基于 K−means 聚类电力大数据审计证据发现研究

李春艳

（云南省能源投资集团有限公司，云南昆明 650228）

摘　要　研究通过 K−means 聚类算法进行电力大数据审计证据发现的技术过程，以寻找一种普适性的电力大数据审计证据发现模式，改变以往就特定问题开发相应系统的被动状态。采集电网的运行、调度、营销数据，使用回归法、差分法、导数法等进行数据治理，增加数据的丰度，进而使用 K−means 聚类算法为核心算法的迭代分析法寻找数据中的特征数据点，进而发现相应问题的数据审计证据。经过测试，在较大数据集迭代 30 次的离线数据分析基础上，对数据的分析敏感度超过 85%，在较小数据集迭代 70 次的在线数据分析基础上，对数据的分析敏感度超过 91%。证实基于 K−means 聚类分析的数据审计过程可以作为电网电力运行相关大数据审计的重要技术过程。

关键词　聚类分析；电力大数据；数据审计

Research on Power Big Data Audit Evidence Discovery Based on Means Clustering

Li Chunyan

(Yunnan Provincial Energy Investment Group co.,LTD,Kunming Yunnan 650228)

Absrtact　This paper studies the technical process of power big data audit evidence discovery through K-means clustering algorithm, in order to find a universal power big data audit evidence discovery mode, and change the passive state of corresponding system development for specific problems in the past. The operation, dispatching and marketing data of power grid are collected, and regression method, differential method and derivative method are used for data governance to increase the abundance of data. Then, K-means clustering algorithm is used as the core algorithm of iterative analysis method to find the characteristic data points in the data, and then find the data audit evidence of the corresponding problems. After testing, on the basis of offline data analysis with 30 iterations for large data sets, the analysis sensitivity of data is more than 85%, and that of online data analysis based on 70 iterations of small data sets is more than 91%. It is confirmed that the data audit process based on K-means clustering analysis can be used as an important technical process of big data audit related to power grid operation.

Keywords　Big Data; Power Audit; Evidence Analysis

电力审计过程是基于电力运行大数据，对电力运行过程进行综合数据分析，以寻求相应的数据证据，为后续的电力运行策略优化或者电力事故追责提供数据支持。电力大数据的来源，主要来自电网运行过程中 CT 互感器采集的电力二次的电流、电压波形数据，即 U-t、I-t 录波图数据，电力营销的收费数据，开关量即开关位置时序数据，调度、运行、检修、巡线等日志数据等[1]。将这些数据进行分析，寻求离群数据，或获得短期内的数据剧烈变化沿，将成为电力大数据审计的重要证据来源[2]。当前的智能电网数据管理中，上述绝大部分数据被存储在电力 IDC 的数据仓库系统中，且数据在使用过程中被人工智能系统、数据服务系统等进行实时的不同深度的数据挖掘，形成了较多深度挖掘数据，这些衍生数据也将为电力大数据审计提供数据来源[3]。

数据审计是发现事故原因和管理体系、技术体系

收稿日期：2021−4−15

作者简介：李春艳 (1975−07)，女，彝族 云南个旧 ,高级会计师 大学 研究方向：审计、财务、投资管理。

隐患的重要途径，也是当前电力运行事故追查的重要途径。系统分析电力数据审计的关键要素，寻找电力数据审计和证据发现过程的通用的可行路径，是本文的重要创新点。本文研究中，使用 K-means 算法对数据进行离群分析，在原始数据离群数据中可以获得直接的瞬时问题数据，在对原始数据进行差值分析或趋势分析后，对不同程度的挖掘数据进行离群分析可以进一步得到问题数据[4]。这些问题数据将成为电力大数据审计的重要证据数据[5]。本文以 K-means 数据分析法为核心数据分析方法，对不同来源的数据进行分析，以探讨电力大数据审计过程中的审计证据获取方法。

1 数据来源及审计目标

1.1 电力大数据的数据来源

智能电网条件下，各变电站已经建立了完善的大数据系统，对电网中的分布式探头系统可以采集各重要状态点的电流、电压、温度数据，及重要设备的震动噪声数据等。这这数据被存储在电力 IDC 数据机房的数据仓库系统中，均可以作为电力大数据审计的数据来源[6]。

当前技术条件下，电力二次数据系统的运行模式主要为 CBA 模式，即在电力二次 IDC 系统中布局基于云计算（C.C.）大数据（B.D.）人工智能（A.I.）算力资源模块，进而实现对电力二次数据的充分利用[7]。综合分析当前电力物联网及外网服务系统的数据来源，可以得到图 1。

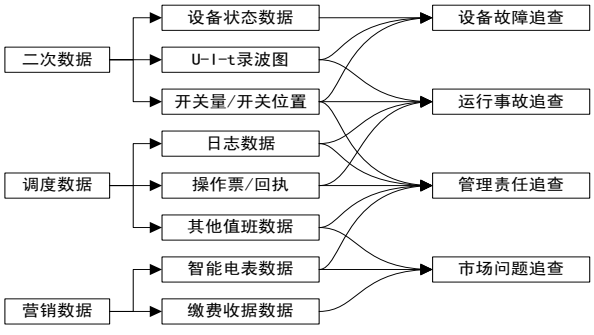

图 1 电力审计数据的来源分析

图 1 中，二次数据主要来自电力二次的运行数据，该数据以时域离散数据为主，主要来自以 CT 线圈采集的电流、电压录波图数据，部分数据来自电网设备的监控数据，包括负荷监控、震动监控、噪音及电弧监控、温度监控等；调度数据主要来自调度系统中的操作票数据及操作票的执行回执数据，各部门（含运行、巡线、运行、检修等）的工作日志数据，其他值班部门（如营销、市场监察、技术等）的值班日志数据等；营销数据主要为智能电表系统记录的实时负荷数据，电费数据，电费缴费的收据数据（主要包括电费缴费量、欠费量、缴费来源账户信息等）。通过对上述数据进行关联分析，发现不同数据层次下（含直接数据、差值数据、均值数据、同环比数据等）的离群数据，并进行数据离群原因分析。通过该分析，可以对事故及故障原因、管理责任、市场问题（含窃电）等进行审计证据采集。

1.2 电力数据审计的目标

数据审计的核心目标，是发现电力数据中的强特征数据，此类强特征数据，一方面是偏离正常运行状态的离群数据或者故障数据，一方面是变化趋势脱离常规变化趋势的强特征数据[8]。如何在电力数据审计过程中发现此类数据特征，成为当前电力数据审计过程中的前沿课题。

上述数据中，大部分数据属于离散时序数列数据，部分日志等数据产生的非标准化数据，也可以通过数据标准化的途径，使其转化为离散时序数列数据[9]。对一般化的时序数列数据来说，其数据特征如图 2。

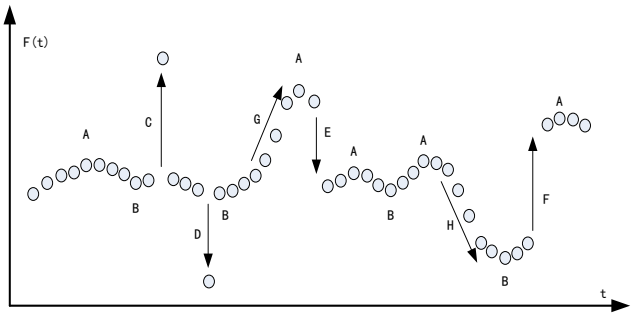

图 2 数据审计特征分析目标示意图

图 2 中，A、B 分别为周期峰值和周期谷值，C 为正离群数据，D 为负离群数据，E 为向下数据断崖，F 为向上数据断崖，G 为连续数据急上升沿，H 为连续数据急下降沿。其中，A、B 特征可以通过对回归方程求导直接获得；C、D 数据可以直接通过对原始

数据 K-means 聚类算法获得；E、F 可以通过数据时域差值数据的 K-means 聚类算法获得；G、H 可以对回归方程求导后对斜率变量进行 K-means 聚类算法获得。可以看到，除 A、B 数据特征外，其他数据特征的提取，均使用到了 K-means 算法。

2 聚类算法对电力大数据审计的实际意义

部分研究认为电力大数据属于标签数据，即时序指针变量被认为是电力大数据的数据标签，但从统计学意义视角分析，时序指针变量在电力大数据中仅提供数据序列标识，并不能提供任何统计学意义。即在 U-I 录波图中，电压 U 数据与电流 I 数据之间可以互为标签数据，但在 U-t 录波图数据和 I-t 录波图数据中，t 数据仅作为数据的序列标签，并不能提供数据标签的统计学意义[10]。所以，在聚类算法的选择方面，需要采用无标签数据的聚类方法。无标签数据的聚类方法指数据源仅 1 列数据，并无其他相关数据，此时对该列数据执行聚类计算，实现一维数据的聚类。

对输入数据 $X=[x_1,x_2,...x_n]$无标签数据集来说，采用 K-means 聚类对其进行 k 个数据中心的聚类计算：

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|^2 \tag{1}$$

式中：k 为数据中心的数量；C 为聚类后的分组数据簇，$C=[C_1,C_2,...,C_k]$，式中的 C_i 为聚类结果中的第 i 个 C 簇；μ_i 为 C_i 的数据中心点，其表达式为：

$$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x_i \tag{2}$$

式中：变量定义与公式（1）一致。

在计算机算力辅助下，对特定数据集的原始数据进行充分遍历和充分迭代，最终得到 K-means 聚类分析的近似解。其具体步骤依照下述 4 个步骤进行计算机软件开发：

1）在样本 X 中选择 k 个聚类点，并定义各聚类点的中心点进行预定义 $\mu=[\mu_1,\mu_2,...\mu_k]$；

2）计算样本 X 中每个数据点与 k 个簇中心点之间的距离：

$$Dist(X_i)=\sqrt{x_i^2-\mu_i^2} \tag{3}$$

式中：$Dist(X_i)$为点距离函数，x_i 为样本 X 中第 i 个元素的实际值；μ_i 为第 i 个聚类中心的坐标值；

将公式（3）计算结果存入最近的簇中，即 $x_i\in\mu_{nearest}$；

3）利用公式（2）对聚类中心进行重计算；

4）循环迭代上述第 2、3 步，直至数据收敛；

在电力 IDC 计算资源算力的支持下，且数据审计过程为离线数据审计，审计占用的 IDC 时间没有明确限制，所以可以通过加深上述循环迭代深度的情况下，获得更详细的 K-means 聚类结果[11]。

3 其他辅助数据分析方法

前文分析中，如果对图 2 中所有数据特征进行分析，需要牵扯到对数据的回归分析、求导计算、差值数列计算、以及针对差值数据或求导后的斜率数据差值计算等[12]。其中：

3.1 数据的回归分析方法

数据回归分析的主要目标是在序列数据中提取特征矩阵，即将序列数据格式转化为特征矩阵数据格式。电力大数据中绝大部分数据为时序序列数据，所以大部分数据在进行聚类分析之前，有必要进行数据回归分析。在 SPSS 数据分析系统中，对数据进行基于曲线估计的回归分析计算方法均已经做到对计算过程的充分封装，所以，在实际计算过程中，只需要选择对常规回归分析的基函数选择即可完成相应计算过程[13]，其常用基函数详见表 1。

表 1 回归分析的常用基函数表

函数类型	基函数	待回归变量
线性	$Y=A\cdot X_i+B$	A，B
多项式	$Y=\sum_{j=0}^{n}A_jX_i^j$	A_j
对数	$Y=A\cdot\log_e(C\cdot X_i+D)+B$	A，B，C，D
幂	$Y=A\cdot e^{C\cdot X_i+D}+B$	A，B，C，D
指数	$Y=A\cdot X_i^j+B$	A，B

表 1 中，多项式函数中 j 因子是多项式的阶数，j 可以取任何整数，当 j=1 时，该函数的本质是线性

函数，当 j=2 时，该函数的本质是二次函数，当 j=3 时，该函数的本质是三次函数；对数函数和幂函数中，e 为自然常数，在实际运作中，往往取其近似值 e=2.718281828；指数函数中，j 为指数函数的阶数，其统计学意义同多项式函数中的阶数定义。

3.2 数据的差值计算方法

差值计算的目的是通过残差算法将序列数据的特征充分放大，使离群数据的离群距离得到放大，以方便后续分析。差值计算中，主要有两种差值计算方案，其一是数列中相邻数据的直接差值计算，其二是按照一定周期进行差值计算[14]，其差值计算的方程式为：

$$\Delta x_m = x_i - x_{i-m} \tag{4}$$

式中：m 为差值计算的差值周期，默认值为 m=1，此时差值结果为相邻数据的直接差值计算，但 m 可以取值为其他自然数，即可以在差值计算过程中定义不同数据间距之间的差值。当 Δx_m >0 时，认为数据处于上升沿，当 Δx_m <0 时，认为数据处于下降沿。对 Δx_m 进行聚类，可以对数据变化趋势进行深度数据分析。

3.3 非标数据的标准化处理方法

数据差值计算可以应用于所有时域离散序列数据，即对所有录波图数据、设备监控数据、电力营销数据等，均为直接的时域离散序列数据。而关于操作票、日志等也可以进行进行非标数据的标准化[15]。本文设计中，拟采用分段投影法进行非标数据的标准化计算，详见表 2。

表 2 分段投影法的数据标准化策略表（局部）

数据赋值	1	2	3	4	5
开关状态	通	——	——	——	断
设备故障状态	烧毁	故障停机	状态检修	带故障运行	完好
温度状态	室温	——	正常	——	超温
日志状态	任务失败	任务放弃	局部完成	基本完成	任务完成

表 2 中，对设备故障，工作日志等，根据其实际表达的非标语言诠释程度进行 5 级分段投影的量化，比如在设备故障状态时，当设备完好时定义为 5，设备完全烧毁时定义为 1，日志状态方面，当任务圆满完成时定义为 5，任务彻底失败时定义为 1。该两类数据均可以实现在 1-5 的 5 段分段投影中均有数据表达。而在设备温度状态时，当设备处于室温时定义为 1，此时设备大概率处于不开机状态，而设备超温时定义为 5，设备温度正常时定义为 3。开关状态的投影中，开关通定义为 5，开关断定义为 1。

3.4 大数据审计聚类算法的整体流程设计

上述算法的核心部分为数据聚类分析模块，核心目标是将特征数据阵列进行聚类分析，最终形成数据审计报告。而为了实现该聚类分析，需要对数据进行诸多前置处理，包括数据标准化处理、差值序列提取、序列数据回归及回归特征提取等，此类模块的很目的是将电力运行大数据转化为特征矩阵数据[16]。上述流程如图 3 所示。

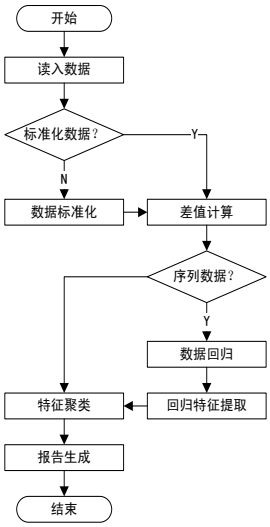

图 3 大数据审计聚类算法的整体流程图

图 3 中，该流程主要分为 6 个阶段：①数据读取阶段将电力大数据中的二次数据、调度数据、营销数据等读入到系统内（如前文图 1）；②数据标准化阶段，对标准化数据可以直接进入下一部差值计算，对非标准化数据需要经过该步进行标准化处理；③差值计算的目的是通过差值序列进一步强化序列变化规律特征，放大离群数据的偏移量（如前文图 2）；④数据回归阶段是将序列数据整理为特征矩阵数据，直接生成特征矩阵数据的无需进行数据回归；⑤聚类分析阶段是将特征矩阵内的离群数据进行提取并标记；

⑥报告生成阶段是将所有离群数据汇总为台账，并实现与原始数据的对应关系。

4 大数据审计聚类算法的实际应用效果

为了验证该算法在实际电力大数据中的应用效果，分别从静态数据审计和动态数据审计两个方面对数据审计效果进行分析。仿真数据为我单位 2020 年全年的实际运行数据，仿真环境为 matlab 下加载 simulink 控件构建的仿真环境，仿真结果如下：

4.1 静态数据审计结果

数据审计系统属于大数据分析系统的范畴。而所有大数据系统的设计原则，是调用最少的 IDC 算力资源实现更稳定更高响应效率的数据处理过程。即应该考察相关算法支持下的大数据分析系统算力占用量和系统响应周期表现。

静态数据的审计过程，主要面向较长周期内，一般周期在 1 年以上，甚至某个项目周期或者某个管理人员任期内的数据，进行离线分析，寻找相应的问题证据。所以，首先应对特定数据量下的系统响应时间进行仿真实测，其次对仿真过程对问题证据的发现效率进行实测。静态数据的离线审计对审计速度要求不高，但对审计精度要求较高。

按照 30 次迭代的 K-means 分析过程，系统响应时间的测试结果如图 4。

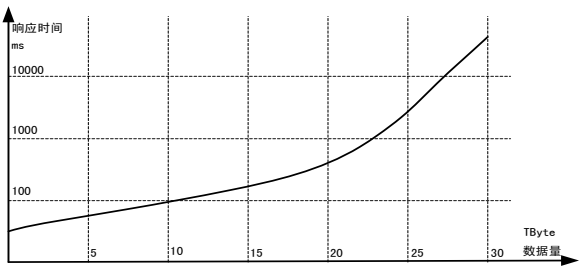

图 4 系统响应时间实测结果图

图 4 中，纵轴（响应时间）为对数轴，横轴（数据量）为线性轴。当数据量超过 27TByte 时，系统响应时间达到 10s 以上，当数据量超过 23TByte 时，系统响应时间达到 1s 以上。即在该 IDC 系统中，当数据量超过 20TByte 时，响应时间的上升速率显著加快。图 3 右侧未展示数据中，当数据量超过 55TByte 时，其响应时间超过 12h。

在占用最小系统算力资源的前提下，实现更准确的数据分析结果，即通过将 K-means 迭代次数压缩到 30 次时，考察本文算法对相应数据的分析结果与实际运行结果之间的差异性。特别是静态分析的数据量较大，而数据算法相对简单，迭代深度较低，主要考察了系统在强数据弱算法条件下的表现能力。对审计精度进行分析，选择 5 组数据，按照 30 次迭代的 K-means 分析过程，分析事故原因、管理责任等，以人工介入的综合判读结果作为比较数据，分析其实际数据分析结果，最终得到表 3。

表 3 数据分析精度对比表（25−30TByte）

数据分组	1	2	3	4	5
运行事故原因	12/15(80.014)	18/(77.816)	19/(84.211)	13/(84.612)	14/(85.7)
设备事故原因	8/9(88.9)	8/8(100)	7/8(87.5)	9/9(100)	8/9(88.9)
管理责任事故原因	4/4(100)	4/5(80.0)	4/4(100)	5/6(83.3)	4/4(100)
窃电行为	6/6(100)	6/7(85.7)	6/6(100)	7/7(100)	5/6(83.3)
平均敏感度	92.2%	85.9%	92.9%	92.0%	89.5%

表 3 中，5 次数据分析测试中，分别获得数据敏感度结果 92.2%、85.9%、92.9%、92.0%、89.5%，即除第 2 次测试外，所有测试结果的敏感度均达到了 89%以上。其中多项测试均达到了 100%的准确率。

4.2 动态数据审计结果

数据动态审计是在电网系统运行过程中，根据实时数据给出动态数据审计结果，此过程要求响应时间足够短，所以无法进行大数据量长周期的数据审计，但要求数据精度更高。与静态数据审计过程的强数据弱算法的分析策略相比，动态数据审计要求使用弱数据强算法的数据分析策略。此时，考察系统调用算力的极限处理能力，以确保在系统设计周期内及时给出高质量的数据分析结果。部分研究通过引入模糊阵列的方式或引入模糊神经网络的方式增加数据审计精度，但本文研究中通过将表 3 数据精度分析中 30 次迭代的迭代次数扩展到 70 次，以增加数据精度。70 次迭代计算的前提下，系统响应时间显著增加，在数据量达到 12.5TByte 时，系统响应时间就达到了 1s 级别，而在图 3 中，系统响应时间达到 1s 时，其数据量为 23TByte。

70 次迭代计算的前提下，继续比较 5 次数据分析精度，表 3 中 5 次分析的数据量均在 25-30TByte 之间，而表 4 中 5 次分析的数据量在 10-15TByte 之间。详见表 4。

表 4 数据分析精度对比表（10—15TByte）

数据分组	1	2	3	4	5
运行事故原因	13/14 (92.9)	12/12(100)	12/13 (92.3)	14/15 (93.3)	14/14(100)
设备事故原因	8/8(100)	9/9(100)	8/9(88.9)	8/8(100)	7/8(87.5)
管理责任事故原因	4/4(100)	4/4(100)	5/5(100)	5/5(100)	4/5(80.0)
窃电行为	5/5(100)	5/6(83.3)	5/5(100)	6/6(100)	6/6(100)
平均敏感度	98.2%	95.8%	95.3%	98.3%	91.9%

表 4 中，5 次测试的平均敏感度均达到了 90%以上，其中第 1 次和第 4 次测试的平均敏感度均超过了 98%。相比较表 4 与表 3，通过加大 K-means 分析的迭代深度，可以有效提升数据审计的敏感度。但是，加大 K-means 迭代深度的方式，会大幅度增加系统的响应时间，如果要保障更大的数据分析量，则需要提高 IDC 资源的投入量，增加了数据维护成本。

5 总结

传统的电力数据审计过程，一般采用按需求进行大数据系统开发的审计模式，很少有通用的数据审计策略。本文研究的核心创新点是发现一种普适性的数据审计算法，该算法综合采集电网的运行、调度、营销数据，使用回归法、差分法、导数法等进行前置数据治理，增加数据的丰度，进而使用 K-means 聚类算法为核心算法的迭代分析法寻找数据中的特征数据点，进而发现相应问题的数据审计证据。经过测试，在较大数据集迭代 30 次的离线数据分析基础上，对数据的分析敏感度超过 85%，在较小数据集迭代 70 次的在线数据分析基础上，对数据的分析敏感度超过 91%。证实该数据审计过程可以作为电网电力运行相关大数据审计的重要参考技术过程。

参考文献

[1] 郑亚会,郑文林.如何查询D5000系统无功数据和线路重过载[J].农村电工,2021,29(05):38.

[2] 芦博,袁富佳,赵升月,王丹丹.基于大数据架构的综合能源监控系统平台技术研究[J].供用电,2021,38(05):64-69.

[3] 李俊,张葵葵.基于大数据技术的电力安全管理实践策略探讨——评《电力安全生产大数据分析与应用》[J].中国安全生产科学技术,2021,17(04):189.

[4] 王能胜,吴玉玲,江涛,王黄磊.融合关联与矩阵算法的变电设备状态监测方法[J].计算机测量与控制,2021,29(04):30-34.

[5] 刘丽,梁大鹏,李海滨,李顺昕,申惠琪,聂文海.应用大数据分析的电网业务需求规划管理平台[J].信息技术,2021(04):141-147.

[6] 郭文东,马奎,车靖阳.大数据分析的电力设备运行安全性综合评估[J].信息技术,2021(04):159-163+169.

[7] 桑瑞.大数据时代电力企业内部审计信息化[J].现代商贸工业,2020,41(26):125-126.

[8] 蔡玺,李兴,祝唯微,权朝阳,杨翠.基于大数据的电力营销数字化审计应用研究[J].电子世界,2020(14):75-76.

[9] 田桂申,白雪娇.电力企业内部审计应对新冠疫情的相关思考[J].中国内部审计,2020(07):7-9.

[10] 吴国萍.浅谈电力企业内部审计技术方法创新[J].广西电业,2020(05):37-40.

[11] 段超,黄敏杰.电力数据在审计开发商偷逃土地增值税中的应用[J].中国内部审计,2020(05):47-49.

[12] 黄松,尚颖,马薇,吴婷婷.大数据挖掘技术在电力审计风险防范中的应用研究[J].中国内部审计,2020(05):32-39.

[13] 程凡.大数据时代电力企业内部审计信息化探讨[J].商讯,2020(06):98+100.

[14] 毛胜,叶琴.试析大数据环境下电力企业的数据式审计模式[J].现代经济信息,2020(03):103+105.

[15] 陈冬梅,厉雨.新时代电网企业全业务大数据审计研究[J].中国内部审计,2019(12):26-32.

[16] 孙刚,吕娜.大数据时代电力企业内部审计信息化问题窥探[J].价值工程,2019,38(34):52-53.