

基于混沌关联维特征的电能表计量多维数据聚类方法

窦圣霞,程志强

(国网宁夏电力有限公司 营销服务中心(计量中心),银川 750001)

Multi-dimensional data clustering method for electric energy meter measurement based on chaotic correlation dimension characteristics

DOU Shengxia, CHENG Zhiqiang

(Marketing Service Center(Metrology Center), State Grid Ningxia Electric Power Co., Ltd., Yinchuan 750001, China)

摘要:针对传统电能计量多维数据的聚类分析算法存在的计算过程复杂、效率低下、系统耗能高等问题,设计出一套适用的解决方案。该方案基于大数据平台进行计算存储,通过混沌关联维聚类分析方法对电能计量数据进行多维分析。该方法吸收了传统大数据聚类算法和混沌特征提取的优点,通过选取交互信息量的第一个最小值作为最佳时间延迟,采用虚假最近邻点算法选择最佳的嵌入维数 u 重构相空间。然后基于相空间重构提取出混沌关联维特征,并融合聚类算法对电能计量多维数据进行聚类。实验证明,本研究的混沌关联维聚类分析法效率高、过程简单且能耗低。

关键词:电能计量;多维数据;混沌特征提取;关联维;聚类算法

Abstract: Aiming at the problems of complex calculation process, low efficiency and high energy consumption of the traditional energy metering multi-dimensional data clustering analysis algorithm, a set of suitable solutions is designed. The solution is based on a big data platform for calculation and storage, and multi-dimensional analysis of electric energy measurement data is carried out through the chaotic correlation dimension clustering analysis method. This method absorbs the advantages of traditional big data clustering algorithms and chaotic feature extraction, selects the first minimum value of the interactive information as the best time delay and uses the false nearest neighbor algorithm to select the best embedding dimension u reconstruction phase space. Then, based on the phase space reconstruction, the chaotic correlation dimension characteristics are extracted and the clustering algorithm is combined to cluster the multi-dimensional data of electric energy measurement. Experiments prove that the chaotic correlation dimension clustering analysis method in this study is highly efficient, simple in process and low in energy consumption.

Key words: electric energy measurement; multidimensional data; chaotic feature extraction; correlation dimension; clustering algorithm

0 引言

在大数据智能电网环境下,电能计量装置相关技术处于不断创新阶段,使得电能计量多维数据采集量急剧上涨。然而,针对海量电能计量多维数据的有效管理、计算机系统性能优化、计量装置易发生故障等成为了新的挑战^[1]。因此,研究出有效的电能计量多维数据分析技术成为亟待解决的问题。

为了解决上述存在的问题,文献[2]公开了基于Hadoop模型的MapReduce算法配合Hive数据仓库储存技术的计算分析系统,这种方案虽然利用大数据技术对计量数据进行了深彻的挖掘分析,但该算法计算效率低下,Hive数据仓库储存技术复杂的读写信息会转换成MapReduce的计算过程,降低了分

析效率。文献[3]采用Hbase数据库分布式存储数据,虽然提高了分析效率,但依然存在MapReduce计算的缺陷。文献[4]采用基于LaVIEW的大数据分析聚类算法,应用位置标记方法,通过循环分批读取,解决了大数据块文本数据的快速聚类难题。这种方法虽然聚类过程简单迅速,但是需要建立一个数据模型专门进行管理,比较复杂。文献[5]公开了基于分布式技术的多维数据分析方法,利用云计算在数据存储、数据管理和虚拟化等方面的技术优势,构建了基于云计算的大数据管理和处理模式,对多维数据进行分析。该方法虽然具有较好效果,但处理效率仍较为低下。针对上述存在的缺陷,本文采用一种合适的解决方案,具体内容如下。

1 总体方案设计

本研究基于ADE7953电能计量芯片配合微处

收稿日期:2021-12-08;修回日期:2022-01-10

基金项目:国家电网有限公司科技项目(5229YX210006)

理器,实现电能计量多维数据的采集,并通过优化电能计量算法提高采集系统的精准度。本文技术亮点在于采用混沌关联维聚类分析法对电能计量数据进行多维分析。通过将混沌特征提取与大数据聚类算法巧妙结合在一起,充分发挥双方优势,高效且精确地分析电能计量多维数据。分析系统架构如图1所示。

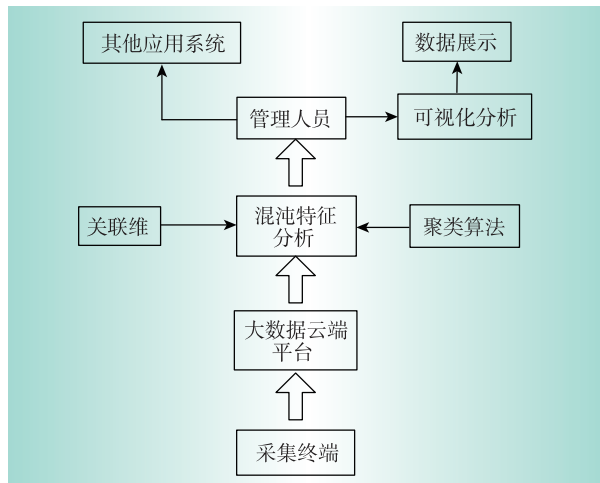


图1 电能计量多维数据分析系统架构图

Fig.1 Architecture diagram of energy metering multi-dimensional data analysis system

如图1所示,采集终端从各个新型智能电能表中收集电能计量数据,并通过优化电能计量算法提高采集系统的精准度。对电能计量数据进行预处理,通过新型智能电能表的通信技术可以实时远程传输数字信号至大数据云端平台。大数据平台应用计算机处理相关技术,对电能计量多维数据进行高效率的处理^[6]。接着基于关联维和聚类算法的混沌特征分析法对电能计量多维数据进行多维分析。最后将分析结果传输至管理人员,由管理人员针对分析结果进行可视化分析,并通过数据挖掘出隐藏信息,以便进行进一步的决策与判断,还可以将其应用到其他系统中。

2 混沌关联维聚类分析法

本研究采用混沌关联维聚类分析法,该方法将混沌特征提取与大数据聚类算法糅合,充分进行优势互补,合理地分析电能计量多维数据。混沌理论可以用公式进行表达如下。

存在一个函数 $f(x)$ 和一个紧性度量空间 M ,其中 $x \in M$ 。在 x 中有一个领域 $c(c > 0)$,领域 c 中存在任意值 y ,使得

$$F(f^n(x), f^n(y)) > z \quad (1)$$

式中: $n > 0$; z 为初始值敏感性, $z > 0$ 。在度量空间 M 上存在任意两个开集 A, B ,使得

$$f^k(A) \cap B \neq \phi \quad (2)$$

式中: $k > 0$ 。从式(2)得出函数 f 的值在度量空间 M 中密集,有 $f(x): M \rightarrow M$,定义 f 为度量空间 M 上的混沌^[7-9]。

在电能计量多维数据中,混沌特征一般表现为无规则非周期性,非常复杂,因此可以利用混沌特征参数来描述数据非周期性无规则特性,例如 Lyapunov 指数、关联维数、近似熵、复杂度等参数。其中 Lyapunov 指数和关联维数能直接反映数据信号的周期性或混沌性,是普遍使用的混沌特征参数^[10-11]。本研究采用混沌特征参数关联维数来反映电能计量多维数据的混沌程度,下面将阐述具体过程。

2.1 相空间的重构

电能计量多维数据序列一般是无规则非周期性序列,相空间重构是无规则非周期性序列的关键。相空间重构的原理是在一个时间延迟点上进行测量,并构成多相空间中的一个点,重复该过程构成若干个点,即重构出多相空间。在多相空间中能够提取电能计量多维数据的混沌关联维特征^[12]。关于相空间重构具体步骤如下。

首先假设原始时间序列为 $\{a_1, a_2, \dots, a_n\}$,则相空间重构矩阵 A 为

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_k \\ a_{1+r} & a_{2+r} & \cdots & a_{k+r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1+(m-1)r} & a_{2+(m-1)r} & \cdots & a_{k+(m-1)r} \end{bmatrix} \quad (3)$$

式中: $k = n - (u-1)r$; r 为时间延迟; u 为最佳嵌入维数。

从式(3)可以看出,相空间重构矩阵中 r 和 u 是两个重要参数,关于参数 r 和 u 的取值是后续步骤中重点阐述的内容^[13]。

接着关于时间延迟 r 的最佳选择,主要方法有线性自相关函数法和平均互信息法。本研究采用平均交互信息法,通过原始时间序列定义出平均交互信息公式,从中选择第一个相关极小值的时间序列间隔,作为最佳时间延迟 r 。关于延迟时间中交互信息量 O 公式为

$$O(r) = - \sum_{mn} P_{mn}(r) \ln \frac{P_{mn}(r)}{P_m P_n} \quad (4)$$

式中: m 和 n 均为某区间; mn 为区间 m 和区间 n 共有的分布; P 为分布概率。若 $O(r)$ 为 0,则 $a(t+r)$ 和 $a(t)$ 相互独立导致无法预测, $O(r)$ 越小, $a(t+r)$ 和 $a(r)$ 相互独立越明显。最终, $O(r)$ 取极小值,此刻 r 是最

佳时间延迟。

最后关于嵌入维数 u 的最佳选择,主要方法有虚假邻点法和奇异值分解法。本研究选用虚假邻点法,通过对随机两个邻近的时间序列点进行维数增加,比较相邻距离的增大量,以此来确定是否互为虚假邻点。假设在 u 维度空间,两个邻近的时间序列点处于时间序列

$$A(n) = \{a(n), a(n+r), \dots, a[n+(u-1)r]\} \quad (5)$$

邻近点 $a(n)$ 与 $a(n+r)$ 之间的距离为 B_1 ; 当维数增加到 $u+1$ 时,距离为 B_2 。如果 B_2 远远大于 B_1 , 则说明 $a(n)$ 与 $a(n+r)$ 互为虚假邻点。

然而在时间序列太小的情况下,不易判定虚假邻点,这种情况下可采用相对度量法。实现公式如

$$R \leq \frac{\|a_{n+r} - a_n\|^{u+1} - \|a_{n+r} - a_n\|^u}{\|a_{n+r} - a_n\|^u} \quad (6)$$

若式(6)成立,则说明 $a(n)$ 与 $a(n+r)$ 互为虚假邻点。一般情况下 R 值为 15,通过不断重复计算直到虚假邻点不再随着维数 u 变化而出现,此刻维数 u 是最佳嵌入维数。

2.2 基于关联维的混沌特征提取

在重构相空间后,进行下一步提取电能计量多维数据的混沌特征,关于混沌特征参量选用关联维数。关联维数主要是通过关联积分计算电能计量多维数据的混沌程度,它的原理在于计算原始时间序列中给定两个中心点之间的距离,并给出一个距离标准值。计算任意一点与中心点之间的距离,超过这个标准值的其他任意一点都与中心点不具有相关性;小于标准值,且大于中心点之间的距离的点与中心点具有相关性。这些相关的点数目越多,则代表多项空间内关联维数越多。下面将通过算法实现该过程。

在多相空间的重构过程后,关于原始时间序列矩阵 A_m 可表示为

$$A_m = (a_m, a_{m+r}, \dots, a_{m+(u-1)r})^T \quad (7)$$

设立两个中心点 a_m 和 a_n , 在关联维数为 u 的相空间内计算 a_m 与 a_n 之间的距离,不超过 L 的点集集合 Q 为

$$Q = \sum_{m \neq n} H(L_i - \|a_m - a_n\|) \quad (8)$$

式中: H 为赫维赛德函数; L_i 为相空间内任意一点 i 到中心点的距离,大于两个中心点之间的距离。关于关联函数 $D(L)$ 的表达式为

$$D(L) = \frac{2}{Q(Q-1)} \sum_{m=1}^i \sum_{n=m+1}^i H(L - \|a_m - a_n\|) \quad (9)$$

式(9)中含有集合 Q 的分数主要作用是对关联

函数的去重。下面计算给定的距离标准值为

$$|a_{m-(v-1)r} - a_{n-(v-1)r}| \quad 1 \leq v \leq u \quad (10)$$

式中: v 为符合要求的关联维数,通过这个关联维数范围参数表示出电能计量多维数据的混沌程度,为下一步大数据聚类分析算法提供条件。

2.3 聚类分析算法的实现

在提取出电能计量多维数据混沌特征后,需要实现聚类分析算法。聚类分析算法指将电能计量多维数据样本进行分类,同一个类簇中的样本具有相似性,不同类簇中的样本之间具有差异性。下面将根据混沌特征进行聚类分析,具体算法流程如下。

(1) 输入 n 个电能计量多维数据样本集 $\{x_1, x_2, \dots, x_n\}$, 从该集合中随机选取 N 个凝聚点,这些凝聚点组成的集合为 $\{z_1, z_2, \dots, z_N\}$ 。

(2) 计算任意电能计量多维数据样本点与凝聚点之间的距离

$$\|x - z_n\| = \min(\|x - z_n\|) \quad (11)$$

式(11)用于描述 x 和 z_n 之间的最短距离。同时假设电能计量多维数据样本集合 w_n 中存在 N_n 个电能计量多维数据样本。

(3) 根据关联函数 $D(L)$ 求和得到凝聚点 z_n 的表达式为

$$z_n = \frac{1}{N} \sum_{n, x \in w_n} x D(L) \quad (12)$$

(4) 分类。假设 x 和 z_n 之间的最大距离为

$$Y = \max(\|x - z_n\|) \quad x \in w_n \quad (13)$$

用 d_1 代表分类距离,如果 x 和 z_n 之间的最大距离大于分类距离,说明 w_m 点并不适合作为凝聚点,要在此聚类区间分成两个凝聚点,可描述为

$$\begin{cases} z_{m1} = z_m + \lambda Y \\ z_{m2} = z_m - \lambda Y \end{cases} \quad (14)$$

式中: λ 为一个大于 0 的常数。若 x 和 z_n 之间的最大距离小于分类距离,说明 w_m 点适合作为凝聚点,能够很好的进行聚类,则不需要进行步骤(5),直接迭代结束。

(5) 合并。假设 z_m 和 z_n 之间的最小距离为 l , 关于 l 的表达式为

$$l = \min(\|z_m - z_n\|) \quad (15)$$

用 d_2 表示合并距离,若最小距离 l 小于合并距离,说明分类得到的 w_m 和 w_n 之间具有相似性,将其进行合并,得到 z_m 点和 z_n 点的凝聚点 z_{mn} 为

$$z_{mn} = \frac{N_m z_m + N_n z_n}{N_m + N_n} \quad (16)$$

若最小距离 l 大于合并距离,则重新进行步骤(4)。综上所述,基于关联维数混沌特征参量,本研

究应用了电能计量多维数据的聚类,将相似的混沌特征样本分成了同一类,成功实现了混沌特征提取与大数据聚类分析法的结合。

3 实验与分析

下面将通过仿真实验进行分析,通过对比验证本文所研究的混沌关联维聚类分析法的优势。

电能计量数据样本信息主要通过电能计量数据库提取,使用大数据云端平台进行数据的预处理,最终得到电能计量多维数据。由于数据库中电能参数种类过多,因此选择其中几个重要参数列出,如表1所示。

表1 电能计量多维数据样本参数

Table 1 Parameters of electric energy metering multi-dimensional data sample

电能参数	平均值
电压/V	236.93
电流/mA	142
有功功率/W	14.82
累计电量/kWh	0.83
频率/Hz	50

本研究分别采用 LaVIEW 算法和混沌关联维聚类分析法(chaos correlation dimension cluster analysis method,CCDCAM),针对不同数据量的电能计量多维数据进行聚类,通过实验统计出 LaVIEW 算法和 CCDCAM 算法计算所耗时间。将统计得到结果进行对比如表2所示。

表2 两种算法聚类耗时对比

Table 2 Time-consuming comparison of two algorithms for clustering

数据量/GB	耗时/s	
	LaVIEW 算法	CCDCAM 算法
20	137	108
200	5 545	1 956
400	12 875	4 465
800	28 583	9 985
1 024	33 737	10 247

从表2可以看出,在不同电能计量多维数据量的条件下,本文所研究的混沌关联维聚类分析法所耗时间更短。在数据量为 1 024 GB 的环境下,本研究的 CCDCAM 算法耗时少 2 倍多,这也直接表明了本研究的混沌关联维聚类分析法更加高效。

在耗时对比实验后,本文又分别采用这两种算法处理相同的数据量统计系统所损失的能量,并进一步统计两种算法的关联维数,MATLAB 输入统计结果得出仿真曲线如图2和图3所示。

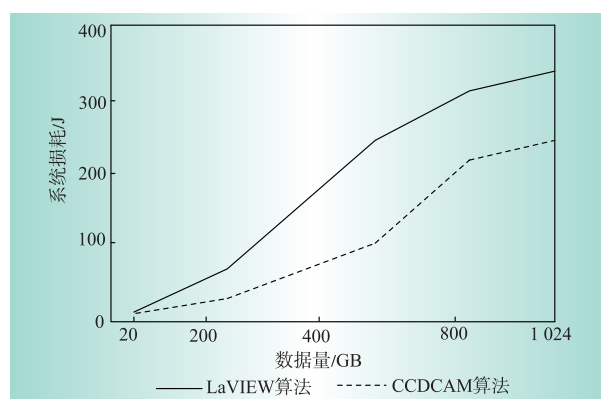


图2 两种算法聚类能耗对比

Fig. 2 Energy consumption comparison of two algorithms for clustering

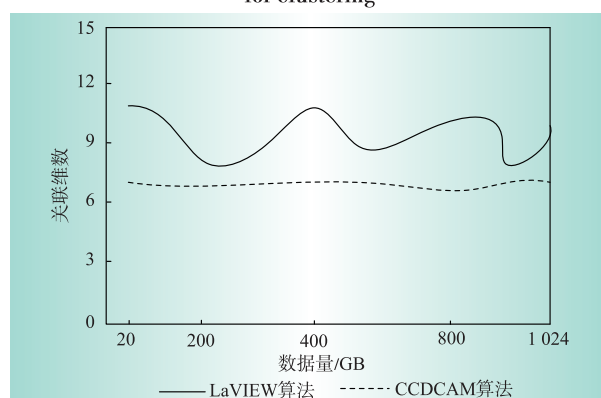


图3 两种算法聚类关联维数变化对比

Fig. 3 Comparison of correlation dimension changes of two algorithms for clustering

从图2可以得出,在不同电能计量数据量的环境下,CCDCAM 算法曲线均在 LaVIEW 算法曲线之下,这体现了本研究的算法系统损耗更低。如图3所示,本研究所采用的混沌关联维聚类分析法关联维数更加稳定。以此证明了本研究算法稳定和性能要求低的优点。

为了验证 CCDCAM 算法比 MapReduce 算法直接计算处理效果更好,本文通过对比来表现这两种方法处理数据的差异。根据表1数据样本,实验分别统计对同类电能计量数据不同数据量分析处理的精准度,得到对比结果如图4所示。

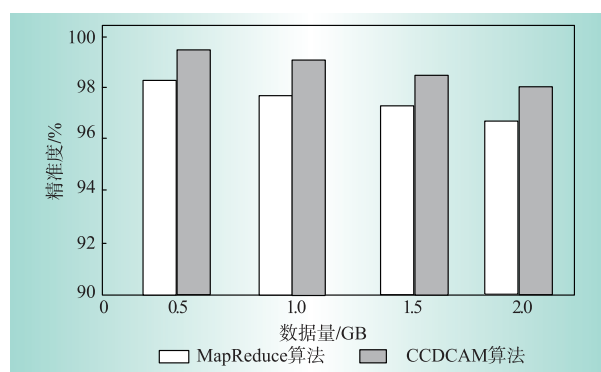


图4 不同方法精准度对比图

Fig. 4 Comparison of the accuracy of different methods

本文所采用的 CCDCAM 算法比传统 MapReduce 算法处理数据精准度要更高,随着电能计量数据量不断增加,处理电能计量数据的精准度也会略微降低,但是相同数据量情况下 CCDCAM 算法比使用 MapReduce 算法优势大。因此,本文所采用的混沌关联维聚类分析法具有更良好的实用性,适合于电能计量大数据的聚合应用。

4 结束语

本研究首先通过在一个时间延迟点上进行测量,并构成多相空间中的一个点,重复该过程构成若干个相点,重构出多相空间。其次在重构相空间后,进行下一步提取电能计量多维数据的混沌特征,通过关联维表现混沌程度。最后将电能计量多维数据样本进行分类,基于混沌特征将相似的样本分为一类,成功实现混沌特征提取与大数据聚类分析法的结合。实验表明,相比传统的聚类分析算法,本研究所采用的分析算法适用性较高,能有效地提高电能计量多维数据聚类分析效率。D

参考文献:

- [1] 罗忠波. 电能计量自动化系统在智能节能服务中的应用[J]. 大科技, 2017(15): 94-95.
LUO Zhongbo. The application of electric energy metering automation system in intelligent energy-saving services[J]. Super Science, 2017(15): 94-95.
- [2] 蒋佳军. 基于 Hadoop 平台的电力大客户用电时序数据分类研究[D]. 广州: 华南理工大学, 2015.
JIANG Jiajun. Research on time series data classification of electric power customers based on Hadoop platform[D]. Guangzhou: South China University of Technology, 2015.
- [3] 董宇, 李欣宇. 电力谐波对电能计量影响的分析[J]. 科技资讯, 2015, 13(35): 120.
DONG Yu, LI Xinyu. Analysis of the influence of power harmonics on electric energy measurement[J]. Science & Technology Information, 2015, 13(35): 120.
- [4] 姜延文. 大数据分析下多维离散数据高效聚类方法仿真[J]. 计算机仿真, 2019, 36(2): 205-208.
JIANG Yanwen. Simulation of efficient clustering method for multi-dimensional discrete data under big data analysis[J]. Computer Integrated Manufacturing Systems, 2019, 36(2): 205-208.
- [5] 娄乾. 用电计量装置在线监测与诊断系统的研究与开发[D]. 北京: 华北电力大学, 2015.
LOU Qian. Research and development of online monitoring and diagnosis system for electricity consumption metering device [D]. Beijing: North China Electric Power University, 2015.
- [6] 朱力. 电能计量装置管理系统的设计与实现[D]. 北京: 华北电力大学, 2017.
ZHU Li. Design and implementation of management system for electric energy metering devices [D]. Beijing: North China Electric Power University, 2017.
- [7] HERMASSI H, HAMDI M, RHOUMA R, et al. A joint encryption-compression codec for speech signals using the ITU-T G. 711 standard and chaotic map [J]. Multimedia Tools & Applications, 2017, 76(1): 1 177-1 200.
- [8] 李海东. 非线性负荷计量系统研究[D]. 济南: 山东大学, 2017.
LI Haidong. Research on nonlinear load measurement system [D]. Ji'nan: Shandong University, 2017.
- [9] KORDOV K. A novel audio encryption algorithm with permutation-substitution architecture [J]. Electronics, 2019, 8(5): 530.
- [10] HAMDI M, RHOUMA R, BELGHITH S. An appropriate system for securing real-time voice communication based on ADPCM coding and chaotic maps [J]. Multimedia Tools & Applications, 2017, 76(5): 7 105-7 128.
- [11] 黄腾飞, 李帮义, 熊季霞. 基于小波方差分解的混沌时间序列噪声估计和阈值去噪[J]. 计算机应用, 2013, 33(3): 890-895.
HUANG Tengfei, LI Bangyi, XIONG Jixia. Noise estimation and threshold denoising of chaotic time series based on wavelet variance decomposition [J]. Journal of Computer Applications, 2013, 33(3): 890-895.
- [12] 行鸿彦, 刘刚. 混沌系统相空间重构参数选取的仿真研究[J]. 计算机仿真, 2018, 35(7): 212-216.
XING Hongyan, LIU Gang. Simulation research on phase space reconstruction parameter selection of chaotic system [J]. Computer Integrated Manufacturing Systems, 2018, 35(7): 212-216.
- [13] 沈力华, 陈吉红, 曾志刚, 等. 基于鲁棒极端学习机的混沌时间序列建模预测[J]. 物理学报, 2018, 67(3): 32-42.
SHEN Lihua, CHEN Jihong, ZENG Zhigang, et al. Chaotic time series modeling and prediction based on robust extreme learning machine [J]. Acta Physica Sinica, 2018, 67(3): 32-42.

作者简介:

窦圣霞(1974), 女, 山东东平人, 学士, 高级工程师, 主要研究方向为电能计量技术及分析管理;

程志强(1973), 男, 河北任丘人, 学士, 高级工程师, 主要研究方向为电能计量技术及分析管理。

(责任编辑 史倩芸)