徐恕贞，司贺杰，罗君梅. 基于改进 K 均值聚类的超高维连续数据异常缺失检测算法[J]. 智能计算机与应用，2025，15(7)：144-148. DOI：10.20169/j.issn.2095-2163.24120903

基于改进 K 均值聚类的超高维连续数据异常缺失检测算法

徐恕贞，司贺杰，罗君梅
（郑州科技学院 信息工程学院，郑州 450064）

摘 要：超高维数据集是指在数据集中协变量的个数呈现出与样本量相关的指数增长态势。由于维度众多，不同维度数据之间存在着非线性的相互作用和依赖关系，传统 K 均值聚类初始聚类中心选择的随机性会导致聚类结果不稳定。在处理超高维连续数据时，随着数据规模和复杂度的增加，这种随机性会导致难以准确地识别出数据中的异常缺失值。本文提出一种基于改进 K 均值聚类的超高维连续数据异常缺失检测算法，采用降维处理技术，将超高维连续数据转换为低维连续数据，以降低计算复杂度并保留关键信息；在低维空间中对 K 均值聚类算法进行改进，优化初始聚类中心的选择方式；利用改进后的 K 均值聚类算法对低维连续数据进行聚类分析，通过计算各数据点与聚类中心的距离，准确识别并检测出数据中的异常缺失值。实验结果表明，本文提出的算法在超高维连续数据的异常缺失检测中展现出了较高的准确性，为超高维数据的异常检测提供了一种有效且高效的方法。

关键词：改进 K 均值聚类；超高维数据；连续数据；数据异常；缺失检测
中图分类号：TP391 **文献标志码**：A **文章编号**：2095-2163(2025)07-0144-05

Anomaly missing detection algorithm for ultra-high dimensional continuous data based on improved K-Means Clustering

XU Shuzhen, SI Hejie, LUO Junmei
(School of Information Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450064, China)

Abstract: Ultra-high-dimensional data set means that the number of covariables in the data set shows an exponential growth trend related to the sample size. Due to the large number of dimensions, there are nonlinear interactions and dependencies between data of different dimensions, and the randomness of initial cluster center selection in traditional K-Means Clustering will lead to unstable clustering results. When dealing with ultra-high dimensional continuous data, as the data size and complexity increase, this randomness can make it difficult to accurately identify abnormal missing values in the data. Therefore, this study proposes a high-dimensional continuous data anomaly detection algorithm based on improved K-Means Clustering. By using dimensionality reduction techniques, ultra-high dimensional continuous data is converted into low dimensional continuous data to reduce computational complexity and preserve key information. Improve the K-Means Clustering algorithm in low dimensional space and optimize the selection of initial clustering centers. Using the improved K-Means Clustering algorithm for clustering analysis of low dimensional continuous data, during the clustering process, the distance between each data point and the cluster center is calculated to accurately identify and detect abnormal missing values in the data. The experimental results show that the proposed algorithm exhibits good accuracy in anomaly detection of ultra-high dimensional continuous data, providing an effective and efficient method for anomaly detection of ultra-high dimensional data.

Key words: improve K-Means Clustering; ultra-high dimensional data; continuous data; data anomaly; missing detection;

0 引 言

在众多领域如气象学、金融分析、图像识别等，数据的维度不断增加，形成了大量的超高维数据集。超高维数据集的广泛存在，使得对其进行有效的分析和处理成为了一个亟待解决的问题。超高维数据集是指在数据集中，协变量的个数呈现出与样本量相关的指数增长态势，即随着样本量的增加，数据的

作者简介：徐恕贞(1990—)，女，硕士，助教，主要研究方向：数据挖掘。Email：xxx632144@163.com；司贺杰(1993—)，女，硕士，助教，主要研究方向：知识图谱，高等教育；罗君梅(1987—)，女，硕士，助教，主要研究方向：数据挖掘。

收稿日期：2024-12-09

维度会以极快的速度增长。与传统的低维数据相比，超高维数据使数据的结构变得极其复杂，数据的分布规律也更加难以捉摸[1-2]。在大规模的网络数据中，由于数据的实时性和动态性以及维度之间的复杂关联，异常缺失值会被正常的数据波动所掩盖。同时，随着超高维连续数据的数据规模和复杂度的增加，传统的检测方法很难将其准确识别出来。因此，有效检测和处理异常缺失数据是超高维连续数据处理的关键。

近年来，众多学者开展了缺失数据检测算法的相关研究。文献[3]以大坝的变形监测数据为研究对象，针对不同尺度变形分量数据的缺失问题，设计一种基于随机森林的检测算法，具有较高的检测精度。但处理超高维变形监测数据时，采用基于相关性的特征选择方法过于简单，可能遗漏与异常缺失相关但相关性不高的特征，影响检测准确性。文献[4]针对多元时间序列数据中缺失值，设计一种基于注意力重新表征的检测算法，具有良好的检测性能。在数据维度极高时，存在复杂交织的子序列关系，可能使注意力权重计算不准，导致对关键信息聚焦偏差，影响异常缺失值检测准确性。文献[5]针对长期监测数据中缺失值，设计一种基于时间序列压缩分割的检测算法。在超高维长期监测数据中，数据在多维度有复杂周期性或趋势性，若按简单规则分割，无法准确捕捉特征关系，影响异常缺失值检测。鉴于超高维连续数据异常缺失检测的重要性和传统算法的局限性，本文提出了一种基于改进 K 均值聚类的算法，旨在实现对超高维连续数据中异常缺失值的有效检测。

1　超高维连续数据降维处理

网络中的连续数据具有显著的“超高维”特性，其维度之高远远超出了传统数据的范畴。由于超高维连续数据的维度极高，数据点在高维空间中的分布变得极为稀疏，传统的数据处理方法在这样的环境下往往会陷入“维度灾难”，导致算法的效率急剧下降，甚至无法有效地运行。在异常缺失检测过程中，大量的特征使数据的模式和规律变得更加复杂，异常缺失值淹没在海量的数据中，难以准确地识别和定位。鉴于超高维连续数据的这些独特特性，为了简化数据处理流程、提高检测效率，并尽可能保留原始数据中的关键信息，对其进行降维处理就显得尤为必要[6-7]。本文引入了主成分分析法(PCA)，进行超高维连续数据的降维处理。

首先，收集超高维连续数据，并将其整理成矩阵形式 $\boldsymbol{Z}_{m\times n}$。在这个矩阵中，每一行代表 m 个超高维数据样本，每一列则代表 n 个样本特征。由于原始超高维连续数据矩阵内各特征的量纲和取值范围可能不同，为了消除这种差异对 PCA 结果的影响，需要对数据进行标准化处理[8]，公式如下：

$$\boldsymbol{Z}'_{m\times n}=\frac{\boldsymbol{Z}_{m\times n}-P_n}{C_n} \tag{1}$$

其中，$\boldsymbol{Z}'_{m\times n}$ 表示标准化处理后的超高维连续数据矩阵，P_n、C_n 分别表示原始超高维连续数据矩阵内特征的均值与标准差。

基于上式完成原始超高维连续数据矩阵的标准化处理后，需要计算其协方差矩阵[9]。协方差矩阵反映了原始数据中各特征之间的相关性，是 PCA 降维处理的关键步骤之一。协方差矩阵的计算公式：

$$\boldsymbol{F}=\frac{1}{m\times n}\boldsymbol{Z}'^{\mathrm{T}}_{m\times n}\boldsymbol{Z}'_{m\times n} \tag{2}$$

其中，$\boldsymbol{F}$ 表示超高维连续数据的协方差矩阵，$\boldsymbol{Z}'^{\mathrm{T}}_{m\times n}$ 表示矩阵 $\boldsymbol{Z}'_{m\times n}$ 的转置矩阵。

对协方差矩阵 $\boldsymbol{F}$ 进行特征分解，得到一系列特征值及其对应的特征向量，这些特征值和特征向量在 PCA 中扮演着重要角色：特征值表示主成分方向的方差大小，反映了该方向上的超高维连续数据变异程度；特征向量则表示主成分的方向，即超高维连续数据在该方向上的投影能够最大程度地保留原始数据的变异信息[10-11]。

根据实际需求和数据特性，选择前 K 个最大的特征值对应的特征向量作为主成分，这些主成分能够最大限度地保留原始数据中的关键信息，同时去除冗余和相关性。将选定的 K 个特征向量按列排列，构成主成分矩阵 $\boldsymbol{Q}_K$。将原始超高维连续数据矩阵 $\boldsymbol{Z}_{m\times n}$ 投影到由 $\boldsymbol{Q}_K$ 定义的低维空间中，即可得到降维后的数据矩阵 $\boldsymbol{X}$[12]：

$$\boldsymbol{X}=\boldsymbol{Z}_{m\times n}\times\boldsymbol{Q}_K \tag{3}$$

这个降维后的数据矩阵 $\boldsymbol{X}$ 是超高维连续数据在低维空间中的表示，其保留了原始数据中的关键信息，同时降低了数据维度和复杂性。

采用主成分分析法(PCA)，可以完成超高维连续数据的降维处理，得到低维连续数据，为后续异常缺失检测提供有力支持。

2　改进 K 均值聚类检测低维数据异常缺失

在超高维数据中，特征数量庞大，直接进行异常检测可能导致计算复杂度高、检测效果差。降维处

理可以有效减少特征数量,降低计算复杂度[13]。在降维后的低维空间中,采用传统 K 均值聚类算法在初始聚类中心的选择上存在随机性,可能导致聚类结果的不稳定,从而影响异常缺失检测的准确性[14-15]。为此,本文对 K 均值聚类算法进行改进,通过优化聚类中心选择和聚类迭代过程,更准确地划分数据类簇,进而更有效地识别出异常缺失值,以此降低检测误差。

为了更合理地选择初始聚类中心,本文采用了基于密度的聚类中心选择方法。首先,计算每个低维数据点的局部密度:

$$\rho_i = \sum_{j \neq i} \exp\left(\frac{\| X_i - X_j \|^2}{2H^2}\right) \tag{4}$$

其中,ρ_i 表示第 i 个低维数据点的局部密度;X_i、X_j 分别表示低维连续数据中第 i 个和第 j 个数据点;H 表示高斯核函数的带宽参数。

基于公式(4)求得的低维数据点局部密度,本文选择局部密度最大的 N 个数据点作为低维数据的初始聚类中心,完成低维数据初始聚类中心的优化选择后,即可进行聚类分析。在每次聚类迭代过程中,对于每个低维数据点,需要计算其与各个聚类中心之间的欧氏距离,公式如下:

$$D(X_i, U_l) = \sqrt{\sum_{w=1}^{W} (X_{iw} - U_{lw})^2} \tag{5}$$

其中,$D(X_i, U_l)$ 表示第 i 个低维数据点 X_i 和第 l 个初始聚类中心 U_l 之间的距离,W 表示数据维度。

根据公式(5)所求各低维数据点和各个初始聚类中心之间的距离,将其归属到距离最近的聚类中心,以此完成初始聚类。根据数据点的归属结果,更新每个聚类中心为所属数据点的均值,公式如下:

$$U_l^{'} = \frac{1}{| M_l |} \sum_{X_i \in M_l} X_i \tag{6}$$

其中,$U_l^{'}$ 表示更新后的聚类中心;M_l 表示归属到第 l 个初始聚类中心的低维数据点集合;$| M_l |$ 表示集合中数据点的数量。

根据公式(6)更新低维数据聚类中心后,重复上述步骤再次聚类。在低维数据的聚类迭代过程中,检查聚类中心是否改变,若改变则返回“数据点归属计算”步骤继续迭代;若聚类中心不变或达到预设迭代次数,则停止迭代,完成低维数据聚类分析并得到结果[16]。最后,依据数据分布特性、聚类中心密集程度等设定合理聚类阈值。根据聚类结果,计算低维数据点到所属聚类中心的距离,与阈值比较,从而得到数据异常缺失检测结果。具体判断方式如下:

$$Y_{X_i} = \begin{cases} 0, D(X_i, U_l) \geqslant \gamma \\ 1, D(X_i, U_l) < \gamma \end{cases} \tag{7}$$

其中,Y_{X_i} 表示低维数据点 X_i 的检测结果,γ 表示阈值。

如果待检测的低维数据点到其所属聚类中心的距离超过设定阈值,则将其视为异常的缺失点,输出检测结果为 0;否则将其视为正常数据点,输出检测结果为 1。

3 仿真实验

3.1 仿真设置

为了验证基于改进 K 均值聚类的超高维连续数据异常缺失检测算法的有效性和优越性,本文设计了仿真实验,将其与以下两种算法进行比较:文献[3]提出的基于 RF 的检测算法与文献[4]提出的基于注意力重新表征的检测算法。仿真对比实验硬件环境为 Intel Core i7 处理器,16 GB 内存以及 NVIDIA GTX 1080 Ti 显卡,所有算法均在 Windows 10 操作系统下运行。

在实验过程中,为了全面模拟实际应用中可能遇到的超高维连续数据情况,人工生成了 4 个超高维连续数据集,分别为弧形数据(DS1)、圆环形数据(DS2)、块状数据(DS3)和混合数据(DS4),其中的混合数据(DS4)由多种形状的数据组合而成,包括弧形数据(DS1)、圆环形数据(DS2)、块状数据(DS3)以及其他不规则形状数据。弧形数据占比 40%,圆环形数据占比 30%,块状数据占比 20%,其他不规则形状数据占比 10%。这些数据集的维度极高,数据包含了大量丰富的特征信息,每个样本在高维空间中具有复杂的结构和分布,是基于对实际应用场景的深入理解和抽象而生成的,具有很强的现实意义和应用价值。以金融分析领域为例,金融市场数据涉及到众多的因素,如股票价格、汇率、利率、成交量等,这些数据的实时变化形成了超高维的数据集。通过对金融数据的异常缺失检测,可以及时发现市场中的异常波动和风险因素,为投资者和金融机构提供决策依据。为了确保本文所提出的算法具有泛化性,在生成仿真数据集时,充分考虑不同数据分布、形状和异常情况的多样性,以尽可能涵盖实际应用中可能遇到的各种情况。超高维连续数据集概况见表 1。

表 1　超高维连续数据集概况

Table 1　Overview of ultra-high-dimensional continuous data sets

数据集	样本数量/组	缺失异常样本数量/个	类数目	形状
DS1	10 000	20	6	弧形
DS2	10 000	24	8	圆环
DS3	10 000	22	8	块状
DS4	10 000	18	10	混合

当特征维度超过样本量一个数量级时，可视为超高维数据。本文数据集中单样本包含 10 000 个样本数量（特征维度为 100 000），满足特征维度≫样本量的超高维标准。

3.2　聚类效果分析

本文实验数据的设计充分考虑了金融领域的数据特征。弧形数据（DS1）模拟了金融数据中的时间序列特性，如股票价格的周期性波动；圆环形数据（DS2）模拟了金融数据中的循环模式，如季节性交易量的变化；块状数据（DS3）模拟了金融数据中的异常事件，如市场崩盘或异常交易；混合数据（DS4）综合模拟了金融数据中的多种特征，如时间序列、循环模式、异常事件以及噪声数据。用本文设计的算法对 4 个超高维连续数据样本进行了聚类处理，结果如图 1 所示。

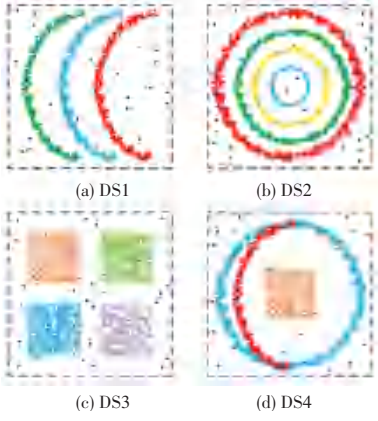

图 1　超高维连续数据集聚类结果

Fig. 1　Results of ultra - high - dimensional continuous data clustering

通过图 1 可以清晰地看到，本文设计的算法在处理形态各异、分布特征多样的超高维数据时，展现出了出色的稳定性，针对弧形数据、圆环形数据、块状数据、混合数据，该算法都能精准地捕捉到数据的分布规律，并据此将数据合理地划分为多个类簇，且各类簇间的边界清晰明了。该算法还能精准地辨识出超高维连续数据中的异常缺失值，即图 1 中的黑色原点标记，这一卓越表现主要得益于在聚类阶段，算法优化了传统的 K 均值聚类方法初始聚类中心的选择方式，使得算法在低维连续数据的聚类分析中表现出色，能够更准确地根据各数据点与聚类中心的距离来检测数据中的异常缺失值，使得本文设计的算法在进行超高维连续数据异常缺失检测时，展现出了优异的稳定性和适应性。

3.3　检测精度对比

将表 1 中的 4 个超高维连续数据集整合为一个包含 10 000 组超高维连续数据样本的大型数据集，并分别应用实验组算法以及对照组中的两种算法对各组样本进行异常缺失值检测，并采用均方根误差 ε 作为评估实验效果的关键指标，公式如下：

$$\varepsilon = \sqrt{\sum_{i}^{I} (Y_i - Y_i')^2 / I} \tag{8}$$

式中：Y_i、Y_i' 分别表示超高维连续数据异常缺失的真实值和检测结果，I 表示存在的异常缺失样本总数。

实验组算法与对照组中两种算法在超高维连续数据异常缺失值检测中的根均方误差对比结果如图 2 所示。

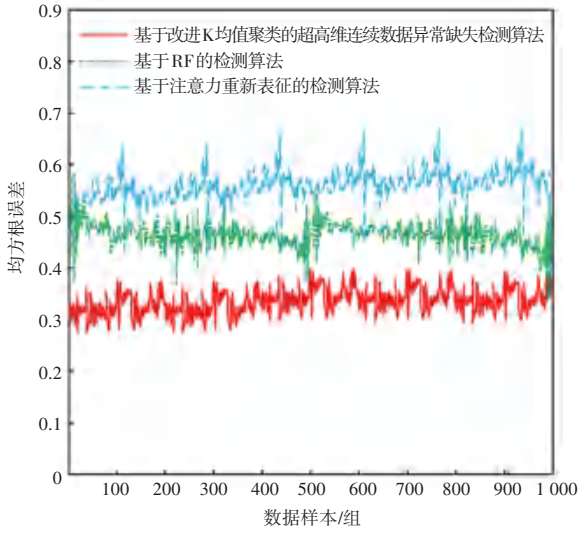

图 2　超高维连续数据异常缺失检测结果误差对比

Fig. 2　Error comparison of detection results of abnormal missing of ultra-high-dimensional continuous data

从图 2 可以看出，实验组算法在超高维连续数据异常缺失检测中展现出了最佳的检测精度，超高维连续数据异常缺失检测结果的均方根误差平均值仅为 0.356，相较于对照组中的两种算法，分别降低了 0.137 和 0.231。由此说明本文采用的改进 K 均值聚类算法可以极大地提升异常缺失值检测的精度，从而更高效地识别出异常缺失值。

3.4　检测性能对比

为了全面评估所设计算法的性能，选取了 PM

(内存峰值)作为关键的评价指标。PM 是指算法在运行期间所占用内存的最大瞬时值,因此应用实验组算法与对照组中的两种算法分别对 10 000 组超高维连续数据样本进行异常缺失检测的过程中,统计上述两种算法运行时计算机资源占用情况的动态变化,占用内存变化曲线的最大瞬时值即为 PM 值,具体结果如图 3 所示。

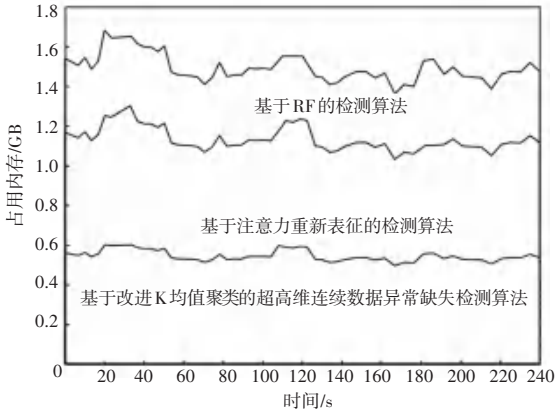

图 3 超高维连续数据异常缺失检测算法性能对比

Fig. 3 Comparison of the performance of ultra-high-dimensional continuous data anomaly missing detection algorithms

从图 3 可以看出,实验组算法在超高维连续数据异常缺失检测任务中内存占用表现稳定,且内存峰值仅为 0.59 GB,相较于对照组中的两种算法,分别降低了 1.08 GB 和 0.69 GB,说明实验组算法采用了先进的降维技术,能够迅速将超高维数据转化为低维表示,极大地减少了数据处理所需的内存空间,提高了算法的运行效率。在聚类阶段,实验组算法对 K 均值聚类算法进行了改进,优化了初始聚类中心的选择方式,不仅提升了异常缺失值检测的精度,还进一步减少了算法运行时的内存消耗,使得实验组算法在内存占用和检测精度方面都取得了显著优势。

4 结束语

本文提出一种基于改进 K 均值聚类的检测算法,旨在解决超高维连续数据中异常缺失值检测这一难题。运用主成分分析法对超高维数据进行降维处理,有效降低数据维度,减少了计算复杂度,而且还保留了关键信息;基于降维后的数据,对K均值聚类算法的初始聚类中心选择策略予以优化,显著提高了算法的收敛速度,并且增强了聚类结果的稳定性。通过将改进后的 K 均值聚类算法应用于降维后的数据,能够实现对异常缺失值的准确检测。在未来的研究上,将着重对算法参数进行进一步优化,增强算法的鲁棒性与适用性。同时,探索把深度学习等前沿技术融入该算法之中,期望能够进一步提高异常缺失检测的性能与效率。

参考文献

[1] 胡翔宇,陈庆奎. 面向车载设备数据流的异常检测方法[J]. 智能计算机与应用,2022,12(11):44-53.
[2] 王锐. 基于改进 LOF 的高维数据异常检测方法[J]. 电信工程技术与标准化,2023,36(3):41-45.
[3] 季骏,鲍中秋,张晓阳,等. 变形分量信息随机森林分析法在缺失数据处理中的应用[J]. 水力发电, 2023,49(7):95-100.
[4] 曾子辉,李超洋,廖清. 缺失值场景下的多元时间序列异常检测算法[J]. 计算机科学,2024,51(7):108-115.
[5] 蒲黔辉,张子怡,肖图刚,等. 基于时间序列压缩分割的监测数据异常识别算法研究[J]. 桥梁建设,2024,54(3):15-23.
[6] 方新怡,万晓霞,史硕,等. 基于稀疏表示的多光谱颜色数据降维方法研究[J]. 激光与光电子学进展, 2021,58(22):547-553.
[7] 崔子才,钟伯成,赵欣阳. 基于 GAN 和特征选择技术的入侵检测数据增强[J]. 智能计算机与应用, 2024,14(3):174-180.
[8] 徐丽燕,徐康,黄兴挺,等. 基于 Transformer 的时序数据异常检测方法[J]. 计算机技术与发展, 2023,33(3):152-160.
[9] 乔非,翟晓东,王巧玲. 面向多维特性数据的缺失值检测及填补方法对比[J]. 同济大学学报(自然科学版),2023,51(12):1972-1982.
[10] 张可. 基于故障树的 IMS 网络运维场景异常检测算法[J]. 现代传输,2023,211(1):61-64.
[11] 崔子才,钟伯成,赵欣阳. 基于 GAN 和特征选择技术的入侵检测数据增强[J]. 智能计算机与应用, 2024,14(3):174-180.
[12] 余嘉茵,何玉林,崔来中,等. 针对大规模数据的分布一致缺失值插补算法[J]. 清华大学学报(自然科学版), 2023,63(5):740-753.
[13] 赵强,刘胜杰,韩东成,等. 基于改进 K 均值聚类的光伏板缺陷检测方法[J]. 红外技术,2024,46(4):475-482.
[14] 周杨,王春林,郭锐. 基于随机森林算法的数据中心运维异常告警方法[J]. 现代电子技术,2023,46(8):143-148.
[15] 王锐. 基于改进 LOF 的高维数据异常检测方法[J]. 电信工程技术与标准化,2023,36(3):41-45.
[16] 周雪峰,徐强,谭艳婷,等. 基于改进灰色聚类算法的云架构数据中心网络异常流量过滤算法[J]. 电信科学, 2023,39(7):90-98.