DOI:10.19651/j.cnki.emt.2106847

基于高斯混合模型的智能电表误差数据挖掘与分析方法

舒珏淋[1] 张 力[2] 胡 建[1]
(1.中电科大数据研究院有限公司 贵阳 550002;2.贵州电子科技职业学院 贵阳 550025)

摘 要:为了合理科学的选择误差最小的智能电表供给用电客户,设计了基于高斯混合模型的智能电表误差数据挖掘与分析方法。首先,分析了高斯混合模型与 EM 算法的基本思路,其次对智能电表误差数据求 3 次标准差作为建模数据,建立了以高斯混合算法为基础的智能电表误差数据模型,最后与传统的 K-means 聚类算法模型进行对比测试。实验结果表明,相对于其他聚类算法,所设计的方法轮廓系数值更大,性能更优。能够用于在大量的数据中寻找误差最小的智能电表,并能够给智能电表厂家反馈产品意见,同时还具有产品市场划分等功能。

关键词:智能电能表;误差分析;高斯混合模型聚类;EM 算法
中图分类号:TP391.1;TN98 **文献标识码**:A **国家标准学科分类代码**:510.4030

Data mining and analysis method for smart meter error data based on Gaussian mixture model

Shu Juelin[1] Zhang Li[2] Hu Jian[1]
(1. CETC Data Research Institute Co., Ltd., Guiyang 550002, China; 2. Guizhou Electronic Technology College, Guiyang 550025, China)

Abstract: In order to reasonably and scientifically select the smart meter with the smallest error to supply electricity to customers, a data mining and analysis method for smart meter error based on Gaussian mixture model is designed. First, the basic ideas of Gaussian mixture model and EM algorithm are analyzed. Secondly, the standard deviation of the smart meter error data is calculated as the modeling data, and the error data model of the smart meter based on the Gaussian mixture algorithm is established. Finally, it is combined with the traditional K-means. Class algorithm model for comparison test. The experimental results show that compared with other clustering algorithms, the designed method has a larger contour coefficient value and better performance. It can be used to find the smart meter with the smallest error in a large amount of data, and it can give feedback to the smart meter manufacturer on the product, and it also has functions such as product market division.

Keywords: smart energy meter; error analysis; Gaussian mixture model clustering; EM algorithm

0 引 言

智能电表是电力系统最为基础,最为广泛的实施场景,因此智能电表的性能影响着电力系统的经济命脉[1]。近年来,随着互联网的逐渐发展,智能电网的规模也逐渐扩大,与此同时电网中的智能电表的数量和种类也发生了跨越式的增长,然而智能电表的种类多样化,对电力公司如何选择智能电表造成了困扰,与此同时智能电表的大规模安装应用,使得用电客户对智能电表的准确度提出了更多的质疑[2]。那么面对如此种类繁多的智能电表,如何去作出选择,是当下急需解决的一个重要问题。智能电表的检查人员在日常工作中会对所发现的误差数据进行记录和归档,方便日后的查看、统计、分析。这些智能电表的误差数据蕴含着大量有用的信息,对误差数据进行挖掘和分析对智能电表的选择和性能的提升有着重要的意义[3-4]。但是当前对于智能电表的误差数据的挖掘和分析仍然缺乏理论性和准确性,主要体现在分析方法简单,仅仅对小量误差数据的统计和简单分类对比,没有对大量的智能电表数据,以智能电表检测误差数据为基础,进行深入挖掘,分析的结果对于智能电表的选择没有太大的实际参考性。目前,在电力领域,许多的学者已经利用数据挖掘技术去解决一些相应的

收稿日期:2021-06-03

实际问题。文献[5]以数据挖掘技术中Apriori的算法建立了基于关联规则的二次设备缺陷模型,之后利用该模型找到二次设备的薄弱部分并分析得出导致薄弱部分的原因,同时还能分析设备的家族性缺陷等功能。文献[6]基于数据挖掘技术,利用主成分分析方法建立了电力设备运行状态的综合评价模型,用于对电力设备是否异常进行快速检出。文献[7]利用电力企业积累的各类数据,采用数据挖掘技术构建反窃电模型,识别用户是否窃电,以给高层进行决策。数据挖掘技术在智能电表中的分析运用并不多,主要侧重于基于数据挖掘实现对智能电表的用电数据的分析。由于当前对于智能电表误差数据的挖掘与分析处理过程中的理论性和准确性不足,主要表现在对于误差数据的处理方法较为简单,仅对于少量误差数据进行统计和简单分类对比,缺乏对大量的误差数据进行深入挖掘,从而导致分析的结果对于智能电表的选择的实际参考性较小。

针对以上问题,本文提出一种基于高斯混合模型的智能电表的误差数据挖掘与分析方法,首先分析了高斯混合模型聚类的基本思路,其次建立了关于聚类算法的智能电表误差模型,最后以某电力局3年的智能电表误差数据为例,阐述了基于高斯混合模型聚类的智能电表误差数据挖掘与分析,实现了对于不同厂家智能电表的合理分类,得出误差最小智能电表类型。

1 聚类与高斯混合模型

1.1 聚　　类

聚类分析是一种分类技术,同时它也是数据挖掘和人工智能领域中的一项重要的数据处理技术[8]。是一种无标签技术,通过元素对象的相似性进行分类,简言之把不同对象类间的相似度趋于相同而分一类[9]。需要注意的几点,聚类与分类是有区别的,聚类是一种无监督的学习方式,不需要进行预先设置条件类的训练实例对象,而分类是一种有监督模式的学习,需要依赖提前设定条件类的训练对象,聚类需要先对所有的个体对象进行适应度评估,再根据个体与簇之间的相似度或者数据本身之间距离决定其划分为若干组,划分的原则是组内的个体之间聚类最小,而不同组之间的距离最大化[10-11]。

1.2 高斯混合模型

高斯混合聚类算法(Gaussian mixture model,GMM)它假定利用某一个给定参数的多元高斯分布生成全部的样本数据,同时它也是一种以概率数值为基础的聚类算法[12]。因此一个高斯混合模型的概率密度函数可以表示为:

$$p(x)=\sum_{i=1}^{k}\omega_i\cdot p(x\mid u_i,\sigma_i) \tag{1}$$

其中,σ 为多元高斯分布的概率密度函数,$p(x\mid\mu,\sigma)$ 为均值向量,k 个不同的多元高斯分布成份,而 ω_i 为高斯混合模型中的权重,且有 $\sum_{i=1}^{k}\omega_i=1$ 每一个 component 都满足 $N_K\sim(\mu_k,\sigma_k)$,$k=1,2,\cdots,k$ 并对应的是一个聚类中心[13-14]。

1.3 EM算法估计GMM参数

EM算法是一种比较典型的迭代型算法,它能有效解决存在隐含变量优化问题。该算法主要分为两步:即求期望(expectation)步骤和最大化(maximization)步骤[15]。

如图1所示,假设 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$,$\boldsymbol{x}$ 是图中所有点,每个点在二维平面上有两个坐标,是二维向量,GMM模型中有3个参数需要估计,分别是 π,μ 和 $\sum$,公式如下:

$$p(x\mid\pi,\mu,\sum)=\sum_{k=1}^{k}\pi_kN(x\mid\mu_k,\sum_k) \tag{2}$$

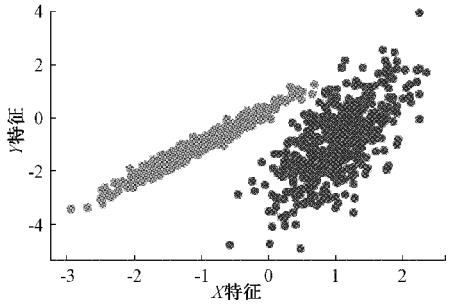

图1 聚类散点图

先求解 μ_k 的最大似然函数,其次对 $p(x\mid\pi,\mu,\sum)$ 进行求对数,最后对 μ_k 求导并令导数为0,即得到最大似然函数。

$$0=-\sum_{n-1}^{N}\frac{\pi_kN(x_n\mid\mu_k,\sum_k)}{\sum_j\pi_jN(x_n\mid\mu_j,\sum_j)}\sum_k(x_n-\mu_k) \tag{3}$$

两边同乘 $\sum_{-1}^{k}$,重新整理可以得到:

$$u_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n \tag{4}$$

其中,

$$N_k=\sum_{n-1}^{N}\gamma(z_{nk}) \tag{5}$$

式中:N 表示点的数量,$\gamma(z_{nk})$ 表示点聚类之后的后验概率,N_k 表示第 k 个聚类的点数量。那么 μ_k 表示所有点的加权平均,因此每个点的权值是 $\sum_{n\ 1}^{N}\gamma(z_{nk})$,跟第 k 个聚类有关。同理,求 $\sum_k$ 的最大似然函数,可以得到:

$$\sum_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^{\mathrm{T}} \tag{6}$$

之后剩下 π_k 的最大似然函数。因此加入拉格朗日:

$$\ln p(x\mid\pi,\mu,\sum)+\lambda\left(\sum_{k=1}^{k}\pi_k-1\right) \tag{7}$$

根据式(7),继续进一步计算可以得到:

$$0=\sum_{n=1}^{N}\frac{N(x_n\mid\mu_k,\sum_k)}{\sum_j\pi_jN(x_n\mid\mu_j,\sum_j)}+\lambda \tag{8}$$

式(8)两边同乘 π_k 可以得到 $\lambda=-N$,进而可以得到化简的表达式:

$$\pi_k=\frac{N_k}{N} \tag{9}$$

EM 算法估计 GMM 参数即最大化式(4)、(6)和(9)。首先指定 π,μ 和 Σ 的初始值,代入贝叶斯公式中计算出 $\gamma(z_{nk})$,然后再将 $\gamma(z_{nk})$代入式(4)、(6)和(9),求得 π_k,μ_k 和 Σ_k;接着用求得的 π_k,μ_k 和 Σ_k 再代入贝叶斯公式得到新的 $\gamma(z_{nk})$,再将更新后的 $\gamma(z_{nk})$代入式(4)、(6)和(9),如此往复,直到算法收敛。

2　基于高斯混合模型数据挖掘方法的设计

针对于市场上用户对于智能电表类型盲目的选择,以及数据挖掘技术在智能电表中的分析运用并不多,主要侧重于在用户用电数据和小量的智能电表数据进行挖掘分析,忽略了从智能电表误差数据作为特征进行挖掘分析,因此本文从智能电表的误差数据方面进行研究,并且设计详细的智能电表误差数据挖掘分析流程。设计流程如图 2 所示。

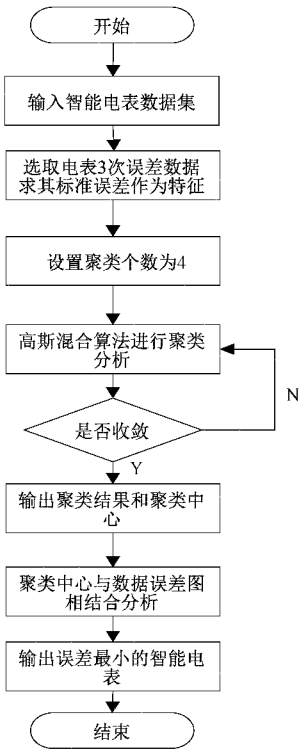

图 2　高斯混合模型数据挖掘方法设计流程

如图 2 所示为基于高斯混合智能电表数据挖掘方法模型,首先输入智能电表数据,根据智能电表检测的 3 次误差数据,求其标准误差,作为数据挖掘的建模数据,其次以标准误差数据进行高斯混合聚类分析,通过设置聚类个数将智能电表类型大体上分为 4 类,最后将各聚类结果的聚类中心与智能电表误差数据图进行相结合分析,从而得到误差最小的智能电表类型,进而把误差最小的智能电表类型推荐给用户。

3　实验与结果分析

3.1　数据集

本文使用的数据集来自于某电力公司从生产系统中导出的 3 年智能电表误差数据,总共有 70 多万条,对该电力公司的智能电表误差数据进行数据预处理后得到了建模数据,共有 766 013 条样本,经过数据独热编码之后,样本中共有 22 个特征,分别为误差数据和智能电表类型。样本中智能电表的类型共有 21 种,DTSI188 智能电表所占数量最大,共有 206 282 台,而 DSSD188S 智能电表所占数量最小,共有 108 台。智能电表误差数据平均值为 0.050 825。

3.2　误差数据的分块处理

由于实验数据是来自于电力公司的智能电表 70 多万条误差数据,数据量较大,因此综合分析决定采用 Python 来处理数据。实验在单台 CPU 3.5 GHz,Intel Core i7,内存 8 GB,硬盘 3 TB Fusion Drive 的计算机上完成,使用的数据挖掘分析工具 juypter notebook,Python 版本为 3.6。采用 Pandas 提供的 IO 工具可以将大文件分为块处理并读取,经过性能测试,得出完整加载智能电表误差数据需要的时间大约为 162 s,使用不同分块大小来读取再调用 pandas 库的 API 连接 DataFrame,将 ChunkSize 设置为 10 万条左右时,速度优化比较明显。下面是统计数据,ReadTime 是数据读取时间,TotalTime 是 Pandas 读取数据进行聚合操作总的时间,根据数据总体数量来看,对 5～20 个 DataFrame 对象进行合并处理,性能表现较好。实验结果如表 1 所示。

表 1　数据分块读取时间

ChunkSize	ReadTime/s	TotalTime/s
1 000	14.631	17.505
10 000	8.859	9.569
100 000	7.563	8.632

如表 1 所示,对于需要使用大量的数据进行分析时,使用 Python 分块处理大量数据,分块参数设置的越大,Python 读取数据的时间越短。因此提高了数据的挖掘分析效率。

3.3　基于高斯混合的智能电表误差数据建模

为了提高智能电表的性能水平,实现电力公司对智能电表的选择购买,巡检人员对智能电表进行 3 次误差检测得到的误差数据,都应及时将误差信息录入生产系统,生产系统管理着历年各种类型智能电表数据。

1)针对某电力公司,收集其不同种类的智能电表误差数据,由于记录的是 3 次误差数据,因此将 3 次误差数据求其标准误差。

$$\sigma=\sqrt{\frac{E_1^2+E_2^2+\cdots+E_n^2}{n}}=\sqrt{\frac{\sum E_i^2}{n}} \tag{10}$$

式中:σ 为智能电表标准误差,E 为智能电表误差。

2)由于误差数据中的电表类型特征是离散型特征,因此在进行聚类时,需要对其数字编码,通过编码后的数据即可用于聚类建模。本文采用 One-Hot 编码对离散型特征

进行处理，使得电表类型数值化[16]。

3)智能电表的数据具有不同的特征，因而不同特征具有不同的量纲，数值间会有差别。因此需要对智能电表的误差数据进行数据预处理，利用数据标准化处理以解决特征不同量纲带来的偏差，具体而言，将智能电表误差数据按照比例进行缩放，使之落入一个特定的区域，便于进行综合分析。

4)利用机器学习的高斯混合聚算法对已经预处理的智能电表误差数据进行挖掘建模，通过得到模型用于不同厂家智能电表的合理分类，得出误差最小智能电表类型，并且推荐给用户，较于实际中用户盲目选择智能电表有着现实意义。

3.4　性能指标测试

本文将聚类算法的轮廓系数作为数据挖掘的性能指标，轮廓系数指标公式如下所示：

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{11}$$

其中，$a(i)$为 i 点与本类其他点不相似度的平均值，$b(i)$为 i 点到其他类平均不相似度的最小值。当 $s(i)$聚类越接近 1，说明样本 i 聚类越合理；$s(i)$越接近 -1，说明样本 i 更应该分类到另外的簇；若 $s(i)$近似为 0，则说明样本 i 在两个簇的边界上。

为了进一步的验证本文提出的基于高斯混合模型的智能电表误差数据挖掘方法的有效性，选择在相同的智能电表标准误差数据集，相同的聚类个数下，将 K-means 聚类算法与高斯混合算法进行比较，以轮廓系数作为评价指标，比较结果如图 3 所示。

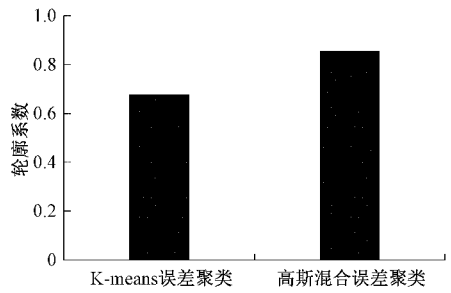

图 3　误差数据聚类比较图

如图 3 所示，相同的智能电表误差数据，高斯混合算法轮廓系数为 0.856，而 K-means 算法轮廓系数为 0.678。由此可以验证高斯混合算法性能要优于 K-means 聚类算法，因为 K-means 算法拟合效果，类的形状不够灵活，拟合结果与实际相差较大，精度有限。而高斯混合聚类针对于特定特征的样本数据进行聚类后得到的簇作为同一类特征数据，后续只针对不同的簇进行建模，建模效果更优。

3.5　应用分析

采用高斯混合模型对智能电表误差数据进行聚类，聚类结果的定量性能评价指标有互信息、同质性和完备性等，但是这些指标并不能指示聚类结果是否达到预期分析目标。本文中，分析目标是确定具有相同误差智能电表的分类，便于挖掘出误差最小的智能电表群体，从而为用电客户提供智能电表选择参考的目的。因此，需要的不是定量的评价指标结果，而是定性地对聚类结果进行分析。样本聚类之后的每一类的样本数目如表 2 所示。

表 2　智能电表聚类类别

样本类别	样本数目
0	37 599
1	143 614
2	78 960
3	457 143

由表 2 可知，聚类的 4 个类中，最大的类中有 457 143 台智能电表，最小的类中有 37 599 台智能电表。由于高斯混合聚类会随机选取初始的聚类中心，因此每次运行的结果可能会不同。为了更好地表达每一个类所代表的智能电表群体的特点，本文进一步进行实验分析，经实验后得出如表 3 所示结果。

表 3　智能电表聚类中心

智能电表类型	类别	聚类中心
MODEL_DTSK719Z	0	4.401 503
MODEL_DSSD188S	0	0.011 875
MODEL_DSZ71	0	0.011 984
MODEL_DTZ71	0	0.036 587
MODEL_DTSI720	1	2.081 126
MODEL_DSSD188S	1	−0.011 875
MODEL_DSZ71	1	−0.011 984
MODEL_DTZ71	1	−0.036 587
MODEL_DSSD71	2	5.572 802
MODEL_DSSD188S	2	0.011 875
MODEL_DSZ71	2	0.011 984
MODEL_DTZ71	2	0.036 587
MODEL_DTSI188	3	0.222 069
MODEL_DTSK188-Z	3	0.124 010
MODEL_DTSK217-Z	3	0.120 085
MODEL_DTSK719S-Z	3	0.095 248

根据聚类的特点，通过观察每一类的聚类中心，发现其群体特点。在智能电表数据已经使用 Z-score 方法进行标准化之后，可以直接通过观察聚类中心在每一个变量上的取值情况来分析每一个聚类中心的含义。如果聚类中心在某一个变量取值大于 0，代表该聚类所代表的群体在该变量取值大于群体平均水平。为了进一步提供群体智能电表的决策条件，本文以智能电表的 3 次误差数据作为特征，加

入智能电表数据，利用高斯混合聚类算法进行聚类分析，可以得到如图 4 所示结果。

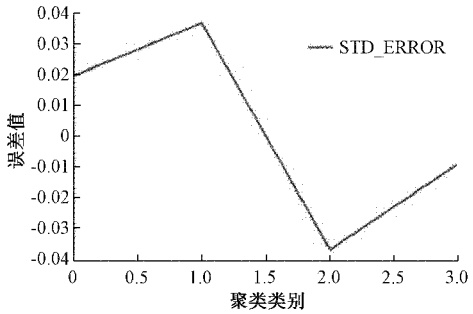

图 4　智能电表误差聚类

聚类中心如表 3 所示，对于智能表群体 0，该群体主要的特点是 MODEL_DTSK719Z 型号的智能表占大多数，高于群体水平，同时根据图 4 可知，该群智能电表的误差偏大，表明该群体的智能电表误差影响较大。对于智能表群 1，该群 MODEL_DTSI720 型号的智能表最多，但是智能表误差与群 0 相比较，则更大，表明该群智能电表很大程度上受到误差影响。对于智能表群 2，MODEL_DSSD71 型号的智能表最多，然而智能表误差为负值，基本对表群没有影响。对于智能表群 3，MODEL_DTSI188 型号的智能表高于整体水平，误差为负值。因此，根据图 2，可以选择智能电表群 2 的 MODEL_DSSD71 型号的智能表作为参考智能表，提供给用电客户。

本文针对某电力公司生产系统中实际数据集进行分析，得出了对智能电表误差数据集进行高斯混合挖掘建模的应用实例，满足于实验利用模型对不同厂家智能电表的合理分类，得出误差最小智能电表类型，并且推荐给用户的结论。通过进一步的实验有效验证了本文所构建模型系统的有效性，能在大量智能电表类型中详细挖掘出误差最小的智能电表，并且推荐给用户，较于实际中用户盲目选择智能电表有着现实意义。

4　结　　论

本文基于高斯混合聚类模型对智能电表误差数据挖掘与分析方法进行了研究，并以算例的方式将该方法应用到某电力公司智能电表误差数据的挖掘和分析中。通过分析智能电表误差数据挖掘的结果，得到如下结论。

1）该方法能够有效分析智能电表误差数据，得出某群体误差较小的智能电表类型，提供给电力公司做出决策，提供给用电客户做出选择。

2）该方法可以作为智能电表市场划分，通过智能电表的误差数据，以及不同类型的智能电表特征对智能电表群体进行划分，以每个智能电表群体的特点进行分析，以供给智能电表厂家作为选择目标市场和制定市场营销策略，提高企业经济效益。

3）该方法能够从大量误差数据中挖掘出智能电表群中误差最大的一类智能电表，以便于给智能电表厂家反馈意见，提高智能电表厂家的产品质量。但由于从生产系统中导出的智能电表误差数据，对智能电表信息的记录只在于误差数据和智能电表类型，因此对智能电表数据进行挖掘分析时，由于缺乏其他特征，不能从影响智能电表性能的各方面因素进行挖掘分析，从而造成分析的局限性。在下一步工作中，将着手解决智能电表其他方面的特征，以保证能从多方面因素对智能电表性能进行综合分析，届时将本文方法应用于生产管理系统的海量数据中，以获得对智能电表的选择和智能电表性能的提高更加有意义的结论。

参考文献

[1]　韩春玲. 基于大数据的智能电能表关键组件技术的研究综述[J]. 电气应用，2019，38(4)：56-63，71.

[2]　黄艳，周文斌，吴晓昱，等. 智能电能表的发展应用及误差调整[J]. 电测与仪表，2012，49(S1)：36-40.

[3]　刘军，何佳，尤金伟，等. 单相智能电能表小电流下误差试验及数据分析[J]. 工业计量，2020，177(6)：35-37，41.

[4]　郭丽娟. 智能电能表全生命周期质量评价方法研究[D]. 保定：华北电力大学，2014.

[5]　张延旭，胡春潮，黄曙，等. 基于 Apriori 算法的二次设备缺陷数据挖掘与分析方法[J]. 电力系统化，2017，41(19)：147-151，163.

[6]　贺川双，杜修明，严英杰，等. 基于数据挖掘和主成分分析的电力设备状态评价[J]. 高压电器，2017，53(12)：34-41.

[7]　刘盛，朱翠艳. 应用数据挖掘技术构建反窃电管理系统的研究[J]. 中国力，2017，50(10)：181-184.

[8]　王鸿玺，李飞，林志文，等. 基于 IK-means 的用电行为研究[J]. 国外电子测量技术，2020，39(1)：54-58.

[9]　白宁. 一种基于 k-均值聚类的异常检测技术[J]. 计算机与现代化，2014，4(1)：93-95，113.

[10]　张丹丹，游子毅，郑建，等. 基于改进的局部异常因子检测的优化聚类算法[J]. 微电子学与计算机，2019，36(11)：43-48.

[11]　王岩，王聪英，申艳梅. 改进的蜂群优化聚类集成联合相似度推荐算法[J]. 计算机工程，2020，46(10)：88-94，102.

[12]　何庆，易娜，江新勇，等. 基于高斯混合模型的最大期望聚类算法研究[J]. 微型电脑应用，2018，34(5)：50-52，75.

[13]　宋磊，郑宝忠，张莹，等. 一种基于高斯混合模型的改进 EM 算法研究[J]. 应用光学，2013，34(6)：985-989.

[14]　王垚，柴变芳，李文斌，等. 一种基于逆模拟退火和高斯混合模型的半监督聚类算法[J]. 南京师大学报(自然科学版)，2017，40(3)：67-73.

[15]　刘帅，刘长良，甄成刚. 基于数据分类重建的风电机组

故障预警方法[J]. 仪器仪表学报，2019，40(8)：1-11.
[16] 张朝龙，何怡刚，杜博伦，等. 基于深度学习的电力变压器智能故障诊断方法[J]. 电子测量与仪器学报，2020，34(1)：81-89.

作者简介

舒珏淋，硕士，助理工程师，主要研究方向为大数据挖掘分析技术等。
E-mail：503989543@qq.com

张力，硕士，助教，主要研究方向为大数据信息安全技术等。
E-mail：16478386498@qq.com

胡建，硕士，助理工程师，主要研究方向为大数据挖掘技术等。
E-mail：993115697@qq.com