta(a)$，

$$\Theta(a)=\arg\max_{b\in B}\sum_{t=1}^{T}\theta_t(a,b)\log\frac{1}{\beta_t}\quad(11)；$$

其中，$\Theta(a)$返回一个迭代T次弱假设后的权重值，T表示迭代次数的最大值。

8.根据权利要求1所述的基于数据驱动的生产线设备故障检测方法，其特征在于：步骤5中所述训练过程，训练和应用带默认参数的随机森林算法；优化模型中的超参数，包括支持向量分类器SVC中的正则化参数和核系数；并使用参数优化方法确定估计量的数量和最大树深度参数；通过网格搜索交叉验证方法对随机森林中的超参数进行优化，包括随机森林中单棵数的最大叶子节点数和最大深度；利用支持向量分类器SVC提供的若干超参数并使用Grid-Search CV方法进行参数调优；在MLP算法中采用Grid Search CV方法中进行参

数优化;所述参数主要包括节点连接性Connection、神经元单元数Number of units、数据点输入维度Input dimension、激活函数。

9.根据权利要求1所述的基于数据驱动的生产线设备故障检测方法,其特征在于:步骤6中所述定义不同的性能指标,包括:

$$Acc=\frac{TP+TN}{FP+FN+TP+TN} \quad (12);$$

$$Prec=\frac{TP}{TP+FP} \quad (13);$$

$$Rec=\frac{TP}{TP+FN} \quad (14)$$

其中,Acc表示分类的准确率,TP和FP分别为真阳性和假阳性,其代表的含义是,设备运转状态数据中,发生真实故障的数据条数占预测故障数据条数的百分比;TN和FN分别为真阴性和假阴性;Prec表示模型预测的精度;Rec表示计算模型的召回率,用于衡量,预测故障占真实故障的百分比;

通过F1来平衡精度和召回率;

$$F1=2\times\frac{Prec\times Rec}{Prec+Rec} \quad (15);$$

选择F1得分作为分类性能指标,最好的模型有最高的F1分数。

10.根据权利要求1-9任意一项所述的基于数据驱动的生产线设备故障检测方法,其特征在于,步骤7的具体实现包括以下子步骤:

步骤7.1:验证采样前后的随机森林算法的实验结果;分别采用:①过采样和随机森林,②欠采样、特征选择和随机森林,③仅使用随机森林;获得三组模型和方法在采用同体量样本进行实验的实验结果;

步骤7.2:验证采样前后的梯度提升树算法的实验结果;分别采用:①过采样和梯度提升树,②欠采样、特征选择和梯度提升树,③仅使用梯度提升树;获得三组模型和方法在采用同体量样本进行实验的实验结果;

步骤7.3:验证采样前后的梯度提升树算法的实验结果;分别采用:①欠采样、特征选择和随机森林,②欠采样、随机森林;获得两组模型和方法在采用同体量样本进行实验的实验结果;

步骤7.4:验证各种模型组合和方式的PR曲线和ROC曲线;分别采用:①随机森林、过采样和特征选择;②梯度提升树、过采样和特征选择;③随机森林、欠采样和特征选择;④梯度提升树、欠采样和特征选择;⑤随机欠采样提升树和特征选择;获得五组模型和方法在采用同体量样本进行实验的实验结果。

一种基于数据驱动的生产线设备故障检测方法

技术领域

[0001] 本发明属于生产设备故障检测技术领域，涉及一种生产线设备故障检测方法，具体涉及一种电气行业中基于数据驱动的生产线设备故障检测方法。

背景技术

[0002] 随着人们的生活越来越依赖工业产品，人们对高质量产品的期望也日益提高。因此，针对电气行业中生产线上的设备，提供一个故障率较低的健康生产线是非常必要的。然而，流水线作业的生产过程中，电气设备出现故障往往又是无法避免的。因为一旦设备出现问题，将会对生产线造成一定量的经济损失，并且高故障率而引发的产品原材料浪费也会导致产品残次率和设备能耗的升高。

[0003] 基于上述现象和原因，一种针对生产线管理的有效质量控制策略亟待研究。在生产线生产的过程中，可采用不同的方法监测电气设备的故障。有时候采用人工检测的方法进行设备故障检测，但这类方法通常效果不佳、价格昂贵且比较耗时。但是，若采用专业的质量测试设备，又可能需要对生产线进行大量的调整，同时也需要很高的前期投资。基于此，Chun等人研制了一种解决方案，该方法可针对每个设备运行环节进行故障检查，并只装运无故障检查的产品(参考文献1)。然而，Kang等人(参考文献2)认为由于设备检测过程不满意、检测质量控制标准差，以及生产环境不断发生变化，设备仍会随机产生一系列故障且无法预测。由于设备故障检测技术不成熟，而导致的结果会引发用户不满并产生经济纠纷(参考文献3)。由此，使得在生产线上利用低成本和高效率方法识别设备故障的技术具有一定的挑战性。

[0004] 为了克服上述利用人工方法进行设备故障诊断的问题，与预测分析相关的技术手段和方法越来越多地应用于不同的领域和场景中(参考文献4-6)。这些预测模型较好的针对工业生产过程中的设备故障进行了预测并取得了较好的效果(参考文献7)。Kang等人(参考文献8)对模型中重要变量的分析有助于发现设备产生故障的根本原因，并且有助于提高未来产品的质量。研究表明，机器学习领域中的相关算法有可能在工业生产线上针对设备故障诊断、评估和预测产品质量等问题产生良好的效果。在工业生产线上，大多数设备类型都会产生大量的工业生产数据。设备故障产生时，往往会在工业互联网中和工业生产线上产生一定的设备异常数据。因此，采用机器学习的相关算法可以较好的利用这些生成的数据构建预测模型。利用这类预测模型，可避免对生产线进行额外的修改同时，减少额外的人工成本投入。

[0005] 目前用于设备异常检测的算法主要为三类：

[0006] (1)基于过程模型的方法：这类方法是将测量系统的输出与数学模型的输出进行比较。然后，利用比较结果的残差对数学模型进行调整和改进。许多研究应用了不同的基于过程模型的方法，包括奇偶方程(参考文献9)，状态观测器(参考文献10)和参数估计(参考文献11-12)。

[0007] (2)基于知识的方法：这类方法是基于规则的，主要依赖于专家知识。这类模型容

易解释且运行效率高。然而，这类方法不够灵活且维护费用昂贵。Angeli等人(参考文献13)研制了一套应用上述方法于设备故障诊断的在线系统。根据Miljkovic等人(参考文献14)的研究成果表明，基于专家知识的方法更适用于定义良好的过程。

[0008] (3)基于数据驱动的方法：这类方法可分为信号分析，频谱分析，模式分析等一些子类方法。Isermann等人(参考文献15)给出了一些通过分析来自传感器的正常和故障信号来识别设备故障的研究。

[0009] 人工检测是在流水线上的设备出现故障之后，安排人员排查整条流水线，然后检修和更换故障的设备，费时费力、检修成本高是其最大的缺陷。

[0010] 自动检测是通过定位设备故障发生的具体设备和故障位置(而非具体要更换的元器件)，但不知道设备存在故障隐患的时间，也不知道具体是哪个元器件出了问题导致的设备故障，无法精准定位时间和故障位置，一旦自动检测到出现故障，需要人为定位故障位置来更换元器件。

[0011] 参考文献：

[0012] [1]Chun,Y.H.(2016).Improved method of estimating the product quality after multiple inspections.International Journal of Production Research,54(19),5686-5696.

[0013] [2]Kang,S.,Kim,E.,Shim,J.,Chang,W.,&Cho,S.(2018).Product failure prediction with missing data.International Journal of Production Research,56(14),4849-4859.

[0014] [3]Kang,S.,Kim,E.,Shim,J.,Chang,W.,&Cho,S.(2018).Product failure prediction with missing data.International Journal of Production Research,56(14),4849-4859.

[0015] [4]K¨oksal,G.,Batmaz,˙I.,&Testik,M.C.(2011).A review of data mining applications for quality improvement in manufacturing industry.Expert systems with Applications,38(10),13448-13467.

[0016] [5]Choudhary,A.K.,Harding,J.A.,&Tiwari,M.K.(2009).Data mining in manufacturing:Areview based on the kind of knowledge.Journal of Intelligent Manufacturing,20(5),501.

[0017] [6]Kusiak,A.(2006).Data mining:Manufacturing and service applications.International Journal of Production Research,44(18-19),4175-4191.

[0018] [7]Lughofer,E.,Pollak,R.,Zavoianu,A.C.,Meyer-Heye,P.,Z¨orrer,H.,Eitzinger,C.,...&Radauer,T.(2017,June).Self-adaptive time-series based forecast models for predicting quality criteria in microfluidics chip production.In:2017 3rd IEEE International Conference on Cybernetics(CYBCONF)(pp.1-8).IEEE.

[0019] [8]Kang,S.,Kim,E.,Shim,J.,Chang,W.,&Cho,S.(2018).Product failure prediction with missing data.International Journal of Production Research,56(14),4849-4859.

[0020] [9]Frank,P.M.(1990).Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy:A survey and some new results.Automatica,26(3),459-474.

[0021] [10]Isermann,R.(2005).Model-based fault-detection and diagnosis-status and applications.Annual Reviews in control,29(1),71-85.

[0022] [11]Isermann,R.(2006).Fault-diagnosis systems:An introduction from fault detection to fault tolerance.Springer Science&Business Media.

[0023] [12]Venkatasubramanian,V.,Rengaswamy,R.,Yin,K.,&Kavuri,S.N.(2003).A review of process fault detection and diagnosis:Part I:Quantitative model-based methods.Computers&Chemical Engineering,27(3),293-311.

[0024] [13]Angeli,C.(2010).Diagnostic expert systems:From expert's knowledge to real-time systems.Advanced Knowledge Based Systems:Model,Applications&Zesearch,1,50-73.

[0025] [14]Miljkovi′c,D.(2011).Fault detection methods:A literature survey.In 2011Proceedings of the34th international convention MIPRO(pp.750-755).IEEE.Natekin,A.,&Knoll,A.(2013).Gradient Boosting Machines,A Tutorial.Frontiers in neurorobotics,7,21.

[0026] [15]Isermann,R.(2006).Fault-diagnosis systems:An introduction from fault detection to fault tolerance.Springer Science&Business Media.

发明内容

[0027] 针对电气行业中的生产线设备故障检测问题的解决既要考虑检测方式灵活,又要控制检测成本的挑战性,同时结合现有检测技术(包括人工检测和自动检测等)的缺陷,本发明提出了一种基于数据驱动的生产线设备故障检测方法。

[0028] 本发明的方法所采用的技术方案是:一种基于数据驱动的生产线设备故障检测方法,包括以下步骤:

[0029] 步骤1:实时采集生产线上的各设备运行数据,包括设备正常运转和非正常运转时的监测数据;

[0030] 步骤2:数据预处理,包括数据分割、数据规范化和数据分类;

[0031] 所述数据分割,是将数据分为训练集、测试集和验证集;所述数据分类,是把已经知道的结果分成有设备故障和无设备故障两类数据;

[0032] 步骤3:针对训练集进行特征选择,所述特征包括电压、电流、功率、频率、电能和需量;

[0033] 采用方差分析特征选择ANOVA f-score的方法用于选择若干相关的特征,利用得分判断某项特征对于因变量而言是否重要,选取特征f得分最高变量的百分比作为训练特征;

[0034] 步骤4:针对训练集采用过采样和欠采样相结合的方法进行数据采样;

[0035] 步骤5:将采样数据输入所述生产线设备故障检测模型中,对模型进行训练;

[0036] 所述生产线设备故障检测模型为:

[0037] $\Delta S_t^l = \left| S_t^l - S_0^{l*} \right|$；

[0038] 其中，$S = \{U | U_l | \overline{U} | \overline{U_l} | I | \overline{I} | P | Q | S | PF | F_r | F_s | EP | EQ | W_{EP} | W_{EQ} | W_{VAh} | P_{xl} | Q_{xl} | S_{xl}\}$；相电压U、线电压$U_l$、平均相电压$\overline{U}$、平均线电压$\overline{U_l}$；相电流I、平均相电流$\overline{I}$；有功功率P、无功功率Q、视在功率S、功率因数PF；各类电力参考值的实际频率F_r、各类传感器的采样频率F_s；有功电能EP、无功电能EQ、组合有功总电能W_{EP}、组合无功总电能W_{EQ}、视在总电能W_{VAh}；有功需量P_{xl}、无功需量Q_{xl}、视在需量S_{xl}；S的取值集合中“|”表示或者，表明S的取值可被集合中的任意值取其一替代；t＝T，T表示对应S取值的特征值采样周期，因此S_0^{l*}表示初始安装时某位置l的原始值，ΔS_t^l表示$S_t^l - S_0^{l*}$的差值变化量，若$\Delta S_t^l > \sigma^l$，σ^l为某待监测传感器可设定的预警阈值；l＝{L_{num}}，l用于标定检测传感器的位置，它用一串字符串编码集合L_{num}表示，num＝{f_n～w_n～pl_n～ep_n～c_n}，f_n表示工厂编号，w_n表示车间编号，pl_n表示流水线编号，ep_n表示设备编号，c_n表示元器件编号；

[0039] 所述训练过程，包括输入、模型通配、数值计算、数据比较、参数优化和输出过程；首先参数输入：S_t^l, S_0^{l*}；然后进行模型通配：

[0040] $S = \{U | U_l | \overline{U} | \overline{U_l} | I | \overline{I} | P | Q | S | PF | F_r | F_s | EP | EQ | W_{EP} | W_{EQ} | W_{VAh} | P_{xl} | Q_{xl} | S_{xl}\}$；

数值计算：$\Delta S_t^l = \left| S_t^l - S_0^{l*} \right|$，数据比较：$\Delta S_t^l (>,\leq) \sigma^l$和参数优化：$\sigma^l \rightarrow \sigma^{l*}$；最后结果输出：$W_t^l = \{1|0\}$；$W_t^l = \{1|0\}$表示输出结果从1或0中二选一，若结果为1，表明t时间l处的元器件预警，需要提醒更换；若结果为0，则表示元器件正常，不需要更换；

[0041] 步骤6：定义不同的性能指标，并利用精度评价生产线设备故障检测模型的优劣，选择性能最好的模型作为测试模型；

[0042] 步骤7：将选择的模型利用已采集整理到的验证集进行验证；并判断模型是否满足要求；分类准确率大于等于阈值，则认为满足模型要求；

[0043] 若模型满足要求，则采用；

[0044] 若模型不满足要求，则弃用。

[0045] 相对于现有技术，本发明的有益效果是：

[0046] 1、本发明既可以知道设备出现的故障时间，又可以知道设备出现故障的具体元器件，同时还可以预测设备将在什么时候出现故障，设备中的元器件的故障概率，及设备和元器件的老化程度；

[0047] 2、本发明提出了电气行业中基于数据驱动的生产线设备故障检测方法，这套方法通过采集电气设备中的状态数据可实现对生产线上电气设备运转状态的检测和预警；

[0048] 3、本发明中所提出来的数据模型和方法：包括数据采样方法、特征选择方法、数据分析和预处理的思路和方法，均可移植到其它生产制造行业设备上进行生产制造型设备的运转状态的检测和预警；

[0049] 4、本发明中提出的基于数据驱动的生产线设备故障检测所使用的模型配置、模型优化和模型验证的实验过程和实验结果所采用的思路和方法同样适用于其它工业制造型企业在做生产线上设备故障检测方面的技术和方法应用。

附图说明

[0050] 图1为发明实施例的流程图；

[0051] 图2为本发明实施例的生产线设备故障检测模型训练原理图；

[0052] 图3为本发明实施例的采样前后的随机森林算法实验结果图；

[0053] 图4为本发明实施例的采样前后GDBT算法实验结果图；

[0054] 图5为本发明实施例的特征选择前后随机森林过采样实验结果图；

[0055] 图6为本发明实施例的各种模型组合和方式的PR曲线和ROC曲线图。

具体实施方式

[0056] 为了便于本领域普通技术人员理解和实施本发明，下面结合附图及实施例对本发明作进一步的详细描述，应当理解，此处所描述的实施示例仅用于说明和解释本发明，并不用于限定本发明。

[0057] 由于电气行业中的工业现场设备运行数据集中存在不同模式，并且是在机器学习等相关领域中的技术帮助下发现的。因此，在数据训练阶段需要使用大量数据点作为本发明研究的研究基础和前置条件。

[0058] 针对电气行业中的生产线设备故障检测问题的解决既要考虑检测方式灵活，又要控制检测成本的挑战性，同时结合现有检测技术(包括人工检测和自动检测等)的缺陷，本发明提出了一种基于数据驱动的生产线设备故障检测方法。该方法的实现过程主要分为数据预处理、特征选择、数据采样、参数优化等阶段。本发明所述数据均为电气行业中设备运转过程中产生的联网状态数据，该类数据被用于本发明所述生产线设备故障检测拟采用方法的数据来源。

[0059] 请见图1，本发明提供的一种基于数据驱动的生产线设备故障检测方法，包括以下步骤：

[0060] 步骤1：实时采集生产线上的各设备运行数据，包括设备正常运转和非正常运转时的监测数据；

[0061] 该阶段是本发明所有工作的数据准备阶段。通过在电气行业中的生产线中的各种设备上加装各种传感器作为设备运转状态的监测装置。当设备启动后开始运转时，这些传感器可通过内置的联网模块实现自组工业互联网络，并实时采集流水线上的各电气设备的运行数据，这些数据中包括设备正常运转和非正常运转(设备故障)时的监测数据集合。

[0062] 电气行业中的生产线设备运行状态数据大多来源于安装于设备生产或运行端的半导体传感器，这些传感器主要用于监测电气设备的运行状态和生产线的制造过程。与此同时，这类传感器产生的制造过程数据亦可反馈并用于电气设备的故障检测。通过前期大量研究基础和结果发现，要想利用流水线上的生产制造设备产生的制造过程数据用于设备的故障检测，需要克服四方面的挑战性问题。分别是：

[0063] ①数据集存在特征-样本比率过高的问题。研究发现，特征占到了数据集合的1/3，这也意味着用于训练每个特征的数据信息较少；

[0064] ②数据集高度不平衡问题，即数据集中设备正常运转的样本数据量远大于设备出现故障时的样本数据量；

[0065] ③数据集中很多特征都存在噪声信息或者不相关的信号，即设备状态数据中，只

有极少数的特征具备绝对相关性，且大部分特征间的关联性较弱。由此可以说明，大多数特征对于因变量来说是有噪声的或不相关的；

[0066] ④电气行业中的设备运转状态数据含有大量的缺失值，且度量不同特征的价值尺度有很大不同。有些传感器的刻度是千，而有些特征的刻度是十进制。

[0067] 步骤2：数据预处理，包括数据分割、数据规范化和数据分类；

[0068] 本实施例中数据分割，是假设生产线设备运行状态数据所构成的数据全集集合为I。那么，将该数据集合采用holdout方法划分为训练集I_{train}和测试集I_{test}（例如，训练集为全集的80％，测试集为全集的20％）。训练模型前，训练集I_{train}中的数据是经过预处理的数据。

[0069] 本实施例中数据规格化，采用平均数插补法进行缺失值的计算和填充。采用最小值最大值度量标准化方法用于数据的标准化。一般而言，最小最大值度量可用于将特征值规范化取值于[0,1]之间，如下公式1，

[0070] $$\overline{F_i^t}=\frac{F_i^t-F_{\min}^t}{F_{\max}^t-F_{\min}^t} \quad (1)$$

[0071] 其中，F_i^t表示电气设备运转状态值中在采样时间片t内的第i个特征值，$F_{\min}^t$表示在采样时间片t内第i个特征值的最小值，$F_{\max}^t$表示在采样时间片t内第i个特征值的最大值；$\overline{F_i^t}$表示采样时间片t内的第i个特征值经过上述公式标准化后的标准化值。

[0072] 本实施例中数据分类，目的是为模型的训练提供先验知识，换句话说，就是提前把已经知道的结果分成有设备故障和无设备故障的两类数据。这样分类的目的是为了告诉要训练的模型，要提取有设备故障的数据就到有设备故障的数据中去提取特征，要提取无设备故障的数据就到无设备故障的数据中去提取特征。

[0073] 本实施例中数据分类，采用支持向量分类器(Support Vector Classifier,SVC)、多层感知器(Multilayer Perceptron,MLP)、随机森林(Random Forest,RF)、梯度提升树(Gradient Boosted Trees,GDBT)和随机欠采样提升树(Random under-sampling boosting Tree,RUS-Boost Tree)等算法进行数据分类。其中，

[0074] 本实施例的支持向量分类器是一种支持向量机的实现，用于执行分类任务。它具有低计算成本和在低样本量的高维数据中具有较好分类效果的方法。如公式2所示，为支持向量机的线性表示形式。

[0075] $f(x)=w^{Tr}\cdot x+b$ (2)

[0076] 其中，x表示输入变量，w表示权重矩阵，b表示偏差，Tr表示矩阵的转置。该公式旨在允许范围φ内最小化误差值。本发明中拟利用如下公式3解决优化问题，

[0077] $$Min(\alpha\cdot\|w\|^2+P\sum_{i=1}^{n}(\delta_i^-+\delta_i^+)) \quad (3)$$

[0078] 其中，P表示惩罚因子，δ_i^+和δ_i^-分别表示与训练数据相关的第i个特征值的正向(+)惩罚和负向(-)惩罚，w表示权重矩阵，α表示动态可调参数，它取值介于0-1之间。Min表示公式3需要满足$(w_i\cdot x_i+b)-y_i<\varphi+\delta_j^+$的可允许范围$\varphi$（$\varphi$为具有取值上限的动态可调参数）内，对公式3的最小化取值。x_i和y_i分别表示训练数据的第i个特征值的输入变量和输出

变量;n表示特征集合总数。

[0079] 本实施例的多层感知器又名多层感知神经网络,由于其先进的深度学习算法可以应对许多难题,且可以增加系统算力,在近些年得以广泛应用。然而,神经网络相关算法一般对训练数据的数据量有一定要求。如下公式4给出了本发明提出的多层感知器,

[0080] $$N(k)=\sum_{j=1}^{m}\eta_j a_j\left(\sum_{i=1}^{n}w_i x_i(k)+w_0\right)+\eta_0 \quad (4)$$

[0081] 其中,η_j表示上述多层感知神经网络在隐藏层中神经元j的权重系数,a_j表示对应神经元j的激活函数,n表示输入层中神经元的数目,m表示隐藏层中神经元的数目,w_i表示以神经元i为神经网络输入的输入层变量$x_i(k)$的权重,$x_i(k)$表示以神经元i为神经网络输入的输入层变量,w_0表示输入层的偏差,η_0表示输出层的偏差。

[0082] 本实施例的随机森林是一种基于树的方法,并在训练过程中用于引导。随机森林由多棵树组成,每棵树只使用所有特征的一个子集。每棵树生成一个预测,因此最终的预测是所有预测的集合。随机森林由三个主要的参数组成,分别是树的大小,预测变量的数目和树的深度。由于随机森林可利用调整每个类的权重处理不平衡的数据,因此针对不平衡数据进行分类,随机森林的性能通常比采用带权重和平衡随机森林的单棵树预测算法更优,且其分类结果也优于其它算法。如公式5所示,给出了本发明拟构造的随机森林算法。

[0083] $$D(x)=\arg\max_{z}\sum_{l=1}^{L}H(d_l(x=Z)) \quad (5)$$

[0084] 从公式5中可以看出,D(x)表示随机森林作为分类模型的组合,d_l表示单棵决策树l对应的决策树分类模型,H表示指示函数,x表示输入变量;Z表示输出变量,小z表示Z的作用域,它是Z的一个集合分布,其含义为所有输出Z的范围取值作用域为小z;L表示决策树全集或者说决策树总数。

[0085] 本实施例的梯度提升树将一系列弱学习器转换为较强的学习器。每棵GDBT树可以改进了之前树算法的预测结果。这就使得GDBT树的表现更加灵活且近些年日益受欢迎。GDBT主要用于执行预后遗传任务,这些任务通常是不平衡的分类问题,且GDBT比随机森林和支持向量机具有更好的分类性能。GBDT算法旨在最小正则化目标函数。如公式6所示,为本发明拟构造的一棵GDBT树。

[0086] $$\Upsilon=\sum_{\mu}\mu(z'_i,z_i)+\sum_{j}\partial(\beta_j) \quad (6)$$

[0087] 其中,假定给定数据集具有n个实例和d个特征,那么$\mu(z'_i,z_i)$是给定的凸型损失函数。其中,$\{(z'_i,z_i)\}_{i=1}^{n}(z'_i\in\mathbb{R}^d,z_i\in\mathbb{R})$,即表示参数$z'_i$,$z_i$构造的参数对分别由特征集合$\mathbb{R}^d$和实例集合$\mathbb{R}$产生。$\partial(\beta_j)=\frac{1}{2}\chi_j\|V_j\|^2$,它是一个正则化项,每一个$\beta_j$是一个与决策树相关的变量。其中,$x_j$表示变量$\beta_j$对应的正则化参数,$V_j$表示变量$\beta_j$对应的叶子节点的权重,i,j分别表示随机森林中不同的节点编号。

[0088] 本实施例的随机欠采样提升树(RUSBT)是一种具有提升和欠采样功能的复杂树,这种树可以节省数据预处理的时间。RUSBT在提升前,通过执行随机欠采样用于对类进行平衡。在提升过程中,RUSBT将多个弱学习器组合成一个强学习器。如公式7所示,本发明构造了一棵随机欠采样提升树,用于计算伪损失值λ_t。公式(7)-(11)中的t表示的含义均为算法

重复迭代次数，公式12中的T表示迭代次数的最大值。

[0089] $$\lambda_t = \sum_{(i,b):b_i \neq b} E_t(i)(1-\theta_t(a_i,b_i)+\theta_t(a_i,b)) \quad (7)$$

[0090] 其中，a_i表示特征空间A中的一个编号为i的点，b_i表示类标签集合B中的一个编号i为类标签(类标签是做数据分类的时候给某个类别人为打的标签，例如，类标签可以为“设备故障”用1表示、“设备正常”用0表示，这里的布尔数值1和0就是类标签)，则训练数据集中的每一个实例都可以用元组(a_i,b_i)进行表示。θ_t表示弱假设(本发明中采用弱分类学习算法Weak-Learn进行训练)。$\theta_t(a_i,b_i)$和$\theta_t(a_i,b)$表示弱假设θ_t的输出。针对实例a_i(它可以是一个置信度评级的数字)来说，$E_t(i)$表示第i个实例迭代t次后的权重。公式8用于计算权重更新参数β_t。

[0091] $$\beta_t = \frac{1-\lambda_t}{1+\lambda_t} \quad (8)$$

[0092] 其中，λ_t表示的含义与公式7中相同。

[0093] $$E_{t+1}(i) = E_{t+1}(i)\beta_t^{\frac{1}{2}(1+\theta_t(a_i,b_i)-\theta_t(a_i,b:b\neq b_i))} \quad (9)$$

[0094] 其中，公式9中的各参数含义已在公式7和公式8中给出，此处不再赘述。

[0095] 公式10用于对第i个实例迭代t+1次后的权重$E_{t+1}(i)$进行规格标准化处理。

[0096] $$E_{t+1}(i) = \frac{E_{t+1}(i)}{\sum_i E_{t+1}(i)} \quad (10)$$

[0097] 公式11用于计算最终假设Θ(a)，

[0098] $$\Theta(a) = \arg\max_{b\in B} \sum_{t=1}^{T} \theta_t(a,b)\log\frac{1}{\beta_t} \quad (11)$$

[0099] 其中，Θ(a)返回一个迭代T次弱假设后的权重值，其它参数在公式(7)-(10)中有提及，此处不再赘述。

[0100] 步骤3：针对训练集进行特征选择，特征包括电压、电流、功率、频率、电能和需量；

[0101] 由于电气行业中的设备运行状态数据具有很高的特征-样本比率，即大量的特征与因变量的相关性很低。在这种情况下，在将数据应用于模型训练之前，需要进行特征选择，以减少不相关特征的数量。因此，本发明采用方差分析特征选择ANOVA f-score的方法用于选择大量相关的特征。其作用在于，利用得分判断某项特征对于因变量而言是否重要。即，更高的f值拒绝零假设，这也意味着变量方差对因变量方差有影响。因而，可选取f得分最高变量的百分比作为训练特征。

[0102] 在本发明中，方差分析ANOVA用于衡量一个特征与所有特征之间的相关性。鉴于此，可利用特征的f统计量满足f分布这一特性，用于显著性检验。

[0103] 在电气行业中，基于数据驱动的生产线设备故障检测方法需要提取用于设备故障诊断的特征包括但不限于以下特征：

[0104] 1)电压：包括相电压U、线电压U_l、平均相电压$\overline{U}$、平均线电压$\overline{U_l}$；

[0105] 其中，相电压U可通过相电压谐波含量H_U给予特征反馈。当上述电压实际值稳定，但利用设备上的传感器采集到的电压数值却在某些时刻出现局部极大值或极小值，从数据

分析的角度显示设备故障，因而需要对采集设备中出现瞬时极值的传感器进行更换；

[0106] 2)电流：包括相电流I、平均相电流$\overline{I}$；其中，相电流可通过相电流谐波含量给予特征反馈。与电压值

[0107] 类似，若电气设备上的电流值与原始值相比，在某时刻出现局部极值时，则可认定此刻需对相应的传感器进行更换。

[0108] 3)功率：包括有功功率P、无功功率Q、视在功率S、功率因数PF；

[0109] 4)频率：包括各类电力参考值的实际频率F_r、各类传感器的采样频率F_s；

[0110] 5)电能：包括有功电能EP、无功电能EQ、组合有功总电能W_{EP}、组合无功总电能W_{EQ}、视在总电能W_{VAh}；

[0111] 6)需量：包括有功需量P_{xl}、无功需量Q_{xl}、视在需量S_{xl}；

[0112] 上述1-6是电气设备在电力行业中实施应用过程中产生的检测特征值；在此，定义原始值，它是指设备安装之初通过各类传感器采集到的测量初始值。电气行业中，一般认为全新设备上的采集装置采集到的数据为准确值，这类测量值可作为各类特征指标的原始参考值被作为参考标准。若检测到的采样值和原始值之间出现明显的特征差异(如，采样极值、采样值缺失、采样值不可读，等等)，则可基本可认定相应的电气元器件出现老化，应定位并予以更换。

[0113] 电气设备除了在电力行业中可按上述模式定义和提取数据的特征变量用作设备故障检测的依据以外，在其它行业，如煤炭、水利水电、化工等行业，同样可采用上述模式定义和提取特征变量，并采用本发明步骤5中的生产线设备故障检测模型实施生产线设备的故障检测方法。

[0114] 步骤4：针对训练集采用过采样和欠采样相结合的方法进行数据采样；

[0115] 由于电气行业中生产线上的设备运行获取到的训练数据高度不平衡性，据调研发现，所有设备运转状态数据中，数据点属于设备异常数据类的仅占7％。针对这种现象，本发明采用过采样和欠采样相结合的方法来改进算法的性能。具体实现该步骤使用到的方法细节，详述如下：

[0116] 本实施例的欠采样，欠采样的采样了所有少数类样本，并随机选取相等数量的多数类样本。然后，它结合两个采样子集，形成一个新的平衡数据集。在该训练集中，正常样本属于多数类，故障样本属于少数类。换言之，欠采样方法的实现原理是，通过丢失一些数据点和有用信息来平衡数据。

[0117] 本实施例的过采样，与欠采样相反，过采样通过复制少数类样本来平衡数据。虽然过采样方法平衡了数据，增加了数据量；但其缺点是，由于数据的机械复制，容易出现过拟合。因此，本发明拟采用合成少数过采样技术SMOTE用于解决过拟合问题。该技术通过合成相似的数据点，即通过选定的某个少数点利用欧式距离找到它的k近邻，然后在其中创建一个或多个新的点，从而克服数据点机械复制所导致的过拟合问题。

[0118] 步骤5：将采样数据输入生产线设备故障检测模型中，对模型进行训练；

[0119] 本实施例的生产线设备故障检测模型为：

[0120] $\Delta S_t^l = \left| S_t^l - S_0^{l*} \right|$；

[0121] 其中，$S = \{U \mid U_l \mid \overline{U} \mid \overline{U_l} \mid I \mid \overline{I} \mid P \mid Q \mid S \mid PF \mid F_r \mid F_s \mid EP \mid EQ \mid W_{EP} \mid W_{EQ} \mid W_{VAh} \mid P_{xl} \mid Q_{xl} \mid S_{xl}\}$；

相电压U、线电压U_1、平均相电压$\overline{U}$、平均线电压$\overline{U_l}$;相电流I、平均相电流$\overline{I}$;有功功率P、无功功率Q、视在功率S、功率因数PF;各类电力参考值的实际频率F_r、各类传感器的采样频率F_s;有功电能EP、无功电能EQ、组合有功总电能W_{EP}、组合无功总电能W_{EQ}、视在总电能W_{VAh};有功需量P_{xl}、无功需量Q_{xl}、视在需量S_{xl};S的取值集合中"|"表示或者,表明S的取值可被集合中的任意值取其一替代;t=T,T表示对应S取值的特征值采样周期,因此S_0^{l*}表示初始安装时某位置l的原始值,ΔS_t^l表示$S_t^l - S_0^{l*}$的差值变化量,若$\Delta S_t^l > \sigma^l$,σ^1为某待监测传感器可设定的预警阈值;l={L_{num}},l用于标定检测传感器的位置,它用一串字符串编码集合L_{num}表示,num={f_n～w_n～pl_n～ep_n～c_n},f_n表示工厂编号,w_n表示车间编号,pl_n表示流水线编号,ep_n表示设备编号,c_n表示元器件编号;例如,一串001～011～035～026～20141021的编码,可用于标识异常元器件出现的位置在工厂编号为001,车间编号为011,流水线编号为035,设备编号为026,元器件编号为20141021(初始使用日期),将上述编码制成二维码,则可实现生产线设备的故障追踪和远程监测。

[0122] 请见图2,本实施例的训练过程,包括输入、模型通配、数值计算、数据比较、参数优化和输出过程;

[0123] 首先参数输入:S_t^l, S_0^{l*};然后进行模型通配:

[0124] $S=\{U|U_l|\overline{U}|\overline{U_l}|I|\overline{I}|P|Q|S|PF|F_r|F_s|EP|EQ|W_{EP}|W_{EQ}|W_{VAh}|P_{xl}|Q_{xl}|S_{xl}\}$;

数值计算:$\Delta S_t^l = \left|S_t^l - S_0^{l*}\right|$,数据比较:$\Delta S_t^l(>,\leq)\sigma^l$和参数优化:$\sigma^1 \rightarrow \sigma^{1*}$;最后结果输出:$W_t^l = \{1|0\}$;

[0125] 在图2中,(>,≤)表示比较结果二选一,σ^{1*}表示参数优化的目标设定值,该取值由数值特征的聚类结果确定。$W_t^l = \{1|0\}$表示输出结果从1或0中二选一,若结果为1,表明t时间l处的元器件预警,需要提醒更换;若结果为0,则表示元器件正常,不需要更换。

[0126] 在训练过程中,训练和应用带默认参数的随机森林算法;优化模型中的超参数(包括支持向量分类器中的正则化参数(用于决定了正则化的强度)和核系数(用于控制核的宽度)),并使用参数优化方法确定估计量的数量和最大树深度参数;通过网格搜索交叉验证(Grid-Search CV)方法对随机森林中的超参数(包括随机森林中单棵数的最大叶子节点数和最大深度)进行优化;利用支持向量分类器SVC(Support Vector Classifier,SVC)提供的若干超参数并使用Grid-Search CV方法进行参数调优;在MLP算法中采用Grid Search CV方法中进行参数优化;参数主要包括节点连接性Connection、神经元单元数Number of units、数据点输入维度Input dimension、建模过程中所需要用到的激活函数(包括Relu、Linear等)。

[0127] 步骤6:定义不同的性能指标,并利用精度评价生产线设备故障检测模型的优劣,选择性能最好的模型作为测试模型;

[0128] 对于二分类问题,本发明采用混淆矩阵定义不同的性能指标,并利用精度评价分类模型优劣的性能指标。如公式(13)-(16),分别给出本发明评估上述所列分类模型用于电气行业中,生产线设备故障检测应用的性能。

[0129] $$Acc=\frac{TP+TN}{FP+FN+TP+TN}\quad(13)$$

[0130] $$Prec=\frac{TP}{TP+FP}\quad(14)$$

[0131] 其中,Acc表示分类的准确率,Prec表示模型预测的精度。TP和FP分别为真阳性和假阳性。其代表的含义是,设备运转状态数据中,发生真实故障的数据条数占预测故障数据条数的百分比。

[0132] $$Rec=\frac{TP}{TP+FN}\quad(15)$$

[0133] 其中,Rec表示计算模型的召回率。TN和FN分别为真阴性和假阴性;公式15用于衡量,预测故障占真实故障的百分比。在实际应用中,提高召回率往往会降低精度,因为高召回率需要较低的阈值。因此,本发明通过引入F1来平衡准确率和召回率。

[0134] $$F1=2\times\frac{Prec\times Rec}{Prec+Rec}\quad(16)$$

[0135] 选择F1得分作为分类性能指标,通过公式15的计算可知,最好的模型应该有最高的F1分数。

[0136] 值得注意的是,公式14中的准确率Prec和公式15中的召回率Rec通常用于评估不平衡分类任务的性能。

[0137] 步骤7:将选择的模型利用已采集整理到的验证集进行验证;并判断模型是否满足要求,分类准确率大于等于95%,则认为满足模型要求;

[0138] 若模型满足要求,则采用;

[0139] 若模型不满足要求,则弃用。

[0140] 本实施例中,步骤7具体包括以下步骤:

[0141] 步骤7.1:验证采样前后的随机森林算法的实验结果(如图3所示)。分别采用:①过采样(Over-Sampling)和随机森林(RF),②欠采样(Under-Sampling)、特征选择(FS)和随机森林(RF),③仅使用随机森林(RF),三组模型和方法在采用同体量样本进行实验的实验结果可以看出:

[0142] 准确率(Accuracy)高低排序为:①②③;

[0143] 精度(Precision)高低排序为:①②③;

[0144] 召回率(Recall)高低排序为:③②①;

[0145] F1得分(F1-Score)高低排序为:①②③;

[0146] 由此可知,采用第①组模型和方法的实验效果普遍较好,但采用第②、③组模型和方法可以弥足①在召回率方面的不足。

[0147] 步骤7.2:验证采样前后的梯度提升树算法的实验结果(如图4所示)。分别采用:①过采样(Over-Sampling)和梯度提升树(GDBT),②欠采样(Under-Sampling)、特征选择(FS)和梯度提升树(GDBT),③仅使用梯度提升树(GDBT),三组模型和方法在采用同体量样本进行实验的实验结果可以看出:

[0148] 准确率(Accuracy)高低排序为:①③②;

[0149] 精度(Precision)高低排序为:①②;

[0150] 召回率(Recall)高低排序为:②①;

[0151] F1得分(F1-Score)高低排序为:①②;

[0152] 由此可知,采用第①组模型和方法的实验效果普遍较好,但采用第②组模型和方法可以弥足①在召回率方面的不足,采用第③组模型在准确率方面仍然可以达到第①组模型和方法几乎相当的效果。

[0153] 步骤7.3:验证采样前后的梯度提升树算法的实验结果(如图5所示)。分别采用:①欠采样(Under-Sampling)、特征选择(FS)和随机森林(RF),②欠采样(Under-Sampling)、随机森林(RF),两组模型和方法在采用同体量样本进行实验的实验结果可以看出:

[0154] 准确率(Accuracy)高低排序为:①②;

[0155] 精度(Precision)高低排序为:②①;

[0156] 召回率(Recall)高低排序为:①②;

[0157] F1得分(F1-Score)高低排序为:①②;

[0158] 由此可知,采用第①组模型和方法的实验效果普遍较好,但采用第②组模型和方法可以弥足①在精度方面的不足。

[0159] 步骤7.4:验证各种模型组合和方式的PR曲线和ROC曲线(如图6所示)。分别采用:①随机森林(RF)、过采样(OS)和特征选择(FS);②梯度提升树(GDBT)、过采样(OS)和特征选择(FS);③随机森林(RF)、欠采样(US)和特征选择(FS);④梯度提升树(GDBT)、欠采样(US)和特征选择(FS);⑤随机欠采样提升树(RUS)和特征选择(FS)。

[0160] 五组模型和方法在采用同体量样本进行实验的实验结果可以看出:

[0161] 在召回率(Recall)方面,③④⑤的召回率均可达到100%,但同等条件下,准确率(Accuracy)高低排序为:④③⑤。而①的召回率次于③④⑤,仅78%左右;最差的是②,仅68%左右;

[0162] 通过比较准确率(Accuracy)和召回率(Recall)的比值可以发现,随着召回率的提升,准确率高低排序为②①④③⑤,

[0163] 由此可知,采用第④组模型和方法的实验效果普遍较好,第①组模型和方法虽然最终可达到较好的准确率,但召回率不高。而采用第④组模型和方法可以弥足①在召回率方面的不足。

[0164] 通过图6中各模型的ROC曲线比较可以看出:

[0165] 第①组模型和方法的FPR(真阳性比例)升高时,TPR(假阳性比例)也随之升高,但两种比例都均不高;

[0166] 第②组模型和方法的FPR(真阳性比例)为0.28以前,均未出现TPR(假阳性比例),但此后TPR激增至0.55,但第②组的FPR最高仅到0.48,但TPR却高达0.8;

[0167] 第③组模型和方法的FPR(真阳性比例)为0.58之前,均未出现TPR(假阳性比例),但此后TPR激增至0.82,在这之后虽然FPR达到了1,但TPR也高达0.92;

[0168] 第④组模型和方法的FPR(假阳性比例)在0.15之前,未出现TPR(真阳性比例)。在此之后FPR到0.23,其TPR才升高到0.43。此后,TPR稳步上升至1,但FPR也最终高达0.92;

[0169] 第⑤组模型和方法的FPR(假阳性比例)在0.44之前,未出现TPR(真阳性比例),其TPR才升高到0.18。但随后在0.18位置,FPR也达到0.57。此后,随着TPR的升高,FPR稳步上升。最终,当TPR为1时,FPR也达到了0.9。

[0170] 本发明的研究内容涵盖电气行业中利用数据驱动模型针对生产线上关于设备故障检测的相关技术和方法。在本发明的阐述中涉及到数据分析、数据采样、数据预处理、数据特征选择、数据分类、数据模型的有效性验证等多个方面的核心算法和关键技术手段。

[0171] 应当理解的是，上述针对较佳实施例的描述较为详细，并不能因此而认为是对本发明专利保护范围的限制，本领域的普通技术人员在本发明的启示下，在不脱离本发明权利要求所保护的范围情况下，还可以做出替换或变形，均落入本发明的保护范围之内，本发明的请求保护范围应以所附权利要求为准。


实时采集生产线上的各设备运行数据，包括设备正常运转和非正常运转时的监测数据；
数据预处理，包括数据分割、数据规范化和数据分类；
针对训练集进行特征选择，特征包括电压、电流、功率、频率、电能和需量；
针对训练集采用过采样和欠采样相结合的方法进行数据采样；
将采样数据输入生产线设备故障检测模型中，对模型进行训练；
定义不同的性能指标，并利用精度评价生产线设备故障检测模型的优劣，选择性能最好的模型作为测试模型；
将选择的模型利用已采集整理到的验证集进行验证；并判断模型是否满足要求；若模型满足要求，则采用；若模型不满足要求，则弃用。

图1

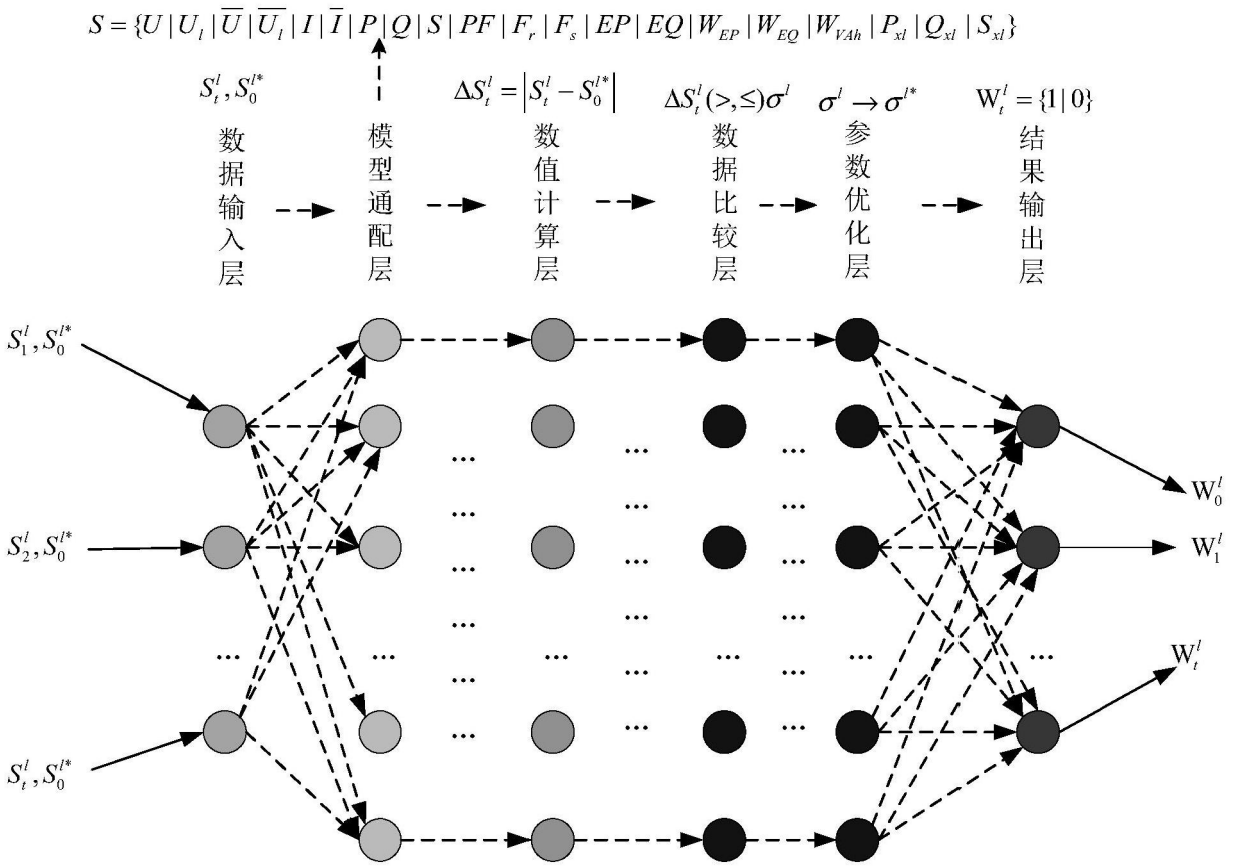

$S=\{U|U_l|\overline{U}|\overline{U_l}|I|\overline{I}|P|Q|S|PF|F_r|F_s|EP|EQ|W_{EP}|W_{EQ}|W_{VAh}|P_{xl}|Q_{xl}|S_{xl}\}$
S_t^l, S_0^{l*}
$\Delta S_t^l = \left|S_t^l - S_0^{l*}\right|$
$\Delta S_t^l(>,\leq)\sigma^l$
$\sigma^l \rightarrow \sigma^{l*}$
$W_t^l = \{1|0\}$
数据输入层
模型通配层
数值计算层
数据比较层
参数优化层
结果输出层
S_1^l, S_0^{l*}
S_2^l, S_0^{l*}
S_t^l, S_0^{l*}
W_0^l
W_1^l
W_t^l

图2

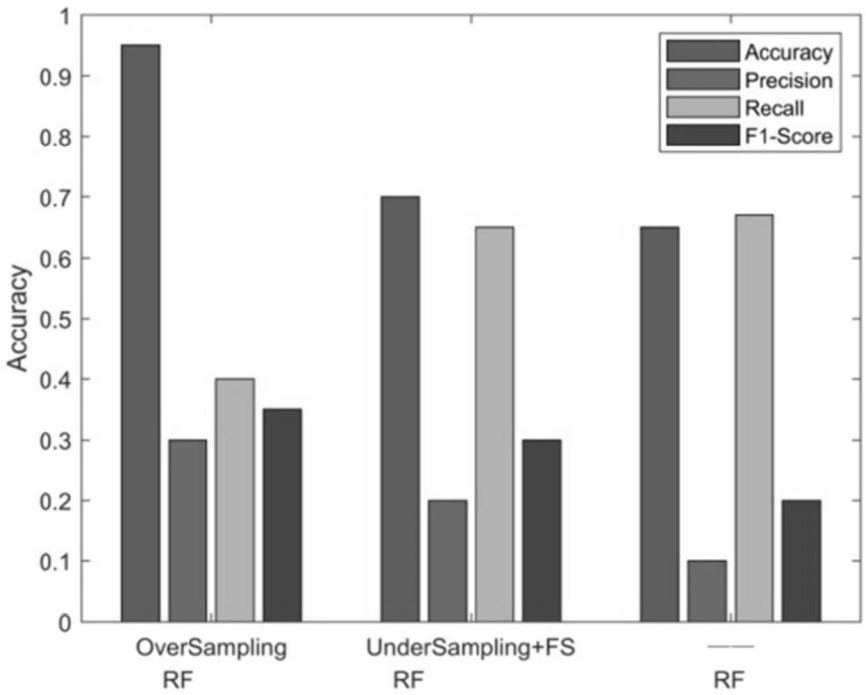

Accuracy
Precision
Recall
F1-Score
Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
OverSampling
RF
UnderSampling+FS
RF
RF

图3

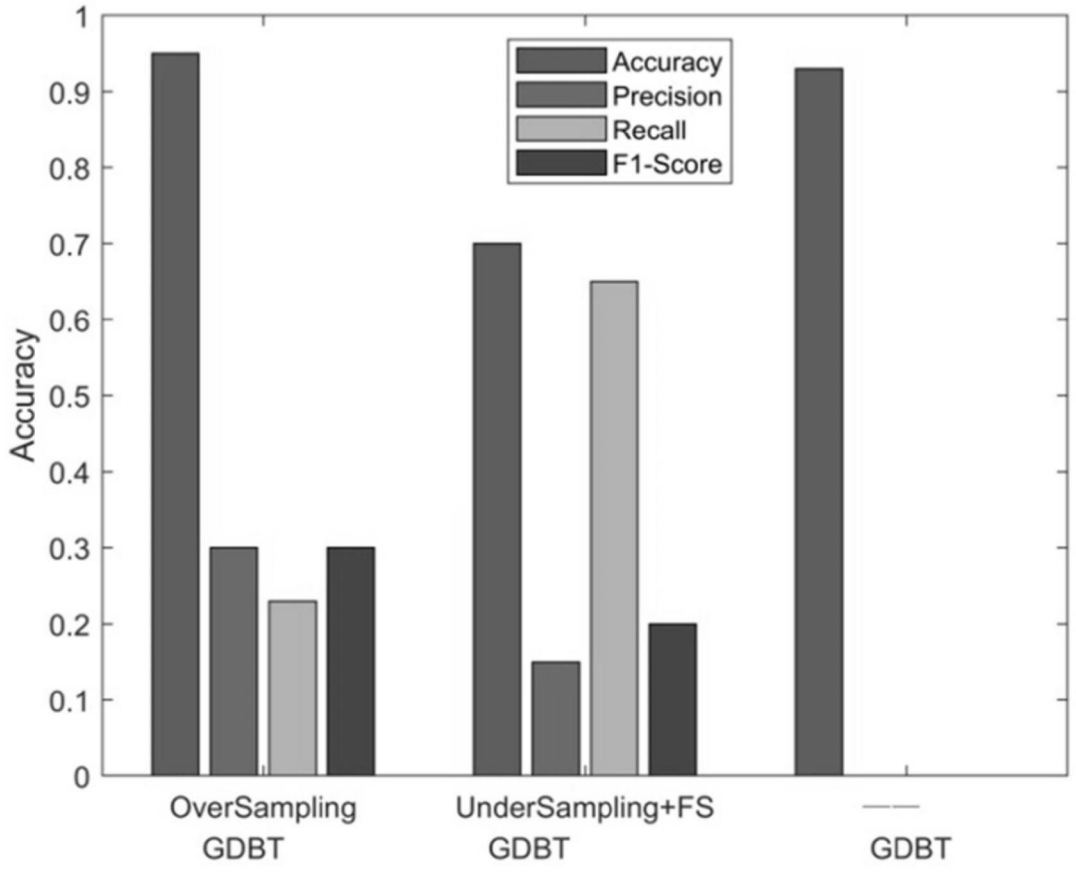

Accuracy
Precision
Recall
F1-Score
Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
OverSampling
GDBT
UnderSampling+FS
GDBT
GDBT

图4

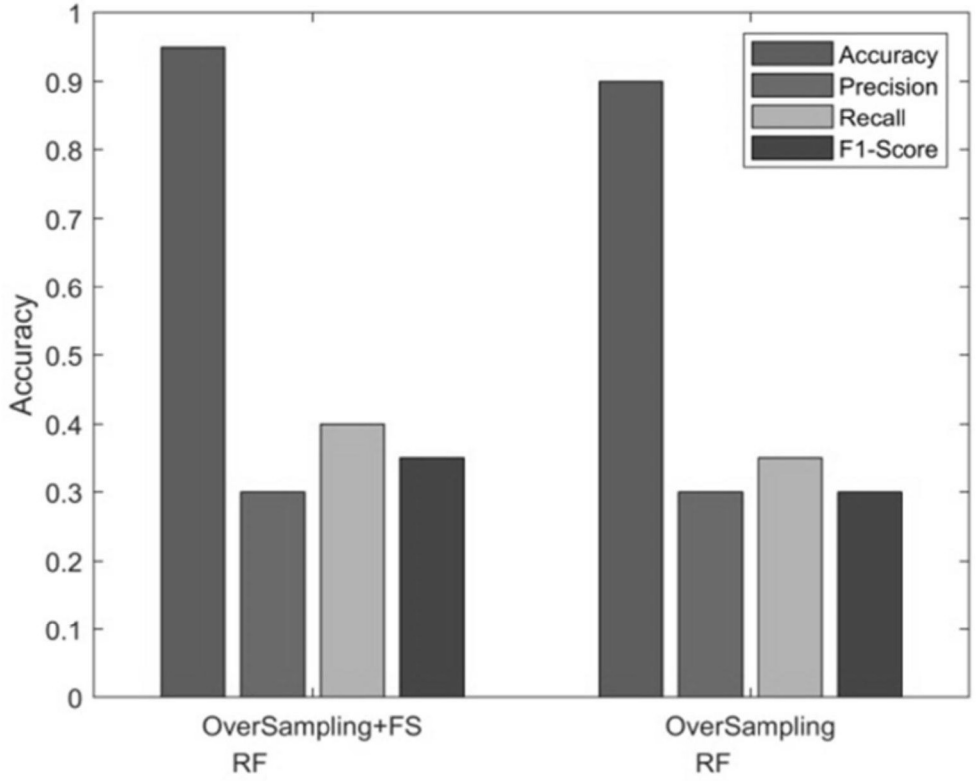

Accuracy
Precision
Recall
F1-Score
Accuracy
OverSampling+FS
RF
OverSampling
RF

图5

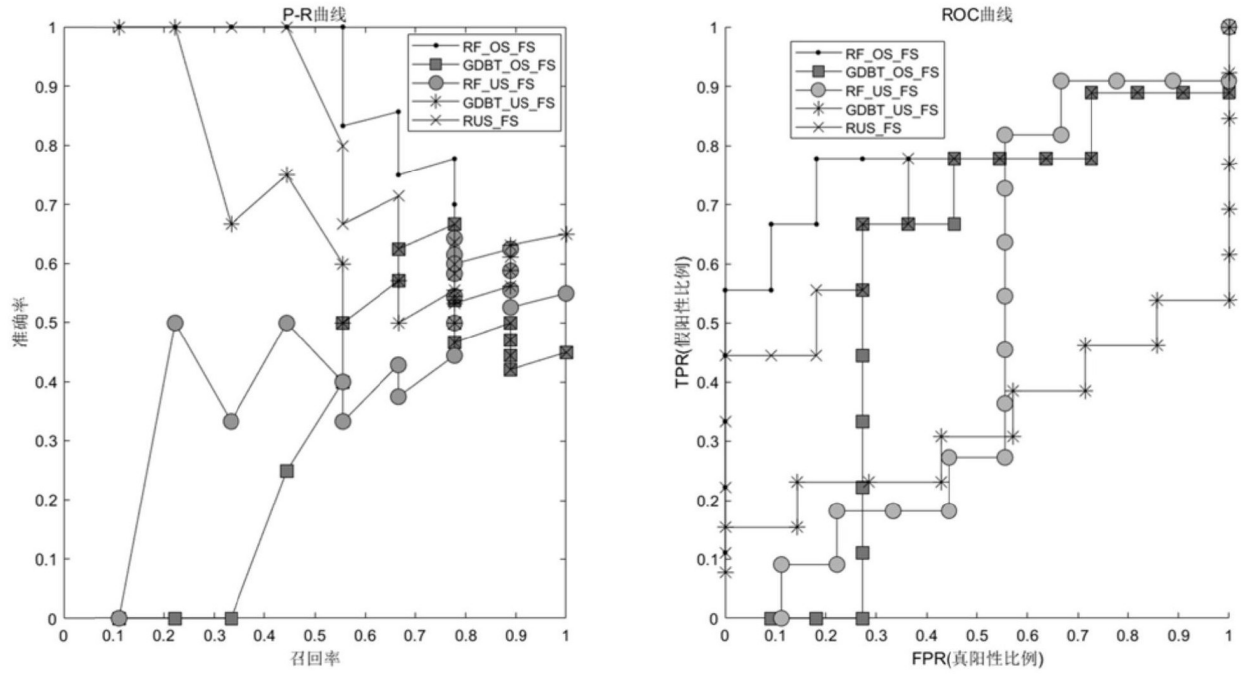

P-R曲线
RF_OS_FS
GDBT_OS_FS
RF_US_FS
GDBT_US_FS
RUS_FS
准确率
召回率
ROC曲线
TPR(假阳性比例)
FPR(真阳性比例)

图6