第54卷第3期 南京航空航天大学学报 Vol. 54 No. 3
2022年6月 Journal of Nanjing University of Aeronautics & Astronautics Jun. 2022

DOI:10.16356/j.1005-2615.2022.03.019

抽油机故障诊断的分布驱动主动学习算法

汪 敏[1],周 磊[1],闵 帆[2],张 响[3],沈佳园[3],韩 菲[4]

(1.西南石油大学电气信息学院,成都610500;2.西南石油大学计算机科学学院,成都610500;
3.浙江浙能天然气运行有限公司,杭州310052;4.新疆油田公司风城油田,克拉玛依834000)

摘要: 抽油机示功图直观显示了抽油机工作情况,但实际工况情况呈现典型的长尾分布特性,类别严重不平衡。传统方法无法准确识别小类别工况,也无法获得井下工作状态准确识别。针对这一问题,提出一种基于分布驱动的多类别长尾数据代价敏感主动学习算法(Cost-sensitive active learning algorithm based on distribution -driven multi-class long-tailed data, CALA)。首先,考虑数据分布特性,以最小化代价为优化目标确定数据的最佳聚类簇数;其次,通过加入预分类误差代价来更新之前得到的最佳聚类簇数;然后,构建集成分类模型作为分类器;最后,通过迭代来平衡数据分布。采用某油田真实的示功图数据进行测试,显著性实验分析证明CALA在小类别工况诊断上具有更好的性能。

关键词: 示功图诊断;代价敏感;主动学习;长尾分布;小类别工况识别

中图分类号: TP181 **文献标志码:** A **文章编号:** 1005-2615(2022)03-0517-11

Distributed Drive Active Learning Algorithm for Fault Diagnosis of Pumping Unit

WANG Min[1], *ZHOU Lei*[1], *MIN Fan*[2], *ZHANG Xiang*[3], *SHEN Jiayuan*[3], *HAN Fei*[4]

(1. School of Electrical Information, Southwest Petroleum University, Chengdu 610500, China; 2. School of Computer Science, Southwest Petroleum University, Chengdu 610500, China; 3. Zhejiang Zheneng Natural Gas Operation Co., Ltd., Hangzhou 310052, China; 4. Fengcheng Factory, Xinjiang Oil Field, Karamay 834000, China)

Abstract: The indicator diagram of the pumping unit visually shows the working conditions of the pumping unit. However, the actual working conditions show typical long-tailed distribution characteristics, and the categories are seriously unbalanced. Traditional methods cannot accurately identify small categories of working conditions, and cannot obtain accurate identification of underground working conditions. Aiming at this problem, a cost-sensitive active learning algorithm based on distribution-driven multi-class long-tail data (CALA) is proposed. First, considering the characteristics of data distribution, the optimal number of clusters for the data is determined by minimizing the cost as the optimization objective. Second, the optimal number of clusters obtained before is updated by adding the pre-classification error cost. Then, a classifier is constructed by integrating the classification models. Finally, balance the data distribution iteratively. Using the real indicator diagram data of an oil field to test, the significant experimental analysis proves that CALA has better performance in the diagnosis of small categories of working conditions.

Key words: indicator diagram diagnosis; cost-sensitive; active learning; long-tail distribution; small category condition identification

基金项目: 国家自然科学基金(62006200);四川省科技计划支持项目(2020YFQ0038,22ZDYF2733)。

收稿日期: 2021-08-01;**修订日期:** 2022-01-06

通信作者: 闵帆,男,教授,博士生导师,E-mail:minfan@swpu.edu.cn。

引用格式: 汪敏,周磊,闵帆,等.抽油机故障诊断的分布驱动主动学习算法[J].南京航空航天大学学报,2022,54(3):517-527. WANG Min, ZHOU Lei, MIN Fan, et al. Distributed drive active learning algorithm for fault diagnosis of pumping unit[J]. Journal of Nanjing University of Aeronautics & Astronautics,2022, 54(3):517-527.

抽油机井一直都是石油开采中的重要组成部分,为了更好地了解抽油机井的工作状况,就必须对其工作时产生的一系列数据进行分析,从而判断抽油机井是否正常工作。通过测量抽油机往复一周所产生的载荷、位移系列数据来绘制地面示功图[1],由不同因素导致的抽油机故障会形成不同形状的示功图。及时准确地对示功图进行诊断,得出抽油机的故障原因,可以减少财产损失和延长零部件的使用寿命。目前以示功图为研究对象对抽油机进行故障诊断是最常见的方法。常见的有BP神经网络[2]、主成分分析方法[3]以及支持向量机(Support vector machine, SVM)[4]等。田增国等[5]提出了一种基于主成分分析的示功图故障诊断系统。该方法是利用降维技术保留大量信息的情况下将原始数据进行压缩,将大量的线性相关属性变量转化成几个相互独立或者不相关的变量。通过计算示功图经过主成分分析后的数据之间的相关系数来判定不同故障。施海青等[6]提出了一种基于支持向量机的抽油机故障诊断方法。该方法采用矢量曲线对数据进行压缩,从而提取井下示功图特征点。采用"一对一"的方式构建多分类支持向量机分类器,能够对多个故障做出识别。杜娟等[7]提出了一种基于卷积神经网络的抽油机工况识别方法。该方法在原有神经网络基础上增添了两个注意力机制模块,能够很好地调节原有模型的过拟合情况,使模型更能关注小类别工况。在工况复杂的抽油机故障诊断实验中,该模型具有良好的泛化能力。文献[8]提出了一种基于稀疏多图正则化极限学习机的抽油机故障诊断方法。该方法通过快速离散曲波变换提取示功图特征,利用图表示学习方法构建类内图和类间图来表示同类数据间的关系以及不同类别数据间的关系。通过稀疏表示,可以使同一类数据的结果输出尽可能相同,不同类别的数据的结果输出尽可能分开。示功图故障诊断测试表明,该模型在抽油机工况识别上有很好的表现。文献[9]采用了适应噪声因子的滤波器以及使用基函数来与之结合的方法。使用近似多边形的傅里叶描述符方法来提取示功图特征,采用径向基函数(Radial basis function,RBF)神经网络,利用指标图数据和生产数据建立故障诊断模型,使用自适应噪声因子来解决模型中的自适应滤波问题。实验表明,模型在示功图故障诊断方面取得不错的表现。

现阶段常用深度学习方法进行故障诊断测试,Peng等[10]开发了一种新型双向门控循环单元(Bidirectional gated recurrent unit,BGRU),在训练阶段对每个训练样本进行加权,以减少类不平衡的影响,然后利用成本敏感的主动学习来选择候选样本。在实际等离子体蚀刻工艺数据集上评估了所提出方法的有效性。Jin等[11]提出一种用于复合故障诊断的新型解耦注意力残差网络,应用在轴承数据集,获得了优越的精度,大大减少了领域专家的标记工作量。Zhang等[12]引入概率主动支持向量机(Probabilistic active support vector machine,Pro-ASVM)的学习方法,根据样本点的概率选择点作为支持向量。应用于轴承振动信号的分类,获得了优异的分类效果。Jian等[13]针对实际工业故障诊断训练集规模较小的问题,提出了一种基于主动和半监督学习的故障诊断新方法。应用于实际的智能维护系统数据,为小训练集下的故障诊断提供了一种有前途且有用的方法。Chen等[14]针对自组织蜂窝网络(Self-organizing cellular networks, SONs)中的故障诊断的多分类问题,提出了一种新的基于主动学习的故障诊断方案。该方案只需很少的标记训练实例即可实现高诊断性能,从而显著降低成本。Punčochář等[15]提出了主动故障诊断(Active fault diagnosis,AFD)领域的基本分类方法。由于实际油田生产过程中存在抽油机井下的故障种类数量多且不同故障类别的数据量不平衡、人为标注的样本少且费时费力等问题,常用的深度学习工况识别模型难以在实际工作中落地。同时,主成分分析方法、支持向量机等传统的方法无法很好的处理不平衡数据分类问题。针对以上方法存在的不足,本文提出一种基于分布驱动的多类别长尾数据代价敏感主动学习算法(Cost-sensitive active learning algorithm based on distribution-driven multi-class long-tailed data, CALA)来解决这一困难且非常有意义的问题。

1 特征提取

本节主要介绍本文示功图的特征提取方法,结合灰度矩阵的知识,提取示功图灰度矩阵的6个特征作为统计特征。

1.1 网格法提取灰度矩阵

本文采用网格法[16]对示功图进行灰度矩阵提取,网格法构建示功图的灰度矩阵主要包含如下步骤:

(1)标准化示功图

为了更好地比较不同工况下的抽油机示功图，消除示功图量纲对收集到的数据的影响，将采集到的示功图数据进行标准归一化。为符合石油工业的习惯，将示功图放进一个长宽比为2∶1的矩形中，满足绘制的地面示功图被矩形内切这一条件。

(2)网格化示功图

将长方形分成多个网格，本文将之划分为20×10大小的网格个数，并将所有网格的初始灰度赋值“0”；若网格内含有示功图曲线，其灰度值赋值为“1”；边界内部网格的灰度值往矩形中心依次递增；边界外部网格的灰度值以矩形边界依次递减。边界搜索方式按列进行。

1.2　特征向量提取

通过对构建好的示功图灰度矩阵[17]进行数理统计，计算灰度均值$\bar{g}$、方差σ^2、偏度ε、峰度P、能量E和熵ξ这6个统计特征作为示功图特征值。

假设灰度矩阵大小为$\boldsymbol{G}(A,B)$，矩阵中任意位置的值$g_{ab}(1\leqslant a\leqslant A,1\leqslant b\leqslant B)$表示示功图网格化后对应位置的灰度。设灰度矩阵中灰度级数为R，设某一灰度级数r的数量为$T(r)$，则该灰度级数在灰度矩阵中出现的概率可表示为$p(r)=T(r)/(A\times B)$。

$$d_1=\bar{g}=\sum_{r=1}^{R}r\cdot p(r) \tag{1}$$

$$d_2=\sigma^2=\sum_{r=1}^{R}\left(r-\bar{g}\right)^2\cdot p(r) \tag{2}$$

$$d_3=\varepsilon=\frac{1}{\sigma^3}\sum_{r=1}^{R}\left(r-\bar{g}\right)^3\cdot p(r) \tag{3}$$

$$d_4=P=\frac{1}{\sigma^4}\sum_{r=1}^{R}\left(r-\bar{g}\right)^4\cdot p(r) \tag{4}$$

$$d_5=E=\sum_{r=1}^{R}\left[p(r)\right]^2 \tag{5}$$

$$d_6=\xi=-\sum_{r=1}^{R}\left[1-p(r)\right]\cdot\log_{10}\left[1-p(r)\right] \tag{6}$$

以统计的6个特征值$\{d_1,d_2,d_3,d_4,d_5,d_6\}$作为最终的分类特征向量。

2　算法设计

本文的数据模型是教师和误分类代价决策系统(TMC-DS)[18]，该决策系统定义成1个四元组

$$\boldsymbol{S}=(\boldsymbol{X},\boldsymbol{y},\boldsymbol{M},t) \tag{7}$$

式中：$\boldsymbol{X}$代表一个数据集向量；$\boldsymbol{y}$代表数据真实标签向量；$\boldsymbol{M}$代表误分类代价矩阵；t代表专家代价为1。CALA算法过程框图如图1所示。

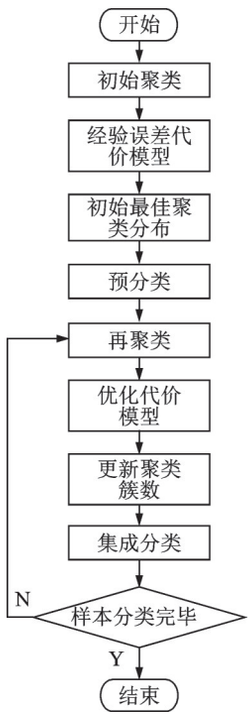

图1　CALA算法流程框图

Fig.1　CALA algorithm flow chart

2.1　获取数据最佳分布

本节设计了一种基于误差统计函数探索数据最佳聚类簇数的方法。依据“物以类聚”的原则，样本间距离越接近，它们的标签就越可能一致的假设[19]。通过对多个结构化数据集进行分析测验，得到拟合误差曲线。其具体步骤如下：

(1) 距离阈值实例对

依据距离阈值λ的相邻实例对$(\boldsymbol{x}_i,\boldsymbol{x}_j)$定义为

$$N_\lambda=\left\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{X}\times\boldsymbol{X}\,\middle|\,\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)\leqslant\lambda\right\} \tag{8}$$

式中：$\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)$代表数据样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$间的欧式距离；$\lambda$为设定归一化距离阈值；$N_\lambda$为满足条件的实例对个数。

(2) 实例对标签统计误差

根据式(8)得到的实例对个数，依据不同的距离阈值定义实例对标签统计误差函数

$$e(\lambda)=\frac{\left|\left\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in N_\lambda\,\middle|\,y_i\neq y_j\right\}\right|}{\left|N_\lambda\right|} \tag{9}$$

式中：$|N_\lambda|$为满足阈值λ下实例对数量；y_i和y_j为样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$对应的真实标签。

(3) 获取经验误差函数

首先选取30个不同样本个数，不同特征个数以及不同类别数量的公开数据集，其次通过式(8)计算不同阈值λ下的实例对个数，然后通过式(9)统计不同阈值λ下的标签统计误差$e(\lambda)$，最后通过多项式拟合得到经验误差函数，即

$$\varphi(\lambda)=-0.016\,41\lambda^3-0.123\,1\lambda^2+0.332\,2\lambda-0.009\,893 \tag{10}$$

拟合曲线相关系数达到0.999 9,符合工程实际。

(4) 优化目标函数

依据得到的经验误差函数,将数据聚类为2到$\sqrt{n}$个簇,根据下面的目标函数得到对应簇数的代价。以最小化代价为优化目标得到对应的聚类簇数k,有

$$f_1(k)=\sum_{i=1}^{k}n_i\varphi(\lambda_i)+kt \tag{11}$$

式中:n为数据样本总数,n_i为对应第i簇的样本个数,λ_i为第i簇的最远两样本距离与数据集最远两样本距离的比值。

2.2 预分类

利用预分类修正基于统计策略得到的最佳簇数。将统计策略得到的最佳聚类簇数中每一簇通过主动学习方法[20]选择最具代表性的样本作为训练集,通过概率预测模型得到样本预分类标签。训练集的选取方式为

$$\boldsymbol{s}^*=\arg\min_{\boldsymbol{x}\in\boldsymbol{C}_i}\left\|\boldsymbol{x}-\boldsymbol{c}_i\right\|_2 \tag{12}$$

式中:$\boldsymbol{c}_i$为第$\boldsymbol{C}_i$簇的聚类中心;$\boldsymbol{s}^*$为该簇交由专家标注的样本。

通过Softmax回归[21],输入任意样本$\boldsymbol{x}_i$,属于样本对应的预测概率为

$$P\left(y'=j_v\middle|\boldsymbol{x}_i;\boldsymbol{\theta}\right)=\frac{e^{\boldsymbol{\theta}_{j_v}^{\mathrm{T}}\boldsymbol{x}_i}}{\sum_{d=1}^{l}e^{\boldsymbol{\theta}_d^{\mathrm{T}}\boldsymbol{x}_i}} \tag{13}$$

其预测标签为

$$j^*=\arg\max_{j_v\in[1,2,l]}P\left(y'=j_v\middle|\boldsymbol{x}_i;\boldsymbol{\theta}\right) \tag{14}$$

式中:l为样本类别数量;$\boldsymbol{\theta}$为Softmax目标函数训练得到的最佳参数。通常通过梯度下降法[22]求解。

2.3 更新最佳聚类分布

通过Softmax回归模型进行预分类,测试样本会得到一个相应的预测标签。将数据再次进行聚类,依照得到的样本预测标签和经验误差曲线构建新的聚类优化目标函数,有

$$f_2(k)=\sum_{i=1}^{k}n_i\left(\partial_1\varphi(\lambda_i)+\partial_2\left(1-p_u(\boldsymbol{C}_i)\right)\right)+kt \tag{15}$$

式中:∂_1和∂_2为权重系数;$p_u(\boldsymbol{C}_i)$为第$\boldsymbol{C}_i$簇的预测标签纯度,定义如下

$$p_u(\boldsymbol{C}_i)=\frac{\max\limits_{q\in[1,2,\cdots,l]}\left|\boldsymbol{x}_i\in\boldsymbol{C}_i\middle|j^*=q\right|}{\left|\boldsymbol{C}_i\right|} \tag{16}$$

式中:$\left|\boldsymbol{C}_i\right|$为第$\boldsymbol{C}_i$簇样本总数;$\left|\boldsymbol{x}_i\in\boldsymbol{C}_i\middle|j^*=q\right|$为第$\boldsymbol{C}_i$簇中预测标签一致的样本个数。

2.4 集成分类

根据找到的最佳聚类簇数,将数据进行聚类,选取每一簇离中心点最近的样本作为训练集,通过Softmax回归得到测试集的预测标签。并且将该训练集同时作为K最近邻算法(K-nearest neighbor,KNN)预测分类模型的训练集,得到测试集的KNN预测标签集合j'。结合二者的预测标签构建决策函数

$$L(\boldsymbol{x}_i)=\begin{cases}\text{分类} & (p_u(\boldsymbol{C}_i=1))\wedge(j_i^*=j_i')\\ \text{继续迭代} & \text{其他}\end{cases} \tag{17}$$

式中:$\boldsymbol{x}_i\in\boldsymbol{C}_i$;$j_i^*$和$j_i'$分别为Softmax和KNN对样本$\boldsymbol{x}_i$的预测标签。

2.5 伪代码及时间复杂度分析

(1)算法伪代码

算法 CALA

输入 决策信息系统$\boldsymbol{S}=(\boldsymbol{X},\boldsymbol{y},\boldsymbol{M},t)$

输出 预测标签集合$\boldsymbol{Y}=\left[y\right]_{n\times1}$

(1)$\boldsymbol{X}_r\leftarrow\varnothing$;$\boldsymbol{Y}\leftarrow\varnothing$;$\boldsymbol{X}_u\leftarrow\varnothing$//初始化

(2)$\boldsymbol{X}_u\leftarrow\boldsymbol{X}$//赋初值

(3)$k\leftarrow k$-means($\boldsymbol{X}$)//获取初始聚类信息

(4)$k_1\leftarrow\arg\min\limits_{2\leqslant k\leqslant\left|X_u\right|}f_1(k)$//得到初始最佳聚类簇数

(5)$\boldsymbol{s}^*\leftarrow$select($\boldsymbol{X},\alpha n$)//得到训练集

(6)$\boldsymbol{X}_r\leftarrow\boldsymbol{s}^*$;$\boldsymbol{X}_u=\boldsymbol{X}-\boldsymbol{X}_r$//得到测试集

(7)$[\boldsymbol{\theta}]_{l\times(m+1)}=\text{Soft max Train}(\boldsymbol{X}_r)$//训练$\boldsymbol{\theta}$模型

(8)$j^*\leftarrow\text{Soft max Classify}(\boldsymbol{X}_u)$/预分类标签$j^*$

(9)while$(\boldsymbol{X}_u\neq\varnothing)$do

(10)$k\leftarrow k$-means$(\boldsymbol{X}_u)$获取聚类信息

(11)$k_2\leftarrow\arg\min\limits_{2\leqslant k\leqslant\left|X_u\right|}f_2(k)$//更新后聚类簇数$k_2$

(12)$\boldsymbol{s}^*\leftarrow$select($\boldsymbol{X}_u,\alpha n$)

(13)$\boldsymbol{X}_r\leftarrow\boldsymbol{s}^*\cup\boldsymbol{X}_r$;$\boldsymbol{X}_u=\boldsymbol{X}_u-\boldsymbol{X}_r$//更新训练集

(14)for $i=1,2,\cdots,\left|\boldsymbol{X}_u\right|$

(15)$j^*\leftarrow\text{Soft max Classify}(\boldsymbol{X}_u)$

(16)$j'\leftarrow\text{KNNClassify}(\boldsymbol{X}_u)$

(17)end for

(18)if $(p_u(\boldsymbol{C}_i)=1)\wedge((j_i^*=j_i'))$//集成分类

(19)$\boldsymbol{Y}\leftarrow\boldsymbol{Y}\cup j^*(\boldsymbol{C}_i)$//最终分类

(20)end if

(21)$\boldsymbol{X}_u\leftarrow\boldsymbol{X}_u-\boldsymbol{C}_i$//下一轮等待样本

(22)if$(\boldsymbol{X}_u==\varnothing)$,then

(23)break;

(24)end if

(25) end while

(26) return $\boldsymbol{Y}=\left[y\right]_{n\times 1}$

(2) 时间复杂度分析

步骤 1～5 为赋值和通过聚类得到数据初始分布信息阶段，计算量主要在于聚类算法，时间复杂度为 $O(kdn)$。步骤 6～8 为选取训练样本和 Softmax 预分类过程，选取训练样本阶段时间复杂度为 $O(n^2)$，Softmax 预分类过程时间复杂度为 $O(n'^2)$，n' 为预分类样本数量，为原始样本总数减去训练样本后的样本个数。$n'<n$，这阶段总的时间复杂度为 $O(n^2)+O(n'^2)=O(n^2)$。步骤 9～25 为更新最佳聚类分布和集成分类过程，更新最佳聚类分布与初始聚类阶段时间复杂度一致为 $O(kdn)$，集成分类过程中，Softmax 分类阶段时间复杂度为 $O(n^2)$，KNN 分类阶段时间复杂度为 $O(n)$，考虑 while 循环过程，则这阶段总的时间复杂度为 $O(kdn\cdot\log_2 n)+O(n^2 log_2 n)+O(n\log_2)=O(n^2\log_2 n)$。其中特征数 $d<n$，聚类簇数 $k<n$，时间复杂度为 $O(kdn)+O(n^2)+O\left(n^2\log_2 n\right)=O\left(n^2\log_2 n\right)$。

3　算法验证

3.1　数据集描述

实验采用来自新疆风城油田 4 个作业区不同抽油机示功图数据对本文算法进行验证分析。其具体信息如表 1 所示。这些数据包含多个类别且都是不平衡数据。其中 A01 是抽油机作业一区常规油井采集的示功图数据，A02 是抽油机作业二区稠油油井采集的示功图数据，A03 是抽油机作业三区超稠油油井采集的示功图数据，A04 是抽油机作业四区 SAGD 油井采集的示功图数据。4 个油田示功图数据包含有正常工作、供液不足、气体影响、气锁、上碰泵、下碰泵、游动阀关闭迟缓、柱塞脱出泵工作筒、游动阀漏、固定阀漏、砂影响＋供液不足和惯性影响这 12 种常见抽油机工况。其中，大部分为正常工作，气体影响工况为最小类别故障工况。A01 中正常工况样本有 4 474 个，气体影响工况有 300 个，不平衡比例为 14.91；A02 中正常工况样本有 4 974 个，气体影响工况有 300 个，不平衡比例为 16.58；A03 中正常工况样本有 5 374 个，气体影响工况有 300 个，不平衡比例为 17.91；A04 中正常工况样本有 5 845 个，气体影响工况样本有 300 个，不平衡比例为 19.48。实际油田工作环境下，抽油机示功图中气体影响这一类工况数据稀少。当发生气体影响时，抽油机泵腔内压力不能正常下降，使得加载速度变慢，采油效率降低。对小类别工况进行准确识别能够及时对故障机械进行维修，减少损失、延长机器设备的使用寿命。

表 1　数据集信息

Table 1　Information of datasets

ID	数据集	样本数量	属性数	类别数	来源
1	A01	10 532	6	12	油田
2	A02	11 032	6	12	油田
3	A03	11 532	6	12	油田
4	A04	12 587	6	12	油田

3.2　评价指标

本文实验采用精度、平均代价 F-Measure 作为评估算法性能的指标，其精度定义为

$$\text{Accuracy}=\frac{\left|\boldsymbol{X}_{\text{t}}\right|-\text{error}}{\left|\boldsymbol{X}_{\text{t}}\right|}\times 100\% \tag{18}$$

式中：$\left|\boldsymbol{X}_{\text{t}}\right|$ 为测试集的样本数量，error 为误分类样本数量。

对于不平衡抽油机故障工况数据而言，刻画不同工况具有不同的误分类代价是很有必要的。对于稀少工况类别数据在实际场景下样本数稀少，误分类的代价应远大于常见工况类别数据误分类代价。本文设定的代价矩阵[23]为

$$\boldsymbol{m}_{ij}=\frac{n_j}{n_i} \tag{19}$$

式中：n_i 和 n_j 分别表示测试集中属于第 i 类和第 j 类的样本数量。平均代价为

$$\text{Cost}=\frac{\sum_{i=1}^{l}\sum_{j=1}^{l}\boldsymbol{A}_{ij}\boldsymbol{m}_{ij}+t\left|\boldsymbol{X}_r\right|}{n} \tag{20}$$

式中：$\boldsymbol{A}_{ij}$ 为将第 i 类误分类为第 j 类的样本数量；$\left|\boldsymbol{X}_r\right|$ 为交由专家标注的样本个数；t 为查询标签代价，实验中设置为 1。

为验证模型在不平衡数据分类上的性能，从准确率(Precision)和召回率(Recall)和 F-measure 分数[24]这 3 个评价指标对模型性能进行综合评判。这 3 种评价指标可以由表 2 的混淆矩阵计算得出。

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}\times 100\% \tag{21}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}\times 100\% \tag{22}$$

$$\text{F-measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\times 100\% \tag{23}$$

式中：TP 和 TN 分别表示真实标签与预测标签全部为正，全部为负的样本数量；FP 表示真实标签为负，预测标签为正的样本数量，而 FN 相反。准确率是针对模型测试结果，表示预测为正实例中有多少真正的正实例；召回率是针对于原始样本具体标签，表示原始样本的正实例有多少被模型预测正确。F-measure 综合兼顾这两个评判标注，是评价算法性能最常用的指标。

表2 混淆矩阵

Table 2 Confusion matrix

真实标签	预测标签	
	正实例	负实例
正实例	TP	FN
负实例	FP	TN

3.3 实验设计

为验证提出的算法模型性能的优越性，将本文提出的CALA算法与基于欠采样技术的代价敏感学习算法(Under-sampling, US)[25]、基于阈值移动调整类别阈值算法(Threshold-moving, TM)[26]、基于过采样技术的代价敏感学习算法(Over-sampling, OS)[27]、增强的自动双支持向量机算法(Enhanced automatic twin support vector machine, EATWSVM)[28]、基于边距的非定性采样主动学习算法(Uncertainty sampling with margin, UM)[29]、基于熵的不确定性采样主动学习算法(Uncertainty sampling with entropy, UE)[30]和基于成本嵌入的主动学习算法(Active learning with cost embedding, ALCE)[31]以及卷积神经网络(Convolutional neural network, CNN)这8种算法进行比较。US、TM、OS和EATWSVM是4种代价敏感不平衡数据处理方法，UM、UE和ALCE是3种代价敏感主动学习算法。

4 实验结果及分析

4.1 与代价敏感不平衡数据处理方法比较

本节实验中，将真实采集到的4个油田的抽油机示功图数据用于模型性能验证。每个数据集选取30%的样本交由专家标注标签进行模型训练，其余样本作为测试集。同样条件下，随机10次重复实验，统计各评价指标结果。结果取均值和标准差如表3所示。

表3 与代价敏感不平衡数据处理方法对比实验结果(均值±方差)

Table 3 Comparison of experimental results with cost-sensitive imbalanced data processing methods(mean±std)

数据集	评价指标	算法				
		US	TM	OS	EATWSVM	CALA
A01	Accuracy	0.79±0.05	0.81±0.01	0.64±0.04	0.56±0.02	**0.83±0.01**
	Cost	0.97±0.02	0.94±0.05	**0.67±0.02**	1.53±0.12	0.90±0.02
	Precision	**0.77±0.04**	0.76±0.03	0.61±0.03	0.70±0.02	0.70±0.01
	Recall	0.68±0.03	0.69±0.01	0.78±0.09	0.52±0.09	**0.88±0.03**
	F-measure	0.68±0.01	0.68±0.02	0.66±0.07	0.57±0.07	**0.73±0.02**
A02	Accuracy	0.80±0.03	0.80±0.02	0.64±0.02	0.67±0.04	**0.83±0.03**
	Cost	0.89±0.05	0.89±0.09	**0.58±0.05**	1.26±0.08	0.90±0.02
	Precision	**0.78±0.04**	0.75±0.03	0.62±0.03	0.53±0.11	0.67±0.04
	Recall	0.69±0.07	0.65±0.02	0.79±0.03	0.70±0.09	**0.87±0.01**
	F-measure	0.68±0.06	0.69±0.05	0.62±0.01	0.50±0.03	**0.72±0.05**
A03	Accuracy	0.78±0.05	0.77±0.02	0.62±0.08	0.51±0.04	**0.80±0.03**
	Cost	0.92±0.07	0.94±0.07	**0.63±0.04**	1.52±0.07	0.97±0.01
	Precision	**0.76±0.02**	0.74±0.02	0.62±0.02	0.53±0.06	0.68±0.05
	Recall	0.67±0.03	0.65±0.02	0.79±0.05	0.72±0.09	**0.87±0.01**
	F-measure	0.66±0.02	0.65±0.03	0.63±0.01	0.55±0.01	**0.72±0.04**
A04	Accuracy	0.67±0.04	0.70±0.02	0.56±0.07	0.45±0.15	**0.81±0.05**
	Cost	0.94±0.05	0.87±0.09	**0.60±0.09**	1.61±0.08	0.85±0.02
	Precision	0.76±0.02	**0.77±0.04**	0.60±0.05	0.71±0.07	0.71±0.04
	Recall	0.66±0.08	0.67±0.07	0.79±0.02	0.40±0.02	**0.88±0.03**
	F-measure	0.66±0.01	0.70±0.04	0.63±0.03	0.46±0.06	**0.75±0.03**

从表3可以看出，在A01、A02、A03和A04数据集中，本文所提出的CALA算法在精度、召回率和F-measure这3种评价指标上展现的性能都优于其余4种对比算法。在代价性能测试上，过采样算法OS表现最好，CALA在4个数据集上的代价排名分别为第二、第四、第四和第二。

为验证本文提出的CALA算法在不同查询比率下的性能，图2显示了CALA与4种代价敏感不平衡数据处理方法在查询比率为30%、35%、40%、45%和50%下的F-measure对比，对于4个真实油井数据集，CALA算法的平均F-measure明显高于其余算法。

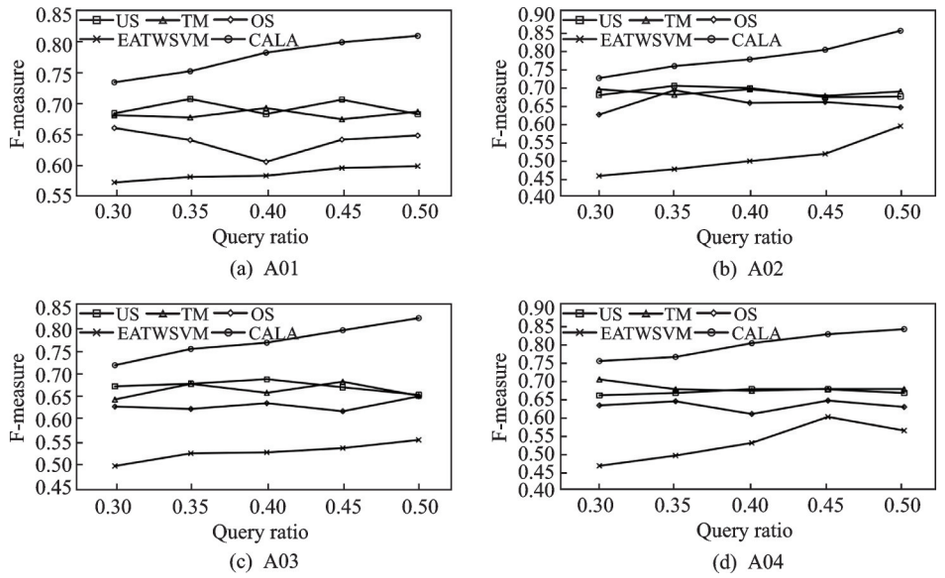

图 2　CALA 算法与 4 种不平衡数据处理算法在不同查询比率下的 F-measure 比较

Fig.2　Comparison of F-measure between CALA algorithm and four imbalanced data processing algorithms under different query ratios

4.2　与代价敏感主动学习算法比较

本节实验中,将真实采集到的 4 个油田的抽油机示功图数据用于模型性能验证。每个数据集选取 30% 的样本交由专家标注标签进行模型训练,其余样本作为测试集。同样条件下,随机 10 次重复实验,统计各评价指标结果。结果取均值和标准差如表 4 所示。

表 4　与代价敏感主动学习算法对比实验结果(均值±方差)

Table 4　Comparison of experimental results with cost sensitive active learning algorithms(mean±std)

数据集	评价指标	算法				
		UM	UE	ALCE	CNN	CALA
A01	Accuracy	0.82 ± 0.01	0.81 ± 0.02	0.82 ± 0.05	0.82 ± 0.01	**0.83 ± 0.01**
	Cost	0.97 ± 0.04	0.85 ± 0.04	0.97 ± 0.09	**0.78 ± 0.07**	0.90 ± 0.02
	Precision	0.78 ± 0.02	0.79 ± 0.05	**0.79 ± 0.03**	0.68 ± 0.02	0.70 ± 0.01
	Recall	0.66 ± 0.01	0.67 ± 0.07	0.67 ± 0.02	0.74 ± 0.05	**0.88 ± 0.03**
	F-measure	0.69 ± 0.03	0.70 ± 0.09	0.69 ± 0.06	0.71 ± 0.01	**0.73 ± 0.02**
A02	Accuracy	0.80 ± 0.02	0.81 ± 0.04	0.80 ± 0.02	0.82 ± 0.04	**0.83 ± 0.03**
	Cost	0.95 ± 0.06	0.94 ± 0.09	0.95 ± 0.04	**0.75 ± 0.09**	0.90 ± 0.02
	Precision	0.77 ± 0.01	0.76 ± 0.10	**0.78 ± 0.02**	0.67 ± 0.06	0.67 ± 0.04
	Recall	0.65 ± 0.04	0.65 ± 0.06	0.66 ± 0.03	0.76 ± 0.02	**0.87 ± 0.01**
	F-measure	0.67 ± 0.03	0.68 ± 0.02	0.67 ± 0.01	0.69 ± 0.05	**0.72 ± 0.05**
A03	Accuracy	**0.83 ± 0.03**	0.82 ± 0.05	0.83 ± 0.07	0.83 ± 0.08	0.80 ± 0.03
	Cost	0.90 ± 0.10	0.90 ± 0.03	0.90 ± 0.02	**0.75 ± 0.02**	0.97 ± 0.01
	Precision	0.77 ± 0.02	**0.78 ± 0.02**	0.77 ± 0.05	0.72 ± 0.07	0.68 ± 0.05
	Recall	0.68 ± 0.03	0.68 ± 0.04	0.67 ± 0.06	0.76 ± 0.05	**0.87 ± 0.01**
	F-measure	0.70 ± 0.05	0.70 ± 0.01	0.71 ± 0.08	0.72 ± 0.05	**0.72 ± 0.04**
A04	Accuracy	0.74 ± 0.02	0.74 ± 0.09	0.74 ± 0.07	0.80 ± 0.09	**0.81 ± 0.05**
	Cost	0.96 ± 0.07	0.96 ± 0.04	0.96 ± 0.05	**0.64 ± 0.01**	0.85 ± 0.02
	Precision	0.77 ± 0.03	0.77 ± 0.06	**0.77 ± 0.02**	0.73 ± 0.04	0.71 ± 0.04
	Recall	0.64 ± 0.02	0.64 ± 0.05	0.64 ± 0.01	0.77 ± 0.07	**0.88 ± 0.03**
	F-measure	0.64 ± 0.04	0.64 ± 0.02	0.64 ± 0.07	0.73 ± 0.01	**0.75 ± 0.03**

从表 4 可以看出,在 A01、A02 和 A04 数据集中,本文所提出的 CALA 算法在精度、召回率和 F-measure 这 3 种评价指标上展现的性能都优于其余 4 种对比算法。A03 数据集上,提出的 CALA 算法在召回率和 F-measure 评价上优于其余对比算法。在代价性能测试上,深度学习算法 CNN 表现

最好,CALA在4个数据集上的代价排名分别为第三、第二、第五和第二。

为验证算法在不同查询比率下的性能,图3分别显示了与3种代价敏感主动学习算法以及深度学习算法在查询比率为30%、35%、40%、45%和50%下的F-measure对比,对于4个真实油井数据集,CALA算法的平均F-measure明显高于其余算法。

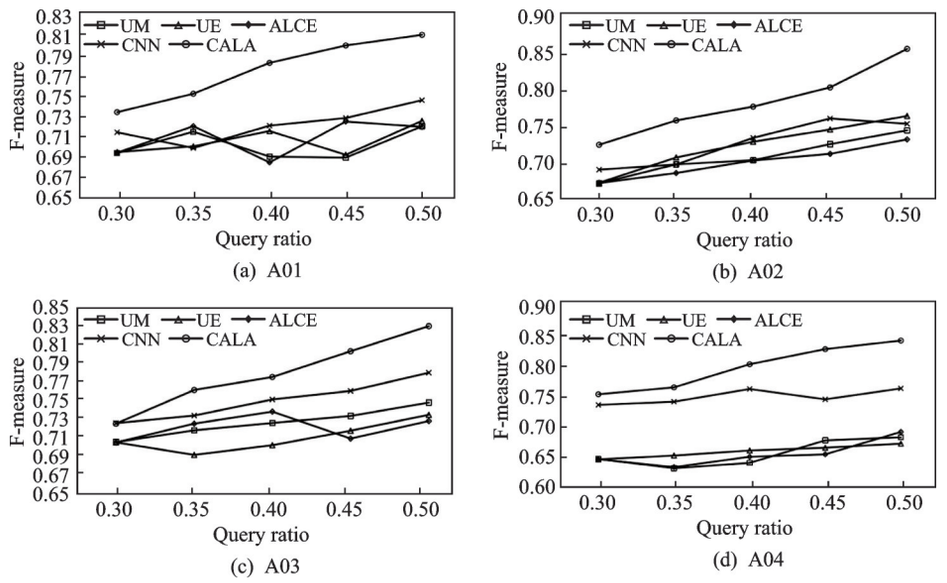

图3 CALA算法与代价敏感主动学习算法以及CNN算法在不同查询比率下的F-measure比较

Fig.3 Comparison of F-measure between CALA algorithm and cost-sensitive active learning algorithm and CNN algoithm under different query ratios

4.3 小类别工况下的模型性能测试

为验证本文算法在小类别上的识别性能,气体影响工况为最小类别工况。其中A01、A02、A03和A04数据集中气体影响工况占比分别为2.85%、2.72%、2.60%和2.38%。表5和表6分别列出CALA算法和8种对比算法在气体影响工况上的性能。表5和表6可以得出,CALA算法在小类别识别方面的准确度和F-Measure优于其余对比算法;在召回率方面,US、TM和UM算法表现较好。

表5 小类别工况下与代价敏感不平衡数据处理方法的对比实验结果(均值±方差)

Table 5 Experimental results compared with cost-sensitive imbalanced data processing methods under small category conditions(mean±std)

数据集	评价指标	算法				
		US	TM	OS	EATWSVM	CALA
A01	Precision	0.82±0.04	0.81±0.03	0.89±0.02	0.74±0.03	**0.99±0.02**
	Recall	**0.97±0.02**	**0.97±0.02**	0.66±0.05	0.54±0.04	0.83±0.07
	F-measure	0.89±0.05	0.89±0.07	0.75±0.06	0.58±0.05	**0.90±0.01**
A02	Precision	0.79±0.09	0.77±0.06	0.93±0.03	0.56±0.03	**0.98±0.01**
	Recall	**0.94±0.07**	0.91±0.02	0.49±0.04	0.75±0.09	0.82±0.05
	F-measure	0.85±0.02	0.84±0.08	0.64±0.05	0.58±0.05	**0.89±0.03**
A03	Precision	0.78±0.02	0.80±0.02	0.87±0.06	0.55±0.01	**0.99±0.04**
	Recall	0.96±0.04	**0.96±0.01**	0.64±0.07	0.72±0.04	0.79±0.08
	F-measure	0.86±0.02	0.87±0.02	0.74±0.04	0.56±0.02	**0.88±0.06**
A04	Precision	0.62±0.09	0.60±0.02	0.60±0.02	0.72±0.02	**0.91±0.02**
	Recall	0.94±0.03	**0.94±0.01**	0.25±0.07	0.44±0.08	0.77±0.05
	F-measure	0.74±0.05	0.73±0.02	0.35±0.09	0.49±0.07	**0.83±0.03**

4.4 模型变换测试

本文算法的核心在于提出的主动查询策略以及基于代价优化目标实现分布优化。因此,本文将KNN算法替换成朴素贝叶斯(Naïve Bayes,NB)算法即CALA_NB。表7为CALA_NB在查询比率为30%下重复10次实验得到的结果。结果表明,将KNN替换成NB之后,算法的效果相差不大,说明本文算法性能适用性能较好。

表 6　小类别工况下与代价敏感主动学习算法的对比实验结果(均值±方差)

Table 6　Experimental results compared with cost-sensitive active learning algorithms under small category conditions (mean±std)

数据集	评价指标	算法				
		UM	UE	ALCE	CNN	CALA
A01	Precision	0.80±0.04	0.82±0.02	0.83±0.05	0.69±0.02	**0.99±0.02**
	Recall	**0.97±0.09**	0.95±0.05	0.96±0.02	0.76±0.01	0.83±0.07
	F-measure	0.72±0.05	0.73±0.03	0.70±0.07	0.73±0.06	**0.90±0.01**
A02	Precision	0.74±0.02	0.76±0.07	0.81±0.01	0.68±0.05	**0.98±0.01**
	Recall	**0.97±0.02**	0.94±0.01	0.93±0.06	0.78±0.04	0.82±0.05
	F-measure	0.88±0.03	0.85±0.02	0.88±0.02	0.72±0.02	**0.89±0.03**
A03	Precision	0.81±0.01	0.82±0.05	0.86±0.05	0.74±0.07	**0.99±0.04**
	Recall	0.92±0.07	0.93±0.08	**0.97±0.08**	0.78±0.03	0.79±0.08
	F-measure	0.87±0.02	0.86±0.02	0.85±0.02	0.76±0.02	**0.88±0.06**
A04	Precision	0.78±0.04	0.79±0.08	0.75±0.04	0.75±0.02	**0.91±0.02**
	Recall	0.92±0.06	0.90±0.07	**0.92±0.01**	0.78±0.03	0.77±0.05
	F-measure	0.77±0.03	0.74±0.04	0.77±0.02	0.76±0.01	**0.83±0.03**

表 7　模型变换测试结果(均值±标准差)

Table 7　Model conversion test results(mean±std)

数据集	算法	评价指标				
		Accuracy	Cost	Precision	Recall	F-measure
A01	CALA	**0.83±0.01**	**0.90±0.02**	**0.70±0.01**	**0.88±0.03**	**0.73±0.02**
	CALA_NB	**0.83±0.01**	**0.90±0.02**	0.70±0.02	0.86±0.02	0.71±0.03
A02	CALA	**0.83±0.03**	**0.90±0.02**	0.67±0.04	0.87±0.01	0.72±0.05
	CALA_NB	0.82±0.04	0.91±0.02	**0.69±0.03**	**0.88±0.02**	**0.73±0.03**
A03	CALA	0.80±0.03	**0.94±0.01**	**0.68±0.05**	0.87±0.01	**0.72±0.04**
	CALA_NB	**0.80±0.02**	0.96±0.02	0.65±0.04	**0.88±0.03**	0.70±0.03
A04	CALA	0.81±0.05	**0.85±0.02**	**0.71±0.04**	**0.88±0.03**	**0.75±0.03**
	CALA_NB	**0.81±0.03**	**0.85±0.02**	0.71±0.05	0.87±0.04	0.73±0.01

4.5　算法适用性分析

为验证算法在12种常见抽油机工况下的不同性能,图4分别显示了CALA在A01、A02、A03以及A04四个数据集用30%查询比例情况下12种工况的精度。其中横坐标1~12分别对应12种抽油机工况。从图中可以看出CALA在各种工况下的识别精度表现都较好。

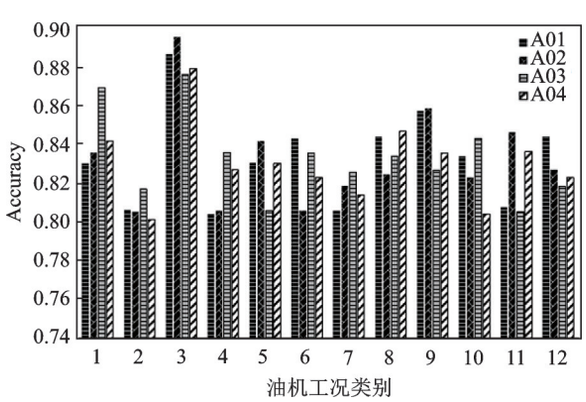

图4　CALA算法在4个油田数据集上的12种工况精度

Fig.4　Accuracy of CALA algorithm for 12 working conditions on four oil field datasets

4.6　模型时间开销对比测试

表8为本文提出算法CALA与其余9种模型在4个实际抽油机数据集上运行的时间开销。本文提出的算法CALA均排名第4,由于使用了集成好的US、TM和OS算法,这3种算法运行速度更快。

表 8　时间开销对比测试结果

Table 8　Time cost comparison test results

数据集	时间/s									
	US	TM	OS	UM	UE	ALCE	CNN	EATWSVM	CALA_NB	CALA
A01	0.84	**0.75**	3.06	44.80	58.03	66.42	21.50	73.52	22.92	20.81
A02	1.16	**0.76**	3.29	50.08	65.60	74.58	26.58	78.20	25.57	24.51
A03	1.19	**0.78**	3.61	55.56	72.68	82.49	32.36	91.52	30.02	29.01
A04	1.20	**0.92**	3.74	75.38	99.03	111.71	35.52	94.88	35.72	29.76

5 结 论

针对抽油机井下工况复杂、种类繁多的特点，本文提出一种抽油机故障诊断的分布驱动主动学习算法。该算法首先利用大量结构化数据构造经验误差函数，结合主动学习查询少量关键样本，通过代价敏感方法优化算法模型，得到工况数据最佳聚类簇数来改善数据分布。有效利用迭代过程中的代价优化函数，使得该算法在抽油机示功图故障诊断方面较对比算法在精度上有较大提高。在小类别工况识别中，本文提出的算法在准确度和F-measure分数上明显优于其余对比算法。针对实际工程环境下未知工况的识别和诊断是下一步将要研究的内容。

参 考 文 献：

[1] 韩光.示功图分析与计量系统研发与应用[J].信息系统工程，2019，311(11)：103-104.
HAN Guang. Development and application of indicator diagram analysis and measurement system[J]. Information System Engineering, 2019, 311(11): 103-104.

[2] 刘卓，罗明良，刘飞，等.基于BP神经网络和不变矩特征的泵功图诊断方法研究[J].制造业自动化，2013(19)：7-9.
LIU Zhuo, LUO Mingliang, LIU Fei, et al. Research on diagnosis method of pump work diagram based on BP neural network and moment invariant features[J]. Manufacturing Automation, 2013(19): 7-9.

[3] 孔祥玉，冯晓伟，胡昌华.广义主成分分析算法及应用[M].北京：国防工业出版社，2018：47-75.
KONG Xiangyu, FENG Xiaowei, HU Changhua. Generalized principal component analysis algorithm and its application[M]. Beijing: National Defense Industry Press, 2018: 47-75.

[4] 王快妮.支持向量机鲁棒性模型与算法研究[M].北京：北京邮电大学出版社，2019：6-12.
WANG Kuaini. Research on robustness model and algorithm of support vector machine[M]. Beijing: Beijing University of Posts and Telecommunications Press, 2019: 6-12.

[5] 田增国，田东哲，姜宝柱，等.基于主成分分析方法的示功图故障诊断系统[J].内燃机与配件，2019，300(24)：155-157.
TIAN Zengguo, TIAN Dongzhe, JIANG Baozhu, et al. Indicator diagram fault diagnosis system based on principal component analysis method[J]. Internal Combustion Engines & Parts, 2019, 300(24): 155-157.

[6] 施海青，张韬，党延辉，等.基于支持向量机的抽油机故障诊断方法研究[J].中国石油石化，2017(11)：72-73.
SHI Haiqing, ZHANG Tao, DANG Yanhui, et al. Research on fault diagnosis method of pumping unit based on support vector machine[J]. China Petroleum & Petrochemical, 2017(11): 72-73.

[7] 杜娟，刘志刚，宋考平，等.基于卷积神经网络的抽油机故障诊断[J].电子科技大学学报，2020，49(5)：751-757.
DU Juan, LIU Zhigang, SONG Kaoping, et al. Fault diagnosis of pumping unit based on convolutional neural network[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(5): 751-757.

[8] ZHANG A, GAO X W. Fault diagnosis of sucker rod pumping systems based on curvelet transform and sparse multi-graph regularized extreme learning machine[J]. International Journal of Computational Intelligence Systems, 2018, 11(1): 428-437.

[9] ZHOU W, LI X, YI J, et al. A novel UKF-RBF method based on adaptive noise factor for fault diagnosis in pumping unit[J]. IEEE Transactions on Industrial Informatics, 2019, 15(3): 1415-1424.

[10] PENG P, ZHANG W J, ZHANG Y, et al. Cost sensitive active learning using bidirectional gated recurrent neural networks for imbalanced fault diagnosis-ScienceDirect[J]. Neurocomputing, 2020, 407: 232-245.

[11] JIN Y R, QIN C J, HUANG Y X, et al. Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network[J]. Measurement, 2020, 173: 108500.

[12] ZHANG Z S, LV W Z, SHEN M H. Active learning of support vector machine for fault diagnosis of bearings[C]//Proceedings of International Symposium on Neural Networks. Berlin, Heidelberg: Springer, 2006: 390-395.

[13] JIAN C X, YANG K J, AO Y H. Industrial fault diagnosis based on active learning and semi-supervised learning using small training set[J]. Engineering Applications of Artificial Intelligence, 2021, 104: 104365.

[14] CHEN M, ZHU K, WANG R, et al. Active learning-based fault diagnosis in self-organizing cellular networks[J]. IEEE Communications Letters, 2020, 24(8): 1734-1737.

[15] PUNČOCHÁŘ I, SKACH J. A survey of active fault diagnosis methods-ScienceDirect [J]. IFAC-Papers OnLine, 2018, 51(24): 1091-1098.

[16] 陈方杰，韩军，王祖武.基于改进网格划分统计的特征点快速匹配方法[J].计算机测量与控制，2019，27：231-235.
CHEN Fangjie, HAN Jun, WANG Zuwu. Fast matching method of feature points based on improved meshing statistics[J]. Computer Measurement & Control, 2019, 27: 231-235.

[17] 钟功祥,邹明铭.往复泵故障示功图灰度矩阵法特征量研究[J].机械科学与技术, 2016(2): 279-284.
ZHONG Gongxiang, ZOU Mingming. Study on the characteristic quantities of the gray matrix method of the reciprocating pump fault indicator diagram[J]. Mechanical Science and Technology, 2016(2): 279-284.

[18] WANG M, MIN F, ZHANG Z H, et al. Active learning[J]. Expert Systems with Applications, 2017, 85: 305-317.

[19] WANG M, LIN Y, MIN F, et al. Cost-sensitive active learning through statistical methods[J]. Information Sciences, 2019, 501: 460-482.

[20] WANG M, FU K, MIN F, et al. Active learning through label error statistical methods[J]. Knowledge-Based Systems, 2020, 189: 105-140.

[21] SALAKHUTDIVOV R, HINTON G. Replicated softmax: An undirected topic model[C]//Proceedings of International Conference on Neural Information Processing Systems.[S.l.]: Curran Associates Inc., 2009.

[22] ALLOCK J, ZHANG S Y. Quantum machine learning[J].National Science Review, 2019, 6(1): 26-28.

[23] ZHOU Z H, LIU X Y. On multi-class cost-sensitive learning[C]//Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference. Boston, Massachusetts, USA: AAAI Press, 2006.

[24] LIU X Y, LI Q Q, ZHOU Z H. Learning imbalanced multi-class data with optimal dichotomy weights[C]//Proceedings of IEEE International Conference on Data Mining.[S.l.]:IEEE, 2013.

[25] LIU X Y, WU J, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2009, 39(2): 539-550.

[26] MALOOF M A. Learning when data sets are imbalanced and when costsare unequal and unknown[J]. ICML, 2003,21: 1263-1284.

[27] ANDO S, HUANG C Y. Deep over-sampling framework for classifying imbalanced data[M].Cham:Springer, 2017.

[28] JIMENEZ-CASTANO C, ALVAREZ-MEZA A, OROZCO-GUTIERREZ A. Enhanced automatic twin support vector machine for imbalanced data classification[J]. Pattern Recognition, 2020, 107: 107442.

[29] TONG S, KOLLER D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research,2001,2: 45-66.

[30] JING F, LI M, ZHANG H J, et al. Entropy-based active learning with support vector machines for content-based image retrieval[C]//Proceedings of IEEE International Conference on Multimedia & Expo.[S.l.]: IEEE, 2004.

[31] HUANG K H, LIN H T. A novel uncertainty sampling algorithm for cost-sensitive multiclass active learning[C]//Proceedings of 2016 IEEE 16th International Conference on Data Mining (ICDM).[S.l.]: IEEE, 2016.

(编辑:刘彦东)