

DOI:10.19344/j.cnki.issn1671-5276.2025.05.011

密度偏差抽样算法的设备样本点选择

陈双, 伍铁军

(南京航空航天大学 机电学院, 江苏 南京 210016)

摘 要:针对数据驱动设备故障诊断进行直接计算需要消耗大量时间和空间资源的问题,提出基于可变网格和 FCM 密度偏差抽样的设备典型样本点选择方法。按可变网格划分的方式将数据集划分为若干个大小不同的网格,根据各网格单元内数据点的抽样概率计算出每个网格内的期望样本数,在网格内部进行 FCM 聚类得到目标样本点个数。实验结果表明:该算法不仅能够保证原始数据集的完整性,还能保持在样本数据集上有相对精确的聚类效果,为后续设备运行工况的识别奠定了基础。

关键词:密度偏差抽样;模糊 C 均值;可变网格;设备典型样本点

中图分类号:TP391.4 **文献标志码:**B **文章编号:**1671-5276(2025)05-0057-05

Equipment Sample Point Selection for Density Biasd Sampling Algorithm

CHEN Shuang, WU Tiejun

(College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: A device typical sample point selection method based on variable grid and FCM density bias sampling is proposed to address the problems of enormous consumption of time and space resources in data-driven device fault diagnosis by direct calculation. The dataset is divided into several grids of different sizes by a variable grid partitioning method, the expected number of samples in each grid are calculated based on the sampling probability of data points within each grid unit, and FCM clustering is performed inside the grid to obtain the number of target sample points. The experimental results show that the proposed algorithm not only ensures the integrity of the original dataset, but also maintains relatively accurate clustering effect, which lays foundation for the subsequent recognition of equipment operating conditions.

Keywords: density biased sampling; fuzzy C-means; variable grid; typical sample points of equipment

0 引言

面对海量和高维的设备运行数据,如果直接聚类分析来获取其运行状态,会使得运行的效率与聚类的质量均无法得到较好的保证,且会有运行时间过长,甚至内存溢出导致无法计算的问题^[1]。由于传感器数据采集误差和设备负荷切换等多种因素^[2],使得历史数据中元素并非全部有用,这种数据不适宜于设备典型工况的建模。

随机抽样(simple random sampling, SRS)是一种常用的数据规模约简方法,它以相同的概率对每个数据进行抽取,具有非常高效的执行效率。然而,该抽样方法存在较强的随机性,其稳定性相对较差,在数据集分布偏差较大的情况下, SRS 方法可能会丢失一些小类别^[3]。为了解决这个问题,提出了基于密度的抽样方法即密度偏差抽样(density biased sampling, DBS)^[4]。该方法让每个原始数据点被抽样的概率取决于所在区域的局部

密度,并根据数据分布情况来生成样本。在设备工况挖掘聚类分析中, DBS 方法能够有效保证样本数据集的完整性和倾斜性,避免原始数据集类别的丢失。因此,该算法能够覆盖设备在正常工况下的全部数据分布情况,显著提升样本选择的质量,适用于海量工业设备数据的分析。

现有的 DBS 算法有基于固定网格划分的 DBS 算法,如 G_DBIS 算法^[5]。但在高维空间中,此方法面对分布不均匀的数据集会产生大量的空网格,这些无用的计算导致其可用性大大降低,且数据集的固定划分往往会忽略数据在各个维度上的分布,导致算法的执行性能不佳。不同于固定网格划分方法,基于可变网格划分的 DBS 算法即 VG_DBIS 算法^[6],其根据相邻数据区间段的密度相似性来合并网格,可以减少网格划分的数量,得到的网格也更能代表原始数据的分布特征^[7]。但是由于合并后的网格单元变大,在网格内部继续使用随机抽样会导致样本选择的典型性不高。

第一作者简介:陈双(1998—),男,江苏淮安人,硕士研究生,研究方向为机械设备状态监测, chenshuangxyz@163.com。

针对以上问题,本文提出了基于可变网格和FCM^[8]密度偏差抽样算法的设备样本点选择方法。该方法先用可变网格划分将原始数据集划分成大小不同的网格,接着在网格内部执行FCM聚类算法来实现抽样,能够有效得到全工况样本数据集,显著提高后续设备典型工况挖掘的聚类精度。

1 基本抽样方法

1.1 模糊C均值聚类算法

模糊C均值(fuzzy C-means, FCM)聚类算法是一种简单的迭代型无监督分类算法,即将给定的数据集分成 c 个类。同时,FCM也可以看作是一种典型样本点提取的方法,其中类心被视为最具代表性的点。虽然能保证点具有较高的代表性,但也导致抽样很难提取到密度小区域的数据点。然而,对设备运行数据来说,样本空间中密度小的区域并非一定是噪声区域,有可能是出现比较少的正常运行工况或者是运行状态的过渡区域。这种情况下所选的样本点用于状态监测时,边缘点将触发报警。而边缘点也是正常工况下的样本点,不应引起预警。此外,FCM算法的类簇数 c 需要人为指定,但设备的真实历史数据集中没有反映类分布情况的先验知识。

1.2 可变网格划分方法

首先利用快速排序法将原始数据集的每个维度等深划分成 k 个区间段,然后对密度相似的相邻区间段执行合并操作,最后对数据集的每一维均执行以上操作^[9]。这样构建的网格空间中网格数量减少致使计算量降低,消耗的总时间也变短,适用于大规模数据集的抽样。

但上述过程只考虑了相邻区间的密度分布,这样可能会出现数据两端的密度分布差别较大时,整维数据合并成一个区间的情况^[10],说明合并区间还应该考虑数据的整体特征。可以在合并区间时进行改进。

初步合并密度相似的相邻区间,得到新的区间长度 $|\bar{I}_{ij}|$,用式(1)重新计算相邻区间段 $|\bar{I}_{ij}|$ 和 $|I_{i(j+1)}|$ 的密度相似性值 $\bar{\varepsilon}_{ri}$ 。

$$\bar{\varepsilon}_{ri} = \begin{cases} \frac{1}{h} |\bar{I}_{ij}| / |I_{i(j+1)}|, & \frac{1}{h} |\bar{I}_{ij}| \leq |I_{i(j+1)}| \\ |I_{i(j+1)}| / \left(\frac{1}{h} |\bar{I}_{ij}| \right), & \frac{1}{h} |\bar{I}_{ij}| > |I_{i(j+1)}| \end{cases} \quad (1)$$

式中 h 为初步合并的区间个数。若此时的密度相似性值大于设定阈值,正式合并相邻区间,并计算

相应的区间密度和密度相似性,反之则不执行合并操作。如此重复下去,直至比较完该维所有的区间段。对每个维度都执行类似操作,就可以得到利用可变网格划分的数据网格空间。

2 改进密度偏差抽样算法

通过统计每个网格内所含数据点的数量,来衡量利用可变网格划分构建网格空间的密度。不同网格中数据点抽样概率的偏差取决于各网格内样本数的偏差。

设第 m 个网格 G_m 的密度为 n_m ,同一个网格内数据点被抽样的概率相等,则该单元格各样本点的抽样概率 $f_m(n_m)$ 为

$$f_m(n_m) = \frac{a}{n_m^e}, \quad e \in [0, 1] \quad (2)$$

式中; $m=1, 2, \dots, G$, G 为网格总数; e 为常量,位于0~1之间,来控制不同网格内的偏差抽样概率。当 $e=0$ 时,各网格单元的抽样概率函数相同;当 $e=1$ 时,将在每个网格中提取数量相同的数据点。

给定的样本量 n 是每个网格单元中所抽取样本的总和,即

$$n = \sum_{m=1}^G n_m f_m(n_m) \quad (3)$$

将式(2)代入式(3)中,可得

$$a = \frac{n}{\sum_{m=1}^G n_m^{1-e}} \quad (4)$$

再将式(4)计算获得的 a 代入式(2)中,可得各网格单元内数据点的抽样概率。

$$f_m(n_m) = \frac{n}{n_m^e \sum_{m=1}^G n_m^{1-e}} \quad (5)$$

式(5)是一个减函数,如果网格中的数据增加,则抽样的概率就相应减小。通过使用可变网格DBS的抽样方法,可以实现对数据集的密度偏差抽样。该方法在抽样时考虑了数据点所处区域的密度情况,将密度集中区域的数据抽取概率降低,而稀疏区域的数据抽取概率增加。

然而,在网格内部随机抽样,可能会出现覆盖不完整的样本集。为此,可以在获得各网格抽样概率后,进一步计算每个网格内预期的样本数量 $\tilde{c}$ 。

$$\tilde{c} = n_m \cdot f_m(n_m) = n_m \frac{a}{n_m^e} = \frac{n_m n}{n_m^e \sum_{m=1}^G n_m^{1-e}} = n \frac{n_m^{1-e}}{\sum_{m=1}^G n_m^{1-e}} \quad (6)$$

通过在每个网格内进行 FCM 聚类,并选择距离聚类中心最近的数据作为样本,可以避免 FCM 算法中参数 c 的指定,显著降低计算量。结合 FCM 算法还可以克服基于可变网格 DBS 算法的随机性,使选取的样本点有较高的代表性。与原 FCM 算法的抽样结果相比,能够获取到稀疏区域的样本点,让覆盖的区域更加全面,更加符合基于设备历史运行数据抽样的要求。

基于可变网格和 FCM 的密度偏差抽样算法的步骤如图 1 所示。

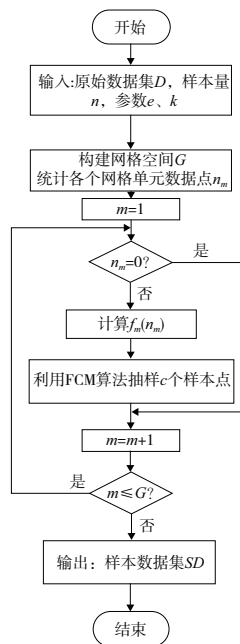


图1 可变网格和 FCM 密度偏差抽样算法流程

3 实验论证

为验证本文所提的抽样算法在设备典型样本点选择方面的可行性,实验采用 MATLAB2018a 分别对本文算法、G_DBS 算法和 SRS 算法在样本质量和执行时间方面进行对比分析。实验选用 5 组标准的 UCI 数据集,详细信息如表 1 所示。

表1 实验数据集

数据集	实例数	维数	类簇数	最大类数据量	最小类数据量
Page-blocks	5 473	10	5	4 913	28
Pendigits	10 992	16	10	1 144	1 055
Shuttle	58 000	9	7	5 581	1 017
LDPR	164 860	5	11	45 587	1 381
Census-Income	299 286	13	2	280 717	18 569

3.1 参数选择

1) 参数值 k 的设置。由于处理的是设备大规模分布不均匀的数据集,可以假设网格总数与数据量呈线性关系,即 k^p 个网格内包含 N 条记录,因此 k 的值可表示为

$$k = \lceil \sqrt[p]{N} \rceil \quad (7)$$

式中:半开 $\lceil \rceil$ 符号表示向上取整; p 表示数据集的维度; N 表示样本个数。得到的网格单元总数将不大于由固定网格划分得到的 k^p 个网格数。

2) 参数值 e 的设置。 e 是一个控制全局密度偏差抽样的参数,可以根据每一维上密度相似性的偏差值来衡量该维数据的偏差程度,再用整体的密度相异性平均值作为 e 的值,如式(8)所示。

$$e = \frac{1}{p} \sum_{i=1}^p \left[\frac{1}{g-1} \sum_{r=1}^{g-1} (1 - \varepsilon_{ri}) \right] \quad (8)$$

式中: g 表示每个维度被等深划分的区间数; ε_{ri} 表示相邻区间段的密度相似性。

3.2 样本质量分析

样本质量包括样本完整性和样本正确性两个方面。样本完整性是指样本数据集所包含的聚类个数是否与原始数据集一致,而样本正确性则衡量了在样本数据集上进行聚类分析的结果是否准确。

1) 样本完整性对比分析

实验选择两个分布不均匀的数据集 Page-blocks 和 Shuttle。为保证参数 k 的值一致,在不同比例下对原样本进行抽样。记录抽样数据集所含的聚类个数,如表 2、表 3 所示。

表2 Page-blocks 数据集实验结果

比例/%	类别数	本文算法	G_DBS	SRS
0.5	5	4	4	4
1.0	5	5	5	4
5.0	5	5	5	4
10.0	5	5	5	5
15.0	5	5	5	5

表3 Shuttle 数据集实验结果

比例/%	类别数	本文算法	G_DBS	SRS
0.5	7	6	5	5
1.0	7	7	7	4
5.0	7	7	7	6
10.0	7	7	7	6
15.0	7	7	7	7

由表 2、表 3 可以看出,对于数据集 Shuttle 和

Page-blocks,本文算法或 G_DBs 算法在抽样比例为 1.0% 就可保证不丢失类,而 SRS 算法需要分别采用 10.0% 和 15.0% 的抽样比例才能保证样本完整性,相对来说 SRS 算法识别样本类别的能力较弱。

因此,相对于 SRS 算法,本文算法在较低抽样比例下就可以保证样本的完备。相对 G_DBs 算法,本文算法虽然减少了网格单元总数,但在抽样时并没有丢失类,依旧保证了原始数据集的分布特征,在数据分布不均匀且抽样比例较小时,识别类的能力较高,适用于复杂工业生产设备的数

2) 样本正确性对比分析

本组实验中采用密度峰值聚类 (density peaks clustering, DPC) 算法^[11] 对各样本数据集进行聚类。为确保实验的客观性,关于 DPC 算法的各项参数和用于抽样的 k 值同样保持一致,并用被正确聚类的数据点个数与样本量的比值来衡量聚类的正确性。对以上 5 个 UCI 数据集的实验结果分别进行对比分析,抽样比例依次为 10%、1%、5%、10% 和 0.1%。图 2 依次给出了各个算法在各自数据集上以相同的抽样比例执行 5 次的实验结果。表 4 给出了各算法实验正确率的平均值,其中加粗为同一组中相对较好的实验结果。

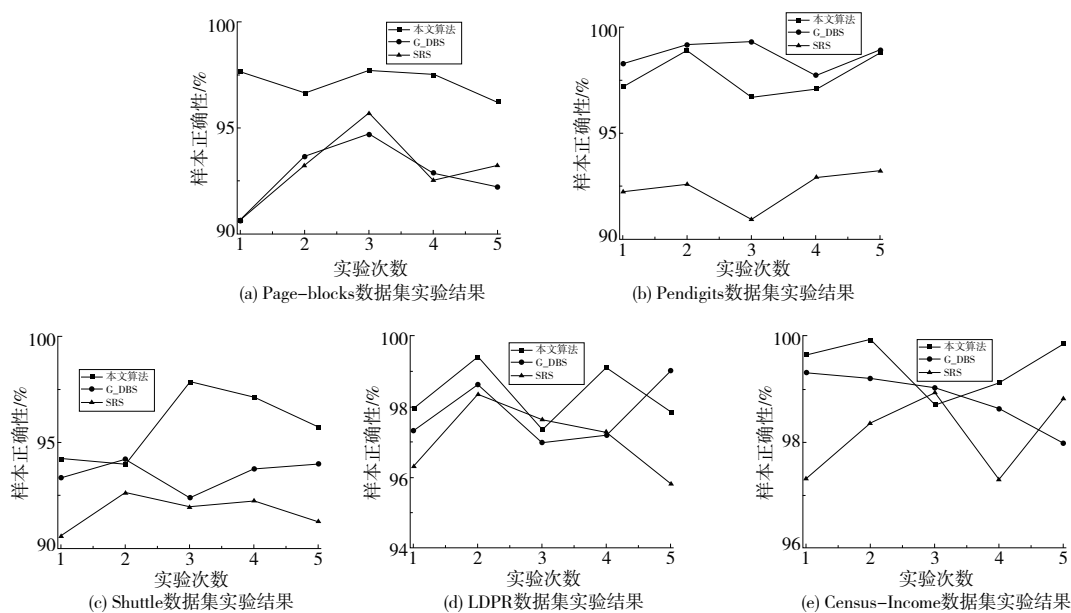


图 2 各算法在各数据集上的实验结果

表 4 样本正确性实验结果 单位: %

数据集	本文算法	G_DBs	SRS
Page-blocks	97.57	92.79	93.07
Pendigits	97.75	98.99	92.88
Shuttle	95.92	93.52	91.89
LDPR	98.93	97.74	97.42
Census-Income	99.54	98.98	98.33

由图 2 和表 4 可得出以下结论:对于 Census-Income、Page-blocks、LDPR 等分布不均匀的数据集,采用本文算法所得数据集的正确性要高于其它两种算法的结果。

对于分布较为均匀的数据集 Pendigits,本文算法聚类的正确率虽然高于 SRS 算法,但低于 G_DBs 算法所获的样本数据集,这表明本文算法

在均匀分布的数据集上执行抽样时,样本正确性会受一定影响。但实际上,设备的正常运行数据点往往集中在多维空间的某个特定区域,即设备的一个典型工况。该局部区域的密度较高,其他区域的密度较低,样本分布是不均匀的,所以本文算法适合于工业设备的实际数据抽样。

相较于 G_DBs 算法,本文算法在减少网格单元数目的前提下,仍有效保证了样本的正确性。在抽样比例相同时,本文算法相对于 SRS 算法既能保证样本数据集的完备,又能有较高的样本聚类正确率。这都说明了本文算法在工业设备的大量数据抽样中具有较大的实用性。

3.3 执行时间对比分析

继续按照 3.2 节介绍的实验方法,并分别记录抽样时间和样本数据集的聚类时间,最终各算法所对应的实验时间平均值如表 5 所示。

表5 各数据集的执行时间

单位:s

数据集	本文算法			G_DBs 算法			SRS 算法		
	抽样	聚类	总时间	抽样	聚类	总时间	抽样	聚类	总时间
Page-blocks	11.57	7.26	18.83	9.64	7.17	16.81	0.013	8.19	8.20
Pendigits	83.95	29.75	113.70	87.29	26.09	113.38	0.034	24.10	24.13
Shuttle	31.16	73.92	105.08	48.62	41.27	89.89	0.070	55.30	55.37
LDPR	79.33	561.30	640.35	767.88	566.03	1 333.90	0.086	671.00	671.09
Census-Income	228.60	885.30	1 113.90	1 007.50	953.34	1 360.80	0.097	1 006.00	1 006.10

SRS 算法的抽样时间无明显变化,始终保持较低的水平。本文算法的抽样时间随着原始数据集的规模和维度增大而增加,但其变化幅度相对于 G_DBs 算法较为稳定,适用于海量、高维数据的抽样。在聚类时间方面,本文算法、SRS 算法和 G_DBs 算法所对应的聚类时间均随着原始数据集规模的增大而增加,但相比较而言,本文算法在高维和数据量大时的聚类时间较短且上升趋势较为缓慢。在总时间方面,SRS 算法在样本较少时进行操作的总时间最短。随着样本数据量和维度变大,本文算法所消耗的总时间小于 G_DBs 算法和 SRS 算法且增加幅度相对平稳。同时考虑到样本完整性与正确性方面,本文算法相比 G_DBs 算法和 SRS 算法,在针对海量和高维分布不均匀的设备数据抽样中均具有一定优势,充分展现其可行性,满足实际设备运行数据建模的需要。

4 结语

本文提出了基于可变网格和 FCM 密度偏差抽样的设备典型样本点选择方法。根据实验结果,该算法在增强数据约简效果的同时能够确保样本集的完整性,可以深入到样本密度较低的区域,从而获取全面的工况数据集。此外,该方法在减少网格单元数目的前提下,依旧有效保证了样本抽样的质量和聚类的精度。尽管样本选择过程的时间消耗依旧较大,但降低了后续设备工况挖掘过程的内存开销,且设备典型工况的挖掘过程是非实时的,使本文算法在实际工程应用中可以接受。

参考文献:

[1] 潘春燕,张仁崇,杨忠保.密度偏差抽样在近邻传播聚

类中的应用[J].商丘师范学院学报,2022,38(12):4-8.

[2] 王林,李云伟,任重,等.基于数字孪生技术的电力设备不间断巡视系统设计[J].机械制造与自动化,2022,51(3):220-224.

[3] GU B, HU F, LIU H. Sampling and its application in data mining: a survey [C]. Singapore: National University of Singapore, 2000.

[4] PALMER C R, FALOUTSOS C. Density biased sampling: an improved method for data mining and clustering[C]//Proceedings of the 2000 ACM SIGMOD international conference on Management of data. Dallas Texas USA: ACM, 2000: 82-92.

[5] 张建锦,吴渝,刘小霞.一种改进的密度偏差抽样算法[J].计算机应用,2007,27(7):1695-1698.

[6] 盛开元,钱雪忠,吴秦.基于可变网格划分的密度偏差抽样算法[J].计算机应用,2013,33(9):2419-2422.

[7] 吕丹,龙华,高杰,等.基于不均匀数据的密度偏差抽样改进算法[J].软件导刊,2018,17(2):77-79,85.

[8] BEZDEK J C, EHRLICH R, FULL W. FCM: the fuzzy C-means clustering algorithm [J]. Computers & Geosciences, 1984, 10(2/3): 191-203.

[9] 金安.一种移动端 APM 系统框架设计与实现[J].软件导刊,2022,21(1):205-209.

[10] 潘春燕,吴有富,李方.一种基于可变网格划分的密度偏差抽样技术及其在聚类中的应用研究[J].凯里学院学报,2017,35(3):16-20.

[11] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492-1496.

收稿日期:2023-11-01