基于随机森林算法的大数据异常检测模型设计

宋世军[1]，樊 敏[2]

（1. 西南交通大学 交通运输与物流学院，成都 610031；2. 西南交通大学 土木工程学院，成都 610031）

摘 要：针对大数据异常检测过程易受边缘数据的干扰，导致大数据异常检测准确率较差的问题，提出了一种基于随机森林算法的大数据异常检测模型。首先，利用改进k-means算法对大数据实行聚类处理，采用主成分分析法提取大数据特征；然后，构建基于随机森林分类器的大数据异常检测模型，将提取的特征输入到模型中，构建决策树，并通过动态更新决策树的权重值提高分类器的分类精度；最后，输出分类结果，完成大数据的异常检测。实验结果表明，本文模型的检测时间约为25 s，大数据异常检测准确率平均值为91%，误报率为4.5%。

关键词：大数据聚类；特征提取；主成分分析法；随机森林分类器；决策树；更新权重

中图分类号：TM714 **文献标志码**：A **文章编号**：1671-5497(2023)09-2659-07

DOI：10.13229/j.cnki.jdxbgxb.20220598

Design of big data anomaly detection model based on random forest algorithm

SONG Shi-jun[1], FAN Min[2]

(1. *School of Transportation and Logistics, Southwest Jiaotong University, Chengdu* 610031, *China*; 2. *School of Civil Engineering, Southwest Jiaotong University, Chengdu* 610031, *China*)

Abstract: Aiming at the problem that Big data anomaly detection process is easily interfered by edge data, which leads to poor accuracy of Big data anomaly detection, a big data anomaly detection model based on Random forest algorithm was proposed. Firstly, the improved k-means algorithm was used to cluster the big data, and the principal component analysis method was used to extract the features of the big data. Then a big data anomaly detection model based on random forest classifier was built, the extracted features was inputted into the model, a decision tree was built, and the classification accuracy of the classifier was improved by dynamically updating the weight value of the decision tree. Finally, the classification results are output to complete the anomaly detection of big data. The experimental results show that the detection time of the proposed model is about 25 s, the average big data anomaly detection accuracy is 91%, and the false alarm rate is 4.5%.

Key words: big data clustering; feature extraction; principal component analysis; random forest classifier; decision tree; update weights

收稿日期：2022-05-18.

基金项目：国家自然科学基金重大专项项目(71942006)；中铁大桥勘测设计院集团有限公司科研项目(KYL202203-0086).

作者简介：宋世军(1981-)，女，博士研究生.研究方向：物流工程，智能计算，信息安全.E-mail：songshijun2022@yeah.net

通信作者：樊敏(1979-)，男，讲师，博士.研究方向：工程项目管理，施工监控，智能建造.E-mail：fanmin@swjtu.edu.cn

0 引 言

大数据能够存储、管理、分析海量规模的数据,被广泛应用于互联网能源、财务学等领域[1]。为了创造安全、高效的环境,人们对大数据异常检测模型提出了更高要求,不仅要求模型效率高,还要求其具有良好的分析性能[2],其中的关键就是对大数据异常检测模型做出升级和改进。因此,研究大数据异常检测模型设计具有重要意义。

万磊等[3]设计了记忆网络检测模型,并将置信度推理算法融入到模型中。首先,从大数据中抽取有效的数字浮动常数和边缘体积两种特征;然后,将特征输入到前置转换器中,通过证据推理输出新的置信度;最后,建立大数据异常检测制度,将新的置信度输入到记忆网络模型中训练,完成大数据的异常检测。该模型没有对大数据实施聚类处理,导致检测时间较长。李清[4]设计了结合粒子群算法与模糊C均值算法的大数据检测模型。首先,利用改进的粒子群算法搜索大数据中的有效函数,计算出大数据的初始中心和位置;然后,利用优化模糊C均值算法得到大数据的初始聚类分布情况;最后,将聚类分布情况输入到模型中,完成大数据异常检测模型的设计。该模型没有提取大数据特征,导致检测结果准确率低、误报率高。丁小欧等[5]设计了基于序列相关性的大数据异常检测模型。首先,对大数据中的多维时间序列实施标准化分段处理,得到相关矩阵模型;然后,根据相关性强度的不同将相关矩阵划分成时间序列团;最后,将所有时间序列团输入到模型中完成大数据异常检测。该模型没有更新大数据的权重值,导致检测结果的误报率较高。

为了解决上述方法中存在的问题,本文提出了一种基于随机森林算法的大数据异常检测模型。利用改进的k-means算法对大数据进行聚类和特征提取,获取大数据异常检测的特征值。通过随机森林算法构建决策树;通过动态更新决策树的权重值检测大数据中存在的异常值和边缘信息,实现大数据异常值的准确检测。

1 大数据预处理

1.1 大数据聚类

基于随机森林算法的大数据异常检测模型,利用改进k-means算法[6]对大数据聚类,获取大数据的基本结构,为特征提取和随机森林分类器提供基础条件,具体步骤如下所示。

(1)改进k-means算法采用密度函数得到大数据中所有数据点的密度值,融入无限大长度原则在高密度数据点中选取初始中心并保证其分布均匀[7]。首先,计算出大数据空间中任意数据点的近邻集合,集合半径为固定常数。计算公式如下所示:

$$\begin{cases} A=\{a_1,a_2,\cdots,a_n\},\ a_i\in A \\ \varepsilon\geqslant H(a_i,a_j) \end{cases} \tag{1}$$

式中:A为大数据集合;a_i为集合中的任意数据点;ε为邻域半径;H为近邻集合;i、j均为大数据点;n为数据点数量。

(2)计算大数据中所有数据点在邻域集合中的密度值$M(a_i)$和集合平均密度值$M(a_{ij})$,将密度值不高于平均密度值的数据点视为稀疏数据点,进行删除,直到所有数据点的密度值都大于平均密度值时,得到大数据密集点集合。计算公式如下所示:

$$\begin{cases} M(a_i)=\left(-H/(2\varepsilon^2)\right)\cdot\sum_{i=1}^{k}M \\ M(a_{ij})\leqslant\sum_{i=1}^{k}M(a_i)\Big/n \end{cases} \tag{2}$$

式中:M为密度值函数;k为邻域中大数据点的个数。

(3)在密集点集合中挑选出拥有最高密度值的点作为首个聚类中心点;取距离最远的大数据点作为中间聚类中心点;取与初始聚类中心点距离值最小的大数据点作为最后一个中心点,直到达到所需的初始聚类中心点个数。

$$\begin{cases} C=\left[M_{\max}(a_i,C_1),\cdots,d_{\min}(a_n,C_n)\right] \\ k\geqslant 3 \end{cases} \tag{3}$$

式中:C为聚类中心点集合;d为距离值。

(4)计算出大数据集合中所有数据点的欧氏距离[8],当获取最小数值时,将此数据点归入到密集点集合中,完成大数据的聚类处理。欧氏距离D的计算公式如下所示:

$$D(a_i,C_j)=\sqrt{\sum_{i=1}^{k}\left|a_i-C_j\right|^2} \tag{4}$$

1.2 特征提取

基于随机森林算法的大数据异常检测模型,将线性判别思想融入到主成分分析(Principal component analysis,PCA)算法中,然后采用PCA

法对聚类处理后的大数据实行逐一线性判别[9]，选择可以反映类间差异的主成分构造大数据特征空间，具体步骤如下所示。

（1）对经过聚类处理后的大数据样本实施PCA分析，得到由若干个主分量构成的子集；然后选择大数据中可以代表类间差别的主分量构造新的特征空间，公式如下所示：

$$\begin{cases} K=\max\left[\mathrm{var}(\boldsymbol{U}_x)/\mathrm{var}(\boldsymbol{U}_y)\right] \\ \boldsymbol{U}=[\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_s] \end{cases} \tag{5}$$

式中：K为大数据的特征向量集合；$\mathrm{var}(\cdot)$为类间差别；$\boldsymbol{U}$为由主分量构成的子集；s为特征向量的个数；$\boldsymbol{x}$、$\boldsymbol{y}$分别为集合中的特征向量。

（2）在二维投影条件下，计算出大数据样本的类间差别和类内差异[10]，公式如下所示：

$$\begin{cases} \mathrm{var}(\boldsymbol{U}_l)=\det\left[\sum_{l=1}^{E}(\bar{\boldsymbol{v}}_l-\bar{\boldsymbol{v}})(\bar{\boldsymbol{v}}_l-\bar{\boldsymbol{v}})\right]^T \\ \mathrm{var}'(\boldsymbol{U}_l)=\det\left[\sum_{l=1}^{E}\sum(\bar{\boldsymbol{x}}-\bar{\boldsymbol{v}})(\bar{\boldsymbol{x}}-\bar{\boldsymbol{v}})\right]x^T \end{cases} \tag{6}$$

式中：var'为类内差异；$\det(\cdot)$为引入的阵列函数；E为大数据样本的类别数；$\bar{v}$为样本的平均分量；$\bar{\boldsymbol{x}}$为大数据在空间上的映射；T为空间维度。

（3）将大数据样本的类间差异与类内差异结合，计算出大数据主成分的判别函数，按照从大到小的顺序完成排列。选取前3个判别函数数值最大的主成分构成新的特征空间。然后将特征空间做局部投影处理[11]，得到样本的新特征。

$$\begin{cases} \bar{\boldsymbol{v}}=\boldsymbol{U}_s^{\mathrm{T}}\bar{\boldsymbol{v}} \\ \bar{\boldsymbol{x}}=\boldsymbol{U}_s^{\mathrm{T}}\boldsymbol{x} \end{cases} \tag{7}$$

（4）在新特征中选取特征向量时，遵循类间差异与类内差异呈对称分布的选取原则，进而可以消除大数据样本中与异常检测模型不相关的差异向量。完成大数据特征的提取。

2　基于随机森林的检测模型

随机森林具有较高的异常检测准确率，对大数据中存在的异常值和边缘信息都具有较高的检测能力[12]，基于随机森林分类器的大数据异常检测模型设计，具体步骤如下：

（1）将PCA分析法提取的大数据特征输入到基于随机森林分类器的大数据异常检测模型中，将特征的权重之和视为一个整体。大数据的初始值可以适当改动，目的是提高模型异常检测能力的精准度。样本权重集合Z如下所示：

$$\begin{cases} Z=\left[\varpi_{1,1},\varpi_{1,2},\cdots\varpi_{1,k}\right] \\ k>1 \end{cases} \tag{8}$$

式中：ϖ为独立权重。

（2）基于随机森林算法的大数据异常检测模型，构建基于大数据样本的决策树[13]，更新样本中的权重值。然后对大数据样本随机抽样，生成训练集合，构建出决策树，并利用原始数据对新决策树进行分类，计算出误差率。决策树分类原理如图1所示。样本权重值更新公式如下所示：

$$Z_{m+1}=\left[\varpi_{m+1,1},\varpi_{m+1,2},\cdots,\varpi_{m+1,n}\right] \tag{9}$$

式中：m为更新次数。

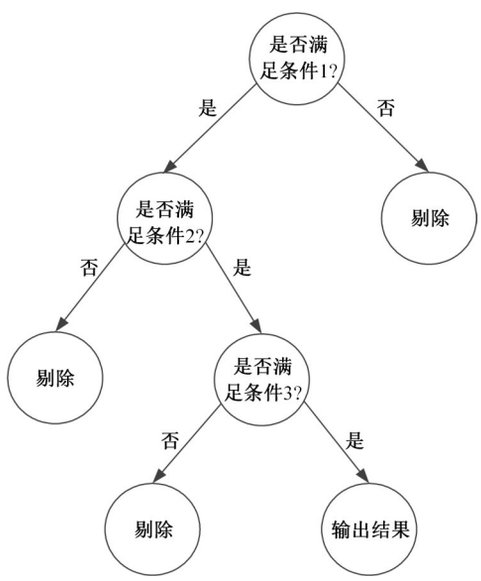

图1　决策树分类原理

Fig. 1　Principle of decision tree classification

（3）利用基于随机森林分类器的大数据异常检测模型，对于正确分类的样本其权重值更新后会降低，即在下一次分类器合成过程中，被检测到的概率会降低；相反，对于错误分类的样本，需要提高适应性才能保证下一次生成分类器后分类正确。因此，引入混淆因子保证每次权值更新后所有大数据的权重之和永远为一个整体[14]，公式如下所示：

$$\begin{cases} \varpi_{m+1,1}=\left(\varpi_{m+1,1}/f\right)\cdot\alpha\cdot e^{\pm\lambda} \\ f=\left[\sum_{m=1}^{Z}\varpi_{m+1,1}\right]\alpha\cdot e^{\pm\lambda} \\ \lambda=\ln\left((1-e_m)/e_m\right)\cdot\frac{1}{2} \end{cases} \tag{10}$$

式中：f为引入的混淆因子；α为一般参数；λ为权重平衡系数；e为正确分类结果下的样本数量。

(4)引入一般参数α的作用是用于区分基于随机森林的大数据异常检测模型分类中的少数类结果和多数类结果。因为少数类占大数据样本比重较小,所以分类器得到的训练经验较少,即少数类样本检测结果错误的代价相对更高,所以通过引入一般参数提高少数类样本的分类精度,进而提升随机森林分类器模型的总体性能[15]。

(5)基于随机森林算法的大数据异常检测模型总体性能提升后,对新的大数据样本依据权重值大小重新抽样,得到新的训练集,生成新的决策树。当所有决策树生成后,计算每棵决策树对应的分类器权重值,公式如下所示:

$$\varpi = 2/\left(X^{-1} + Y^{-1}\right) \tag{11}$$

式中:X为决策树分类正确的数量与所有样本的比值;Y为代价的平均值占样本平均值的比重。

(6)分类器权重值的大小衡量了基于随机森林算法的大数据异常检测模型的检测精准度和所需代价。由式(11)可知,当某棵决策树的分类精准度较高时,其正确分类的数量也越大,即X值越大;当这棵决策树的所需代价越低时,其Y值越大。如果某棵决策树的分类精准度最优、所需代价最低,则其分类性能最佳,对应的权重值也最大,这棵决策树输出的结果即为大数据异常检测结果,从而完成基于随机森林算法的大数据异常检测。

基于随机森林算法的大数据异常检测模型如图2所示。

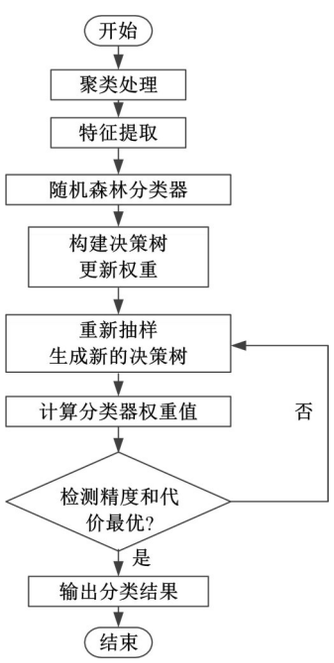

图2 基于随机森林检测模型流程

Fig. 2 Flow chart based on random forest detection model

3 实验及结果分析

为了验证基于随机森林算法的大数据异常检测模型的整体有效性,以模型检测的执行时间、准确率和误报率作为指标,对本文模型、文献[3]模型、文献[4]模型做对比测试。

(1)检测时间

采用本文模型、文献[3]模型和文献[4]模型识别15组大数据,对比不同模型对大数据异常检测的执行时间。执行时间越长,表明模型的执行效率越低;相反,执行时间越短,表明模型的执行效率越高,模型检测结果如表1所示。

表1 不同模型的检测时间

Table 1 Test time of different methods

实验序号	检测时间/s		
	本文模型	文献[3]模型	文献[4]模型
1	20	41	78
2	22	43	78
3	26	41	74
4	27	48	72
5	24	46	89
6	23	68	78
7	27	41	72
8	27	44	81
9	25	47	77
10	24	48	78
11	23	44	79
12	22	47	78
13	24	49	87
14	25	41	75
15	26	54	77

分析表1中的数据可知,针对大数据的异常检测,本文模型的检测时间约为25 s,文献[3]模型和文献[4]模型的检测时间分别在47 s和72 s附近波动;通过对比发现,在不同实验序号下本文模型的检测时间均小于文献[3]模型和文献[4]模型的检测时间,表明针对大数据的异常检测,本文模型的效率高于文献[3]模型和文献[4]模型的检测效率。

(2)准确率

准确率是指针对大数据的异常检测,模型检测出正确样本的数量占总数据样本的比例。准确率是评价大数据异常检测模型检测结果准确性的指标,针对大数据的异常检测,准确率越高,表明模型的准确性越高;反之,表明模型的准确性越低,其计算公式如下:

$$a=\frac{\mathrm{TP}+\mathrm{FP}}{N} \tag{12}$$

式中：a为检测准确率；TP为实际正常，检测结果一致的大数据个数；FP为实际异常，检测结果一致的大数据个数；N为大数据整体样本数量。

本文模型、文献[3]模型和文献[4]模型的准确率测试结果如图3所示。

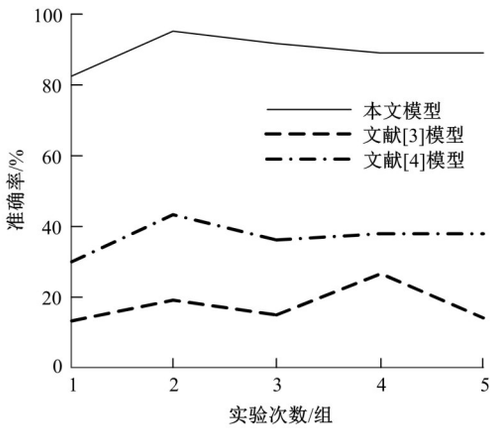

图3　不同模型的准确率

Fig. 3　Accuracy of different models

分析图3可知，在5组实验中，本文模型的准确率均高于文献[3]模型和文献[4]模型的准确率，平均值为91%，并且在不同组实验下，本文模型的检测准确率比较稳定，没有发生大幅度变化，而文献[3]模型和文献[4]模型的检测准确率随着实验次数的增加而发生明显浮动。说明针对大数据的异常检测，本文模型检测出正确样本数量占总样本数量比例高于文献[3]模型和文献[4]模型，本文模型的稳定性较强。

(3)误报率

误报率表示的是模型针对大数据异常检测结果中误差的存在情况，是评价模型精准度的重要指标。误报率越大，表明模型的精准度越低；反之，表明模型的精准度越高。误报率w的计算公式如下所示：

$$w=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{13}$$

式中：TN为实际异常，却被检测为正常的样本数量。

为保证测试结果的准确率，本次测试在10组大数据中完成，本文模型、文献[3]模型和文献[4]模型的测试结果如图4所示。

分析图4可知：本文模型的误报率小于文献[3]模型和文献[4]模型的误报率，平均值为

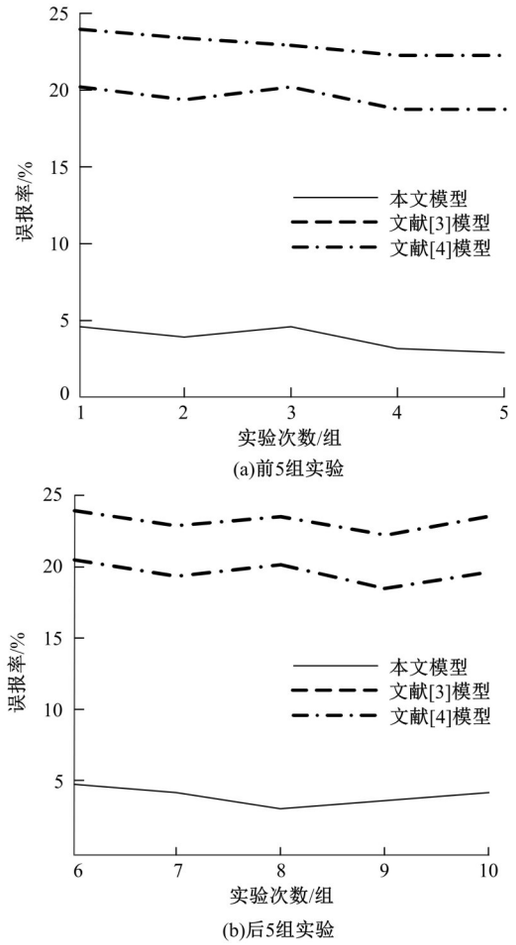

图4　不同模型的误报率

Fig. 4　False positive rate of different models

4.5%，表明本文模型对大数据的异常检测结果误差小于文献[3]模型和文献[4]模型的误差；并且随着实验次数的增加，本文模型的误报率未发生明显波动，而文献[3]模型和文献[4]模型的误报率结果不稳定。本文模型在对大数据进行异常检测前，对海量数据实行了聚类处理，预处理后的大数据在保留原始信息的同时消除了冗余信息；并且本文模型对大数据的权值完成了更新迭代计算，使之在大数据异常检测过程中不受边缘信息的干扰，从而保证检测结果的误差较小。

4　结束语

针对目前大数据异常检测模型存在检测时间长、检测准确率过低和误报率高的问题，提出了一种基于随机森林算法的大数据异常检测模型。该模型首先对大数据实行聚类处理；然后提取大数据特征；最后将提取的特征输入到基于随机森林分类器的检测模型中，构建了决策树，输出分类结

果,完成基于随机森林算法的大数据异常检测。实验结果表明,本文模型降低了执行时间和误报率,提高了模型的检测准确率。

参考文献:

[1] 刘永辉, 张显, 孙鸿雁, 等. 能源互联网背景下电力市场大数据应用探讨[J]. 电力系统自动化, 2021, 45(11): 1-10.
Liu Yong-hui, Zhang Xian, Sun Hong-yan, et al. Discussion on application of big data in electricity market in background of energy internet[J]. Automation of Electric Power Systems, 2021, 45(11): 1-10.

[2] 姜丹, 梁春燕, 吴军英, 等. 基于大数据分析的电力运行数据异常检测示警方法[J]. 中国测试, 2020, 46(7): 18-23.
Jiang Dan, Liang Chun-yan, Wu Jun-ying, et al. Alarm method of power operation data anomaly detection based on big data analysis[J]. China Measurement & Test, 2020, 46(7): 18-23.

[3] 万磊, 陈成, 黄文杰, 等. 基于BRB和LSTM网络的电力大数据用电异常检测方法[J]. 电力建设, 2021, 42(8): 38-45.
Wan Lei, Chen Cheng, Huang Wen-jie, et al. Power abnormity detection method based on power big data applying BRB and LSTM network[J]. Electric Power Construction, 2021, 42(8): 38-45.

[4] 李清. 基于改进PSO-PFCM聚类算法的电力大数据异常检测方法[J]. 电力系统保护与控制, 2021, 49(18): 161-166.
Li Qing. Power big data anomaly detection method based on an improved PSO-PFCM clustering algorithm[J]. Power System Protection and Control, 2021, 49(18): 161-166.

[5] 丁小欧, 于晟健, 王沐贤, 等. 基于相关性分析的工业时序数据异常检测[J]. 软件学报, 2020, 31(3): 726-747.
Ding Xiao-ou, Yu Sheng-jian, Wang Mu-xian, et al. Anomaly detection on industrial time series based on correlation analysis[J]. Journal of Software, 2020, 31(3): 726-747.

[6] 谢桦, 陈昊, 邓晓洋, 等. 基于改进k-means聚类技术与半不变量法的电-气综合能源系统运行风险评估方法[J]. 中国电机工程学报, 2020, 40(1): 59-69, 374.
Xie Hua, Chen Hao, Deng Xiao-yang, et al. Electric-gas integrated energy system operational risk assessment based on improved k-means clustering technology and semi-invariant method[J]. Proceedings of the CSEE, 2020, 40(1): 59-69, 374.

[7] 吴金蔚. φ-混合样本下密度函数在有限点处的联合渐近分布[J]. 信阳师范学院学报: 自然科学版, 2021, 34(4): 541-544.
Wu Jin-wei. The joint asymptotic distribution of probability density function in a finite number of points under φ-mixing samples[J]. Journal of Xinyang Normal University (Natural Science Edition), 2021, 34(4): 541-544.

[8] 张重远, 胡焕, 程槐号, 等. 基于欧氏距离分析的电力变压器绕组变形程度与类型的诊断方法[J]. 高压电器, 2020, 56(1): 224-230.
Zhang Zhong-yuan, Hu Huan, Cheng Huai-hao, et al. Diagnostic method to determine degree and type of winding deformation in power transformer based on euclidean distance[J]. High Voltage Apparatus, 2020, 56(1): 224-230.

[9] 代瑾, 陈莹. 联合线性判别和图正则的任务导向型跨模态检索[J]. 计算机辅助设计与图形学学报, 2021, 33(1): 106-115.
Dai Jin, Chen Ying. Joint Linear Discrimination and graph regularization for task-oriented cross-modal retrieval[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(1): 106-115.

[10] 蔡瑞初, 李嘉豪, 郝志峰. 基于类内最大均值差异的无监督领域自适应算法[J]. 计算机应用研究, 2020, 37(8): 2371-2375.
Cai Rui-chu, Li Jia-hao, Hao Zhi-feng. Unsupervised domain adaptive algorithm with intra-class maximum mean discrepancy[J]. Application Research of Computers, 2020, 37(8): 2371-2375.

[11] 胡善科, 秦玉华, 段如敏, 等. 联合矩阵局部保持投影的近红外光谱特征提取[J]. 光谱学与光谱分析, 2020, 40(12): 3772-3777.
Hu Shan-ke, Qin Yu-hua, Duan Ru-min, et al. Research on feature extraction of near-infrared spectroscopy based on joint matrix local preserving projection[J]. Spectroscopy and Spectral Analysis, 2020, 40(12): 3772-3777.

[12] 吴铮, 张悦, 董泽. 基于改进高斯混合模型的热工过程异常值检测[J]. 系统仿真学报, 2023, 35(5): 1020-1033.
Wu Zheng, Zhang Yue, Dong Ze. Outlier detection during thermal processes based on improved Gaussian mixture model[J]. Journal of System Simulation, 2023, 35(5): 1020-1033.

[13] 谢桦, 陈俊星, 赵宇明, 等. 基于SMOTE和决策树

算法的电力变压器状态评估知识获取方法[J]. 电力自动化设备, 2020, 40(2): 137-142.
Xie Hua, Chen Jun-xing, Zhao Yu-ming, et al. Knowledge acquisition method of power transformer condition assessment based on SMOTE and decision tree algorithm[J]. Electric Power Automation Equipment, 2020, 40(2): 137-142.

[14] 蔡瑞初, 白一鸣, 乔杰, 等. 基于混淆因子隐压缩表示模型的因果推断方法[J]. 计算机应用, 2021, 41(10): 2793-2798.
Cai Rui-chu, Bai Yi-ming, Qiao Jie, et al. Causal inference method based on confounder hidden compact representation model[J]. Journal of Computer Applications, 2021, 41(10): 2793-2798.

[15] 张清华, 庞国弘, 李新太, 等. 基于代价敏感的序贯三支决策最优粒度选择方法[J]. 电子与信息学报, 2021, 43(10): 3001-3009.
Zhang Qing-hua, Pang Guo-hong, Li Xin-tai, et al. Optimal granularity selection method based on cost-sensitive sequential three-way decisions[J]. Journal of Electronics & Information Technology, 2021, 43(10): 3001-3009.