【电子与信息科学 / Electronics and Information Science】

基于改进 k -means 的电力信息系统异常检测方法

黄 林¹, 常 健¹, 杨 帆¹, 李 忆², 牛新征²

1) 国网四川省电力公司信息通信公司, 四川成都 610015; 2) 电子科技大学计算机科学与工程学院, 四川成都 611731

摘 要: 电力信息系统可用于管控电力设备, 检测电力信息系统的异常对维持电力设备的稳定运行具有重要意义, 但传统的异常检测方法难以检测电力信息系统中存在的多个指标综合异常的情况, 为解决该问题, 提出一种基于改进 k -means 算法的异常检测方法. 将数据空间划分为网格, 以网格均值点映射该网格内所有样本点来压缩数据, 减少了计算量; 通过引入基于聚类边界密度和簇密度移动聚类边界的机制, 提高 k -means 算法的准确率, 以准确识别正常模式; 通过计算数据与正常模式的偏离程度, 检测异常. 实验结果表明, 该方法能准确挖掘多指标综合异常, 与其他异常检测方法比较, 检测运行时间由 16.44 s 减少到 0.55 s, 异常检测的准确率提高了 5.2%, 在电力运维异常检测领域具有良好的工程应用前景.

关键词: 电力信息系统; 模式识别; 异常检测; 数据压缩; k -means 算法; 聚类

中图分类号: TP391

文献标志码: A

doi: 10.3724/SP.J.1249.2020.02214

An anomaly detection method for electric power information system based on improved k -means

HUANG Lin¹, CHANG Jian¹, YANG Fan¹, LI Yi², and NIU Xinzheng²

1) State Grid Sichuan Electric Power Company Information and Communication Corporation, Chengdu 610015, Sichuan Province, P. R. China

2) School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan Province, P. R. China

Abstract: The electric power information system is usually used to control the power equipment. The anomaly detection of electric power information system is very important for maintaining the stable operation of power equipment. However, the traditional anomaly detection method is difficult to detect the comprehensive anomalies of multi-indicator in the electric power information system. In order to solve this problem, an anomaly detection method based on improved k -means algorithm is presented in this paper. To reduce the amount of calculation, the data space is compressed by dividing the data space into multiple grids and all the sample points in same grid are mapped by the mean point of grid. In order to accurately identify the normal mode, the accuracy of k -means algorithm is improved by the mechanism of moving cluster boundaries based on the cluster boundary density and cluster density. Then the anomalies are detected by calculating the deviation degree between the data and normal mode. The experimental

Received: 2019-09-19; **Accepted:** 2019-11-26

Foundation: Science and Technology Project of Sichuan Province (2017FZ0094); Science and Technology Project of Chengdu (2017-RK00-00021-ZF); Information and Communication Project of State Grid Sichuan Electric Power Company Information and Communication Corporation (SGSCXT00XGJS1800219)

Corresponding author: Senior engineer HUANG Lin. E-mail: 19798860@qq.com

Citation: HUANG Lin, CHANG Jian, YANG Fan, et al. An anomaly detection method for electric power information system based on improved k -means [J]. Journal of Shenzhen University Science and Engineering, 2020, 37(2): 214-220. (in Chinese)

results show that the proposed method can accurately mine the comprehensive anomalies on multi-indicator. Compared with other anomaly detection methods, the running time of our method is reduced from 16.44 seconds to 0.55 seconds and the accuracy of anomaly detection is improved by 5.2%. Our method has good application prospects in the field of power operation and maintenance anomaly detection.

Key words: electric power information system; pattern recognition; anomaly detection; data compression; k -means method; clustering

电力信息系统负责监控电网各平台的状态,保障电网各平台稳定运行.面向电力信息系统数据设计精准高效的异常检测方案,有助于运维人员快速定位故障,提高运维效率.

异常点与正常发生的事件具有较大差异,极大可能是由非正常机制产生的数据点造成的.传统基于规则的异常检测方法根据专家经验设置规则,检测系统的异常,依赖于预先定义好的数据指标特征.但随着电力基础设施的升级、电力信息系统监控指标及主机数量的不断增长^[1],当面对大量的主机以及多样的采集指标时,阈值设置复杂,工作量呈指数增加.因此,人工智能技术被更多应用于异常检测. k -means 算法作为一种经典的数据挖掘方法,被广泛用于数据的异常检测^[3-13].根据检测思想的不同,可分为两类.一是将数据空间聚类,每一类别由人工标记为正常类或异常类,检测异常.如马雪君^[4]通过自适应遗传算法进行聚类,动态识别异常类,获取异常模式.闫义涵^[5]根据标准差与信息熵动态分裂簇进行聚类,识别异常.该类方法的优点是能快速定位异常类别,但是由于异常类别难以枚举,异常类别可能增加,需要海量数据的支撑.二是将数据空间聚类,识别数据空间中正常模式,然后判断样本点到正常模式的距离是否超过设置的阈值来检测异常.如蒋华等^[6-7]采用基于密度和距离的方法选择初始聚类中心,进行聚类,挖掘正常模式,然后以3倍方差准则识别异常.除此之外,YIN等^[8-13]改进 k -means 算法,提高 k -means 算法全局搜索能力,提高异常检测的准确性,但其聚类过程中均以最近邻分配准则划分类别,未考虑各模式的差异性.本研究采用上述第2类异常检测思想,提出一种新的异常检测框架,首先采用网格映射的方法压缩数据,提高异常检测模型建立效率,然后以 BD k -means (border divide k -means) 算法识别系统中的正常模式,为提高电力信息系统数据异常检测的精确度提供参考.

1 基于 BD k -means 聚类的电力信息系统异常检测方法

基于 BD k -means 聚类的电力信息系统异常检测方法主要分为模式挖掘和异常识别两个过程(图1).

1) 模式挖掘:电力信息系统监控数据经压缩后,使用 BD k -means 算法对历史监控数据进行聚类,挖掘系统存在的正常模式.

2) 异常识别:计算样本点与正常模式的偏离程度,将偏离程度高于阈值的样本点视为异常点.

图1 异常检测框架

Fig. 1 Framework for anomaly detection

为便于描述本研究提出的异常检测方法,表1对参数进行了说明.

表 1 参数说明
Table 1 Parameters description

参数	说明
P	样本点集
$P[l]$	数据空间 P 中第 l 个样本
P_i	簇 i 样本点集
$P_i[l]$	簇 i 第 l 个样本点
w_i	$P[l]$ 的权重
k	聚类簇数
C_t	第 t 次迭代聚类中心集合
C_i	簇 i 聚类中心
s	边界范围
threshold _{i}	簇 i 的异常阈值

1.1 模式挖掘

模式挖掘的目的是从众多的历史数据中, 识别出系统的正常模式. 由于电力信息系统单位时间采集一次数据, 采集频率高、时间跨度大, 具有众多相似或重复数据, 在特征提取过程中将进行大量重复计算. 所以, 模式挖掘过程分为数据压缩和模式特征提取两个步骤.

1.1.1 基于网格的数据压缩

考虑到数据各维度的度量存在不同数量级问题, 数据压缩前, 本研究基于 minmax 准则, 将数据各维度统一规范化, 如式(1).

$$P[l] = \frac{P[l] - \min P}{\max P - \min P} \quad (1)$$

然后, 在数据空间中划分网格, 以各网格内样本点的均值及其权重因子 (各网格内样本点的数量) 作为新的数据集. 如图 2, 各网格的均值点及其权重组成的数据集 $\{(m_1, w_1), (m_2, w_2), (m_3, w_3), (m_4, w_4)\}$ 表示压缩后的数据集. 其中, m_i 为 region _{i} 内样本点均值, w_i 为 region _{i} 内样本点数.

1.1.2 基于 BDk-means 的模式特征提取

BDk-means 算法在传统 k-means 算法的基础上, 引入聚类边界再分配机制 (当边界密度大于簇密度时, 每次迭代使聚类边界向密度小的方向移动), 使边界密度逐步向下收敛, 得到边界清晰的聚类划分, 准确提取正常模式, 主要包含初始聚类中心选择、模式更新和聚类中心更新等 3 个步骤.

为便于描述本研究提出的 BDk-means 算法, 给出以下定义.

图 2 数据压缩示例

Fig. 2 Data compression example

定义 1 簇边界区间: 给定任意两个簇 i, j 和边界范围参数 s , 簇边界区间 (edge _{i, j}) 表示处于直线 L_1 和直线 L_2 之间的平面空间, 其中, $L_1(a_1x + b_1y + c_1)$ 和 $L_2(a_2x + b_2y + c_2 = 0)$ 分别表示线段 C_iC_j 的垂直平分线 L_0 前后平移 s 的直线, 如图 3.

图 3 边界划分示意

Fig. 3 Edge demarcation

定义 2 边界样本: 给定任意两个簇 i 和 j , 其边界样本 ePoint _{i, j} 表示处于平面空间 edge _{i, j} 内的样本点 (x, y) , 如图 3 中介于直线 L_1 和直线 L_2 之间的样本点 $P(x, y)$, 满足约束:

$$\begin{cases} a_1x + b_1y + c_1 > 0 \\ a_2x + b_2y + c_2 < 0 \\ P(x, y) \in P_i, P(x, y) \in P_j \end{cases} \quad (2)$$

1) 初始聚类中心选择

传统 k -means 算法初始聚类中心选择过程具有随机性, 选择不同的初始聚类中心, 算法陷入的局部最优解不同, 可能得到不理想的局部最优解. 因此, 本研究提出一种改进的初始聚类中心选择算法, 首先对于给定数据集建立如图 4 所示的坐标空间, 然后确定坐标空间中与原点距离的最大样本点 $\max P$ 和最小样本点 $\min P$, 并以原点为圆心将坐标空间切分为图 4 所示的 k 个等宽圆环空间, 每个圆环空间 C_i 表示为

$$C_i = \{in_r_i, out_r_i\} \quad (3)$$

其中, in_r_i 和 out_r_i 分别代表圆环空间的外环和内环, 具体定义为

$$\begin{cases} in_r_i = \text{dist}(o, \min P) + (i-1) \times \text{val} \\ out_r_i = \text{dist}(o, \min P) + i \times \text{val} \end{cases} \quad (4)$$

$$\text{val} = \frac{\text{dist}(o, \max P) - \text{dist}(o, \min P)}{k} \quad (5)$$

其中, $\text{dist}(o, \min P)$ 和 $\text{dist}(o, \max P)$ 分别表示原点 o 到样本点 $\max P$ 和 $\min P$ 的距离.

图 4 初始聚类

Fig. 4 Initial clustering

本研究将各圆环内样本点的集合视为初始簇, 并基于加权均值更新聚类中心:

$$c_i^j = \frac{1}{N} \sum_{l=1}^{|P_i|} (w_l \times P_i[l]) \quad (6)$$

$$N = \sum_{l=1}^{|P_i|} w_l \quad (7)$$

其中, $|P_i|$ 为 P_i 集合中样本点的个数; w_l 为经数据压缩后样本点 $P_i[l]$ 的权重.

2) 模式更新

一般来说, 电力信息系统数据模式分布具有差异性 (图 3), 模式 C_1 中样本离中心距离较小, 模式 C_2 中样本离中心距离较大.

传统 k -means 算法以最近邻准则标记样本点类别, 以图 3 中 $C_1 C_2$ 的垂直平分线 L_0 划分两种模式, 将 L_0 至 L_2 中的样本点划分为模式 C_1 , 然而 L_0 两侧样本点分布类似, 均为模式 C_2 中的样本点. 因此, 本研究提出一种边界再分配策略, 首先以最近邻准则划分初始簇, 分配样本点为其最近邻簇 P_m , 获得初步聚类划分 $P_1, P_2, P_3, \dots, P_k$, 样本点所属类别由式 (8) 确定:

$$m = \underset{i}{\operatorname{argmin}} \text{dist}(C, P_i[l]) \quad (8)$$

计算每一样本点与最近邻簇心 C_i 和次近邻簇心 C_j 垂直平分线的距离, 当其小于给定的边界范围 s , 则视该样本点为边界 $\text{edge}_{i,j}$ 的边界样本点 $\text{ePoint}_{i,j}$. 通过比较边界样本点密度 $\text{bDens}_{i,j}$ 与簇 i, j 相对聚类边界 $\text{edge}_{i,j}$ 的密度 $\text{cDens}_{i,j}$ 和 $\text{cDens}_{j,i}$, 当边界密度 $\text{bDens}_{i,j} > \text{cDens}_{i,j}$ 或 $\text{cDens}_{j,i}$ 时, 划分边界样本点到簇相对密度更大的簇, 否则以最近邻准则标记样本点类别.

边界密度 $\text{bDens}_{i,j}$ 为

$$\text{bDens}_{i,j} = \frac{|\text{ePoint}_{i,j}|}{s} \quad (9)$$

其中, $|\text{ePoint}_{i,j}|$ 为边界 $\text{edge}_{i,j}$ 样本点数量; s 为给定的边界范围参数. 簇 i 相对聚类边界 $\text{edge}_{i,j}$ 的簇相对密度为

$$\text{cDens}_{i,j} = \frac{|P_i|}{r_{i,j}} \quad (10)$$

$$r_{i,j} = \text{dist}(c_i, c_j)/2 \quad (11)$$

其中, $|P_i|$ 表示簇 i 的样本点数量.

3) 聚类中心更新

与传统 k -means 算法类似, BDk-means 算法以每类样本点的均值作为聚类中心. 不同点为当每一样本点 $P_i[l]$ 与其他聚类中心 $C_j - c_i^j$ 的最小距离 $\text{dMin}_i[l]$ 大于该样本点与原聚类中心的距离时, BDk-means 算法才更新该样本点所属簇的聚类中心, 以减少陷入局部最优解的概率. 即满足

$$\text{dist}(P_i[l], c_i^{j-1}) < \text{dMin}_i[l], \forall P_i[l] \quad (12)$$

$$\text{dMin}_i[l] = \min(\text{dist}(P_i[l], C^j - c_i^j)) \quad (13)$$

其中, $\text{dist}(P_i[l], C^j - c_i^j)$ 表示样本点 $P_i[l]$ 到 $C^j - c_i^j$ 距离的集合, 才更新聚类中心.

1.2 异常识别

异常识别过程基于已挖掘出电力信息系统的正常模式, 通过设置各模式的异常阈值, 识别异常^[14]. 鉴于电力监控数据中未记录历史故障数据以及各簇分布不服从正态分布, 本研究以各正常模式箱型图的上限作为异常阈值 threshold_i . 根据正常模式聚类中心及各模式异常阈值进行实时数据的异常识别步骤为:

- 1) 电力信息系统监控数据实时采集;
- 2) 计算实时监控数据与各聚类中心的距离 $D_CP_0, D_CP_1, \dots, D_CP_k$;
- 3) 当实时监控样本点与各正常模式中心的距离 D_CP_i 均大于异常距离 threshold_i 时, 则识别其为异常数据, 进行异常预警.

2 实验仿真及结果

2.1 数据集及实验环境

基于 BDk-means 聚类的电力信息系统异常检测实验, 实验数据选取某电力信息系统某一主机网络流入量和网络流出量两个维度近 1 年所产生的约 6 万条数据, 实验环境为 python3.6、win10 64 位、主频 2.3 GHz, 内存 8 Gbyte.

实验中将本研究提出的异常检测方法与传统 k -means、PSO- k -means^[9] 和 Minmax- k -means^[15] 聚类算法的异常检测方法相比较.

2.2 评价指标

2.2.1 模式挖掘评价指标

本实验使用聚类内部评价指标轮廓系数 (SilCoe) 和 S_Dbw_{new} ^[16-17] 评估模式挖掘的准确性.

1) 轮廓系数. 以样本点为单位评估聚类的准确性, 其取值在 $-1 \sim 1$, 计算公式为

$$\text{SilCoe} = \frac{b - a}{\max(a, b)} \quad (14)$$

其中, a 表示样本点与其所属簇其他样本点的平均距离; b 表示样本点与其所属簇的最近邻簇样本点的平均距离.

2) S_Dbw_{new} . 以簇为单位评估聚类的准确性, 其取值为 0 到正无穷, 计算公式为

$$S_Dbw_{\text{new}} = \text{Dens_bw}_{\text{new}} + \text{Scat}_{\text{new}} \quad (15)$$

其中, $\text{Dens_bw}_{\text{new}}$ 为聚类边缘密度与数据簇心密度之比; Scat_{new} 为簇标准差与数据空间整体标准差之比.

通常, 轮廓系数越接近 1, S_Dbw_{new} 越小, 表明挖掘的模式越准确.

2.2.2 异常识别评价指标

本研究提出异常距离 (abnormal distance, AbnDst) 评估异常检测结果的有效性, 计算公式为

$$\text{AbnDst} = \frac{\sum_{P \in P_i} \sum_{AP \in \text{AbnS}} \text{dist}(AP, P)}{|P_i| \times |\text{AbnS}|} \quad (16)$$

其中, $\text{dist}(AP, P)$ 表示异常样本点 AP 至正常样本点 P 的距离, $|\text{AbnS}|$ 表示异常数据集 AbnS 中元素的数量. 一般来说, 异常距离大小与异常可能性成正相关.

2.3 实验结果

异常检测算法对比实验统一采用未压缩的数据, 将本研究所提方法与其他 3 种方法相比较, 以保证对比实验的可靠性. 首先, 对比 4 种算法的聚类图, 不同类别的样本点在图中以不同的灰度显示, 如图 5.

传统 k -means 及 PSO- k -means 算法的各模式边界处样本点数量较多, 说明该部分样本点属于两种模式的可能性相似, 无法准确确认其所属模式. 而 Minmax- k -means 与本研究提出的 BDk-means 挖掘的各模式边界处样本点数量更少, 模式区分更加明显, 挖掘的模式准确率更高.

为进一步验证 BDk-means 算法的有效性, 对比基于 BDk-means 和 Minmax- k -means 异常检测的各指标值, 如表 2.

表 2 算法性能对比

Table 2 Performance comparison of algorithms

指 标	Minmax- k -means	BDk-means (数据未压缩)	BDk-means (数据压缩)
轮廓系数	0.829 35	0.831 74	0.831 74
S_Dbw_{new}	0.070 02	0.059 89	0.059 89
AbnDst	0.906 37	0.953 53	0.959 56
运行时间/s	16.440 39	8.315 89	0.549 76

BDk-means 算法比 Minmax- k -means 算法聚类结果的轮廓系数更接近 1, 具有更小的 S_Dbw_{new} 和更大的异常距离. 另外, BDk-means 算法运行时间为 Minmax- k -means 执行时间的 50%, 原因是

图5 聚类结果对比

Fig. 5 Comparison of cluster results

Minmax- k -means 算法每次迭代均需先后两次对每一样本点进行类别划分, 第1次计算各类别的权重, 第2次确定模式, 而本研究算法第2次仅需对边界处样本点统一划分模式, 使得模式挖掘效率有明显提升.

考虑到电力信息系统数据的特色, 增加数据压缩步骤后, 异常检测模型的运行时间由 8.32 s 降低至 0.55 s, 进一步提高了效率, 且各指标基本保持不变. 可见, 本研究提出的数据压缩方法能有效提高异常检测效率.

综上, 本研究提出的基于 BD k -means 聚类的异常检测方法对电力信息系统的异常检测具有一定参考价值.

结 语

提出一种基于 BD k -means 聚类的异常检测方法. 该方法基于网格映射进行数据压缩, 有效提高了异常检测效率, 并改进 k -means 算法, 解决了传统 k -means 算法聚类中心敏感、无法准确挖掘分布差异过大模式的问题. 实验结果表明, 与现有算法相比, 本研究异常检测方法在准确度及效率方面有了较大提升, 对电力信息系统的异常检测具有重要的指导意义.

基金项目: 四川省科技计划资助项目 (2017FZ0094); 成都市科技资助项目 (2017-RK00-00021-ZF); 国网四川省电力公司信息通信公司资助项目 (SGSCXT00XGJS1800219)

作者简介: 黄 林 (1983—), 国网四川省电力公司高级工程师. 研究方向: 电力信息化. E-mail: 19798860@qq.com

引 文: 黄 林, 常 健, 杨 帆, 等. 基于改进 k -means 的电力信息系统异常检测方法[J]. 深圳大学学报理工版, 2020, 37(2): 214-220.

参考文献 / References:

- [1] 张明明, 袁国泉. 浅谈基于 ITIL V3.0 的电网信息通信运维流程设计[J]. 电信科学, 2017, 33(增刊1): 71-74.
ZHANG Mingming, YUAN Guoquan. Discussion on operation and operation process planning of power network information communication based on ITIL V3.0 [J]. Telecommunications Science, 2017, 33 (Suppl. 1): 71-74. (in Chinese)
- [2] 陈兴蜀, 江天宇, 曾雪梅, 等. 基于多维时间序列分析的网络异常检测[J]. 工程科学与技术, 2017, 49(1): 144-150.
CHEN Xingshu, JIANG Tianyu, ZENG Xuemei, et al.

- Network anomaly detector based on multiple time series analysis [J]. *Advanced Engineering Sciences*, 2017, 49 (01): 144-150. (in Chinese)
- [3] 吴镜锋, 金炜东, 唐 鹏. 数据异常的监测技术综述 [J]. *计算机科学*, 2017, 44(增刊2): 24-28.
WU Jingfeng, JIN Weidong, TANG Peng. Survey on monitoring techniques for data abnormalities [J]. *Computer Science*, 2017, 44(Supply2): 24-28. (in Chinese)
- [4] 马雪君. 大规模网络流量异常检测方法研究[D]. 长春: 吉林大学, 2018.
MA Xuejun. Research on the method of traffic anomaly detection for large-scale network [D]. Changchun: Jilin University, 2018. (in Chinese)
- [5] 闫义涵. 基于 k -means 的入侵检测方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2017.
YAN Yihan. Research of intrusion detection method based on k -means [D]. Harbin: Harbin Industrial University, 2017. (in Chinese)
- [6] 蒋 华, 季 丰, 王慧娇, 等. 改进 k -means 算法的海洋数据异常检测[J]. *计算机工程与设计*, 2018, 39(10): 3132-3136.
JIANG Hua, JI Feng, WANG Huijiao, et al. Improved k -means algorithm for ocean data anomaly detection [J]. *Computer Engineering and Design*, 2018, 39(10): 3132-3136. (in Chinese)
- [7] ZHANG L, DENG S, LI S. Analysis of power consumer behavior based on the complementation of k -means and DBSCAN [C]// *IEEE Conference on Energy Internet and Energy System Integration (EI2)*. Beijing: IEEE, 2017: 1-5.
- [8] YIN C, ZHANG S, WANG J, et al. An improved k -means using in anomaly detection [C]// *The 1st International Conference on Computational Intelligence Theory, Systems and Applications (CCITSA)*. Yilan: IEEE, 2015: 129-132.
- [9] 张仁斌, 许辅昊, 刘 飞, 等. 基于 k -均值聚类的工业异常数据检测[J]. *计算机应用研究*, 2018, 35(7): 2180-2184.
ZHANG Renbin, XU Fuhao, LIU Fei, et al. Industry anomaly data detection based on k -means clustering [J]. *Application Research of Computers*, 2018, 35(7): 2180-2184. (in Chinese)
- [10] 辛 壮, 万 良, 李均涛. 改进的聚类算法在网络异常行为检测中的应用[J]. *计算机技术与发展*, 2019, 29(3): 111-116.
XIN Zhuang, WAN Liang, LI Juntao. Application of improved clustering algorithm in network abnormal behavior detection [J]. *Computer Technology and Development*, 2019(3): 111-116. (in Chinese)
- [11] 邵 伦, 周新志, 赵成萍, 等. 基于多维网格空间的改进 k -means 聚类算法[J]. *计算机应用*, 2018, 38(10): 2850-2855.
SHAO Lun, ZHOU Xinshi, ZHAO Chengping, et al. Improved k -means clustering algorithm based on multi-dimensional grid space [J]. *Journal of Computer Applications*, 2018, 38(10): 2850-2855. (in Chinese)
- [12] ANAND S J V, PRANAV I, NEETISH M, et al. Network intrusion detection using improved genetic k -means algorithm [C]// *International Conference on Advances in Computing, Communications and Informatics*. Bangalore: IEEE, 2018: 2441-2446.
- [13] 王丝丝, 张敬磊, 王建兴, 等. 基于改进群智能 k -means 优化算法的侵犯性驾驶行为异常点检测[J]. *科学技术与工程*, 2018, 18(34): 208-214.
WANG Sisi, ZHANG Jinglei, WANG Jianxing, et al. Outlier detection of reckless driving behaviors based on improved swarm intelligent k -means algorithm [J]. *Science Technology and Engineering*, 2018, 18(34): 208-214. (in Chinese)
- [14] SAVAGE D, ZHANG X, YU X, et al. Anomaly detection in online social networks [J]. *Social Networks*, 2014, 39: 62-70.
- [15] 张孝远, 张新萍, 苏保平. 基于最小最大核 k 均值聚类算法的水电机组振动故障诊断[J]. *电力信息系统保护与控制*, 2015, 43(5): 27-34.
ZHANG Xiaoyuan, ZHANG Xinping, SU Baoping. Vibrant fault diagnosis for hydro-turbine generating unit using minmax kernel k -means clustering algorithm [J]. *Power System Protection and Control*, 2015, 43(5): 27-34. (in Chinese)
- [16] TONG Jianhua, TAN Hongzhou. Clustering validity based on the improved S_DBW index [J]. *Journal of Electronics (China)*, 2009, 26(2): 258-264.
- [17] LIU Y, LI Z, XIONG H, et al. Understanding of internal clustering validation measures [C] // *IEEE International Conference on Data Mining*. Sydney, Australia: IEEE, 2010: 911-916.

【中文责编: 坪 梓; 英文责编: 木 柯】