改进K均值聚类的海洋数据异常检测算法研究

蒋 华 武 尧 王 鑫 王慧娇
(桂林电子科技大学计算机与信息安全学院 广西 桂林 541004)

摘 要 针对海洋Argo浮标监测数据中的异常数据挖掘问题,在改进K均值算法的基础上,提出基于距离为准则进行海洋异常数据判定的异常检测算法。该算法重新定义海洋数据邻近度,并根据数据的规模以及分布情况,区块化、自适应地筛选备选初始聚类中心;在算法迭代过程中,运用簇内,数据对象相对于聚类中心的距离均值,全局考量类簇内,符合异常特征的数据对象进行异常检测。通过仿真数据集和真实数据集分别进行实验验证,对比结果表明:该算法在聚类性能以及异常检测方面都优于对比算法。

关键词 K-means算法,Argo浮标数据,邻近度,区块化,异常检测
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.032

Study on Ocean Data Anomaly Detection Algorithm Based on Improved K-means Clustering

JIANG Hua WU Yao WANG Xin WANG Hui-jiao
(College of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China)

Abstract Aiming at the problem of abnormal data mining in marine Argo buoy monitoring data, an anomaly detection algorithm based on distance criterion was proposed based on the improved K-means algorithm. The algorithm redefines the proximity of ocean data, blocks according to the size and distribution of the data, and adaptively selects alternative initial clustering centers. In the iterative process of the algorithm, using the distance mean of the data objects in the cluster relative to the clustering center, the global consideration is given to the data objects in the cluster according with the abnormal features to detect the anomalies. The simulation dataset and the real dataset are verified by experiments, and the comparison results show that it is superior to the contrast algorithm in clustering performance and anomaly detection.

Keywords K-means algorithm, Argo buoy data, Proximity, Block, Anomaly detection

1 引言

Argo计划为实时观测全球海洋提供了平台,全球数千个Argo浮标旨在快速、准确、大规模地收集丰富的海洋温度、盐度剖面资料。但浮标由于在海上长期工作,其抛弃式、海上漂泊的特性、传感器的不稳定性(例如:其携带的电导率传感器漂移会导致测量数据产生误差,从而影响观测资料的质量)或卫星传送出现误码等因素,都会使得观测结果产生较大的偏差[1]。异常检测是数据挖掘领域的重要研究方向之一,该技术的目的在于在待检测数据集中,高效准确地判别出与正常分布模式存在偏差的疑似异常数据[2]。鉴于以上问题,随着各国对海洋资源开发的不断重视,Argo浮标异常数据检测成为当下众多学者的热门研究课题。文献[3]对比了现有Argo浮标与"最佳匹配"的历史剖面资料,并且与"三倍标准差"相结合;文献[4]引入了梯度相关空间尺度法(GDSM),通过对比理论测试与现场测试浮标数据,对异常数据进行剔除。

K均值算法是数据挖掘中实际应用十分广泛的聚类算法,该算法易于实现,其时间复杂度为$O(nki)$(其中,n为待处理数据集的数据量,k为聚类个数,i为算法的迭代次数),十分高效[5-6]。但K均值算法的缺陷也同样显著:对初始聚类中心的选取依赖性强,其选取结果将直接影响聚类效果[7]。

此外,聚类中心选择的合理性将直接影响异常检测精度,并且为了使选择算法更好地切合数据的实际分布情况,本文在研究海洋Argo浮标数据特性、现有的异常检测算法、K均值算法的基础上,提出基于改进K均值的海洋数据异常检测算法。该算法参照海洋数据邻近度的测定结果,自适应筛选备选序列中合理性较高的数据对象作为新的初始聚类中心,在有效发掘海洋数据集中簇结构的同时,异常数据的剔除也更为准确。实验结果表明,该改进算法聚类与异常检测效果提升显著。

到稿日期:2018-06-06 返修日期:2018-09-15 本文受广西科技重大专项(AA18118025),桂林电子科技大学研究生教育创新计划项目(2017YJCX48)资助。

蒋 华(1963-),男,博士,教授,主要研究方向为信息安全;武 尧(1994-),男,硕士,主要研究方向为信息安全、数据挖掘、海洋大数据,E-mail:907149625@qq.com(通信作者);王 鑫(1976-),男,硕士,副教授,主要研究方向为无线传感网络;王慧娇(1976-),女,硕士,副教授,主要研究方向为无线传感器网络。

2 相关工作

K 均值异常检测算法中初始聚类中心选取的改进方式大体分为 3 种:随机抽样、距离优化和密度估计[8]。但是,这些算法中从真实聚类中心的位置与待选数据对象动态全局考量的却不多,都具有一定的局限性。

文献[9]对待分类数据对象周围的密集性进行排序,选取指定值(K)个密集性最大的数据对象作为初始聚类中心,但单纯依据数据密集性并不能保证选取的初始聚类中心为全局均匀分布;文献[10]提出将数据对象的密集性与最近最大距离保证相结合的方式来选取初始聚类中心,但与文献[9]类似,每次选取初始聚类中心需要反复计算各数据对象的密集性,并且密度参数的确定也十分困难,限制了其性能;文献[11]基于距离最大原则选取初始聚类中心,并且为规避聚类结果随机性的问题,选择使用相距距离最大的两个数据对象作为最初始的聚类中心进行迭代选取。此举一定程度上避免了基于密集性选取的缺陷,但在其初始聚类中心选取的过程中,仅考虑数据对象间的相对距离,未兼顾到与先前聚类中心的邻近程度,数据对象未均匀分布时,会导致初始聚类中心选取过于集中,从而使算法陷入局部最优;韩崇等[12]基于数据流离散点的异常检测算法,将某簇类内部数据对象 S_{ij} 到质心 C_j 的距离 d 与该类簇中全部 S_{ij} 到 C_j 的距离的中位数的比值作为离群点的判断标准。但该算法在运行时需要多次排序查找中位数,时间复杂度和空间复杂度高,并且不能完全反映数据的总体情况。

本文基于上述算法的研究,适当结合其各自优势,对先前算法的缺陷给予改进,并分析 Argo 浮标数据的特点,提出基于改进 K 均值聚类的海洋数据异常检测算法。

3 问题定义

本节将针对文章中提及的参数与概念给予定义。

Argo 浮标数据可用于分析上层海洋热、盐含量的空间分布、季节和年季变化特征。遂需测量温度、盐度等属性,并且这些属性间存在一定的关联性。为此,Argo 浮标数据为包含压力、温度、PH 值等属性的多维数据对象,表示为 $Data=\{x_1,x_2,x_3,\cdots,x_{n-2},x_{n-1},x_n\}$,共有 n 个数据对象,数据对象皆为 m 维数据,第 i 个数据对象可以表示为 $x_i=\{x_{i1},x_{i2},x_{i3},\cdots,x_{im-2},x_{im-1},x_{im}\}$。其中,$x_{im}$ 表示第 i 个数据对象的第 m 维属性。

参数及概念定义如下。

定义 1(质心) 质心为各聚类中心,表示为 $c_1,c_2,\cdots,c_i$ $(i=1,2,3,\cdots,k)$,其中各聚类内部包含 v 个数据对象,记为 $x_{i1},x_{i2},\cdots,x_{iv}$。质心的更新公式如下:

$$c_i'=\frac{\sum_{x_{iv}\in Centers[i]}x_{iv}}{CenterCount[i]} \tag{1}$$

其中,$CenterCount[i]$为聚类 $Centers[i]$中数据集对象的数量,$i(i=1,2,3\cdots k)$为聚类标号。

定义 2(聚类程度) 聚类程度作为评价聚类质量的标准,采用误差平方和 SSE 作为目标函数,最终实现各聚类的 SSE 最小的聚类目标。聚类准则函数如下:

$$J_{SSE}=\sum_{i=1}^{k}\sum_{v=1}^{m}Dist(x_{iv},c_i) \tag{2}$$

其中,v 表示类簇所包含的所有数据对象个数;$Dist(x_{iv},c_i)$表示数据对象 x 与聚类中心 c_i 的距离。

定义 3(距离准则) 采用欧氏距离作为数据相似度的度量。计算第 i 个簇中的第 v 个数据对象 x_{iv} 与该簇聚类中心 c_i 的欧氏距离的方法如下:

$$Dist(x_{iv},c_i)=\sqrt{\sum_{p=1}^{m}(c_i^p-x_{iv}^p)^2} \tag{3}$$

其中,p 为数据对象维度。

定义 4(邻近度) 本文定义的邻近度是指数据集某一划分状态下,不同数据对象相对于当前类簇中心的相对邻近程度,用参数 $Proximity$ 表示。此外,为该参数设置阈值 α 和 β 用于界定范围,公式定义为:

$$Proximity=\left|\frac{\sqrt{\sum_{p=1}^{m}(c_i^p-x_{ia}^p)^2}}{\sqrt{\sum_{p=1}^{m}(c_i^p-x_{ib}^p)^2}}\right| \tag{4}$$

初始中心点选取可能出现如下情形:

情形 1:满足式(5)时,筛选数据对象 x_{ia} 为新的初始聚类中心。

$$\alpha\leqslant Proximity \tag{5}$$

情形 2:满足式(6)时,筛选数据对象 x_{ib} 为新的初始聚类中心。

$$Proximity\leqslant\beta \tag{6}$$

情形 3:以上条件都不满足时,筛选数据对象 x_{ia} 和 x_{ib} 都作为新的初始聚类中心。

鉴于邻近度阈值 α 和 β 的界定将直接影响到初始聚类中心选取的合理性;并且在选取开始的若干次中,由于区块内数据对象的规模较大,数据对象的聚集程度小,此时确定 α 和 β 不具有代表性。因此,本文优先以末次迭代区块内部备选聚类中心的相对位置为指定邻近度阈值的参考依据,通过多次实验优化来确定邻近度阈值。

定义 5(异常数据判断标准) 定义数据对象的异常度,用参数 Abn_x 表示。算法初始化过程完成之后,迭代更新质心,在每个簇类中,若属于该簇的数据对象 x_{iv} 与该簇聚类中心 c_i 的距离大于该簇中所有数据对象与聚类中心距离的均值时,则 Abn_x++,即不等式(7)成立时,数据对象的异常度 Abn_x 自增 1。若 Abn_x 大于给定阈值 λ,则判断其为异常数据,将其从原始数据集中剔除并放入异常数据集中。

$$Dist(x_{iv},c_i)\geqslant\frac{1}{t_i}\sum_{x_{iv}=c_i}Dist(x_{iv},c_i) \tag{7}$$

其中,t_i 表示 c_i 代表的簇拥有的数据对象的数量;$x_{iv}=c_i$ 表示属于 c_i 这个类簇的所有数据对象 x_{iv},$Dist(x_{iv},c_i)$表示数据对象 x_{iv} 与聚类中心 c_i 的距离。

经实验验证:阈值 λ 为 3 时,即不会因为 Abn_x 设置过低导致判定过于敏感,过多正常数据被错误判定;也不会因为 Abn_x 设置过高导致判定过于笼统,过多异常数据对象未被识别。

[1] http://www.argo.org.cn/

此时的判定效果最佳,因此将阈值 λ 设置为 3 较为合理。

4　海洋数据的异常检测

4.1　初始化过程的改进

K 均值算法的初始聚类中心的选取方式为随机选取 k 个数据对象,这样易导致聚类结果的随机性较大。因此,更为合理地选取初始聚类中心,会使聚类效果得到提升。出于上述考虑,本文需先提出两点原则。

原则 1　聚类中心最大距离

聚类算法中,聚类结果始终以数据对象的相似度为度量:类内数据对象相似度最大,类间相似度最小。K-means 算法中,数据对象的相似度通常以欧氏距离为度量,数据对象间的相似度与其相距距离成反比关系。为此,初始聚类中心的选取应考虑数据点间的距离关系,即初始聚类中心间的距离应该本着尽可能大的原则,以保证初始聚类中心的均匀分布。

原则 2　迭代划分,区块化获取

考虑到聚类中心出现在数据对象密集区域的可能性较大,并兼顾全局数据密集程度,因此采用在前一个初始聚类中心选取完成之后选取新的初始聚类中心时,选择在数据对象规模最大的类簇中基于原则 1 迭代划分、区块化获取初始聚类中心的方法。该方法可在保证聚类中心均匀分布的同时,有效降低算法的时间复杂度、空间复杂度。

但仅仅基于上述考量,可能会导致新聚类中心的选取过于集中以及不合理。为此,加入数据对象邻近度(定义 4)用于解决此类问题。图 1、图 2 用于解释上述说法。

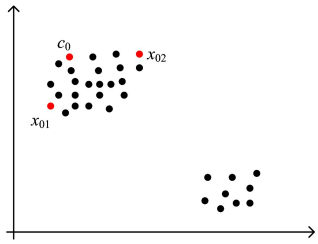

图 1　理想状态下候选初始聚类中心的相距关系

Fig. 1　Candidate initial clustering center distance relations in ideal state

图 1 为理想状态下候选初始聚类中心 x_{01} 和 x_{02} 与先前聚类中心 c_0 的相距关系。c_0 与 x_{01} 和 x_{02} 相距都较远,选取 x_{01} 和 x_{02} 中任何一个作为新的初始聚类中心都符合原则 1 的需求,且都较为合理。

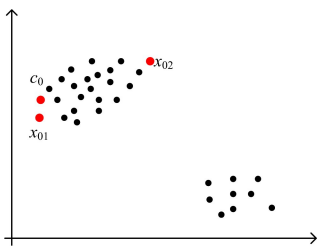

图 2　特殊情况下候选初始聚类中心的相距关系

Fig. 2　Candidate initial clustering center distance relations under special circumstances

图 2 中,c_0 与 x_{01} 过近,与 x_{02} 相距较远,选取 x_{01} 作为新的初始聚类中心不合理,会导致选取的初始聚类中心过于集中。

算法 1　初始化算法

步骤 1　初始化参数。将聚类中心点集 Centers[k]、各聚类包含点数 CenterCounts[k]设置为空。

步骤 2　查找数据对象数量最多的类簇,即:max(CenterCount[c_i]),并将其标记为 M。

步骤 3　若 Centers[k]为空,则说明为首次选取。因此基于原则 1 查找 M 中的数据对象,即 max(Dist(x_{ia},x_{ib}))(x_{ia},x_{ib}∈Data),并将它们放入 Centers[k]中作为最初始的两个最初始聚类中心;若 Centers[k]不为空,则同样先基于原则 1 选取 M 中的数据对象后,将它们放入临时变量中作为备选的初始聚类中心。

步骤 4　利用式(4)计算新初始聚类中心备选数据对象的邻近度 Proximity,并利用定义 4 判定和选取新的初始聚类中心。

步骤 5　重新划分聚类,计算 M 中所有数据对象与 M 中聚类中心的欧氏距离(定义 3),对于数据对象 x_j,若使得 Min(Dist(x_j,c_i))成立,则将数据对象 x_j 划分到 c_i 所代表的簇中。其中,c_i 为 M 的聚类中心。

步骤 6　重复步骤 3－步骤 5 直到最终初始聚类中心点集 Centers[k]放满,从而得到 k 个初始聚类中心。

改进算法以选取更为合理的初始聚类中心,提升聚类效果为最终目的。首先,基于聚类中心距离最大原则,选取最初始的两个初始聚类中心并进行划分;然后,基于原则 2 选取当前状态下数据对象规模最大的区块,使聚类中心选取远离规模小且离群的小簇;最后,基于原则 1,在当前区块内引入数据对象的邻近度,筛选新的初始聚类中心,保证各选取好的初始聚类中心均匀分布。

4.2　海洋数据异常检测

聚类中心选择的合理程度将直接影响异常检测的准确性。基于上述考量,在改进初始聚类中心选取的合理性并有效提升准确性的基础上,加入异常检测机制。

算法 2　基于 K 均值聚类的数据异常检测框架

输入:m 维数据集 Data=$\{x_1,x_2,x_3,\cdots,x_{n-2},x_{n-1},x_n\}$,聚类个数 k,聚类收敛阈值 η,邻近度阈值 α 和 β,异常度 Abn_x

输出:聚类中心点集 Centers[k],各个点所属聚类标号 Lables[n],各数据对象异常度 Abn_x,异常数据对象集合 Exc[n]

步骤 1　初始化参数。将聚类中心点集 Centers[k]、各聚类包含点数 CenterCounts[k]、各个点所属聚类标号 Lables[n]设置为空;各数据对象的异常度 Abn_x 设置为 0。

步骤 2　运用算法 1,实现基于 K 均值聚类的数据异常检测框架中 k 个初始聚类中心的选取,选取结果放在 Centers[]中。步骤 3－步骤 7 将用于详细描述海洋数据集的异常检测过程。

步骤 3　将当前状态下各聚类包含的数据点数放入 CenterCounts[]中,各个数据对象所属聚类标号放入 Lables[]中。

步骤 4　重新划分聚类,计算类簇中所有数据对象与聚类中心 Centers[i]的欧氏距离(定义 3),对于数据对象 x_j,若使得 Min(Dist(x_j,c_i))成立,将数据对象 x_j 划分到 c_i 所代表的簇中。其中,c_i 为 M 中的聚类中心(i=1,2,3,…,k)。

步骤 5 继续迭代，持续更新质心(定义 1)。

步骤 6 每次迭代的同时，运用定义 5 对异常数据对象进行判定，若判定为异常则将其从 Data 中剔除，并放入集合 Exc[]中。

步骤 7 运用准则函数(定义 2)，判定$|J'-J|\leqslant\eta$是否成立(J 是上次聚类的准则函数，J′是本次聚类的准则函数)。若成立，则算法结束，输出 Centers[]，Lables[]，Abn_x，Exc[]；否则重复步骤 3－步骤 7。

为直观地描述基于 K 均值聚类的海洋数据异常检测算法的具体流程，给出该算法的流程图，如图 3 所示。

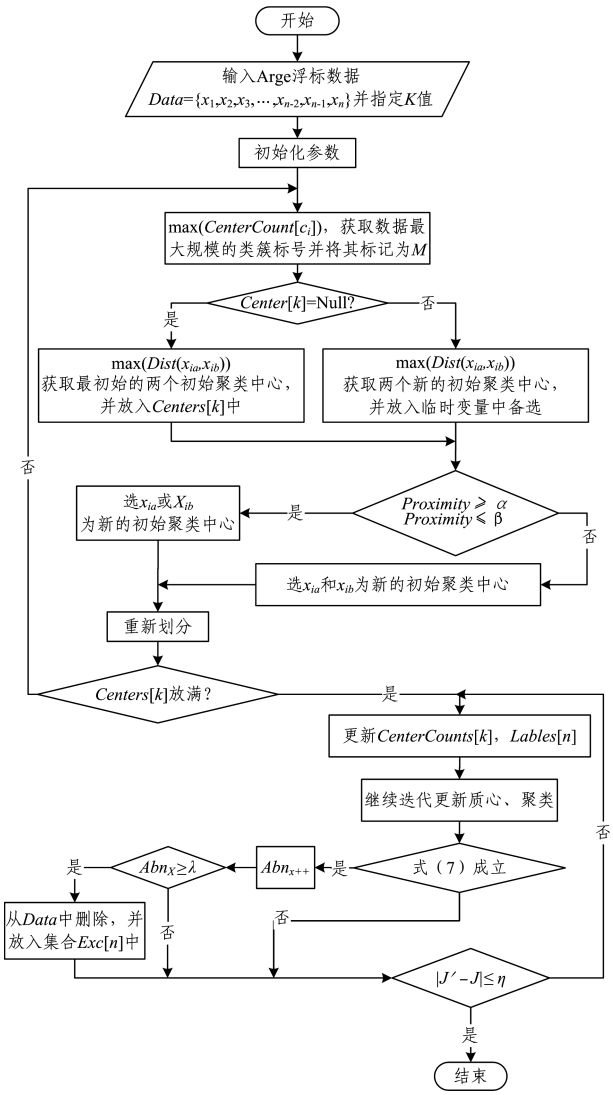

图 3 改进算法的流程图

Fig. 3 Flowchart of improved algorithm

本文提出的算法并不与 DSOBD 算法类似，而是仅将每个数据对象与其他所有数据对象的距离作为异常判定标准的方法；也不采取类似于文献[13-15]中异常判定仅针对数据规模小的类簇的方式，而是在算法迭代过程中，根据各类簇的数据分布特点，运用簇内数据对象相对于聚类中心的距离均值，全局考量该类簇内符合异常特征的数据对象，并对其进行检测和剔除。

5 实验设计与结果分析

为验证改进算法的有效性，采用仿真数据和真实数据分别进行实验分析，并且为避免实验结果的随机性，3 种算法分别运行 10 次取结果的均值。参照指标为检测率、误检率、异常点误差平方和、正常点误差平方和 4 个指标数据，将基于 K 均值的异常检测算法、基于最大距离的异常检测算法与本文改进的异常检测算法进行对比。

本实验环境为：处理器为 Intel(R)Core(TM) i5-3210M；内存为 8GB；硬盘容量为 500GB；操作系统为 Windows10 64 bit；JDK 版本为 1.8。实验所用 IDE 为 MyEclipse 2017，Matlab R2012b，编程语言为 Java。实验参数如表 1 所列。

表 1 实验参数

Table 1 Test parameters

	η	α	β
基于 K 均值的异常检测算法	0.01	null	null
基于最大距离的异常检测算法	0.01	null	null
改进的异常检测算法	0.01	2.75	0.375

5.1 仿真数据集

欲验证算法是否能有效聚类并且有效检测异常数据，可先使用已标注异常标识的数据集进行实验。鉴于上述原因，先运用高斯分布函数生成包括 3 个主要聚类的 1030 个二维数据对象，其中预先加入 15 个与大部分数据具有显著差异性的数据点和 5 个差异性稍小的数据点，共 1050 个二维仿真数据集。

实验结果如图 4－图 6 所示，3 种算法都能有效地划分聚类，但改进算法的数据对象的簇结构划分得更为均匀和清晰；同时，基于 K 均值的异常检测算法(算法 1)、基于最大距离的异常检测算法(算法 2)和本文改进的异常检测算法(算法 3)都能进行异常数据检测。其中，算法 1 检测出了 129 个异常数据对象，算法 2 检测出了 11 个异常数据对象，算法 3 检测出了 27 个异常数据对象。对比实验结果表明：本文改进的异常检测算法能有效地进行异常数据检测。

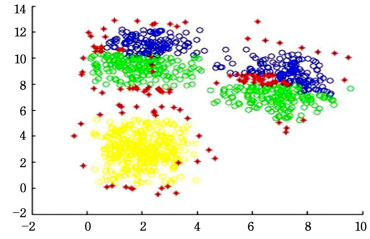

图 4 基于 K 均值的异常检测算法的异常检测结果

Fig. 4 Abnormal detection results of abnormal detection algorithm based on K-mean

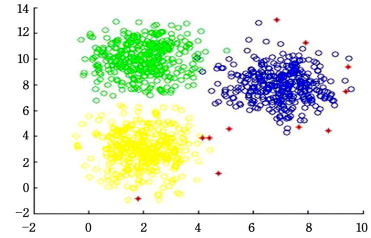

图 5 基于最大距离的异常检测算法的异常检测结果

Fig. 5 Anomaly detection results of anomaly detection algorithm based on maximum distance

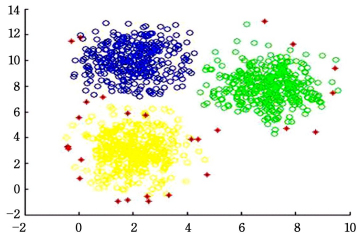

图 6　改进算法的异常检测结果

Fig. 6　Abnormity detection results of improved anomaly detection algorithm

若要验证算法的有效性，仅依靠异常数据对象的数目时说服力不足，因此加入检测率、误检率和异常数据对象的误差平方和。本文将针对上述指标在仿真数据集上对这 3 种算法进行验证，实验结果如表 2 所列。

表 2　仿真数据实验结果

Table 2　Experimental results of simulation data

实验参数（均值）	聚类算法		
	基于 K 均值	基于最大距离	改进算法
异常数据个数/个	129	11	27
检测率/%	75	45	85
误检率/%	11.07	0.29	0.97
异常数据单位误差平方和	68.53	4073.88	54.91

由表 2 可得，基于 K 均值的异常检测算法检测出的异常点数量过多，这说明该算法易发生“错误判定”现象，将正常数据判定为异常数据，误检率将直接受其影响，均值都在 11% 左右，此结果不符合 Argo 浮标监测数据质量控制要求；基于最大距离的异常检测算法中，由于其初始聚类中心选取的随机性仍较大，导致其检测结果的稳定性欠佳。

由表 2 可以看出，相比于基于 K 均值的异常检测算法，改进的异常检测算法的检测率平均要提升 9～10 个百分点；误检率降低也十分明显，平均降低了 9 个百分点。同样，与基于最大距离的异常检测算法相比，改进算法的检测率平均降低了 40 个百分点；在误检率方面，二者相差不多，因为与基于最大距离的算法检测出的异常点的数量较少有关。

改进算法基于两项原则，初始聚类的选取更为合理，理论上聚类效果更优。为验证改进算法能否在有效检测异常数据的同时优化聚类效果，本文采用对比正常数据标准差的方式加以验证。由图 7 可知，相对于其他两种对比算法，改进算法的标准差明显更小，由此可知，改进算法的聚类效果更佳。

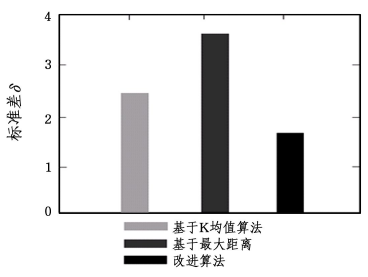

图 7　正常数据标准差

Fig. 7　Standard deviation of normal data

5.2　真实数据集

真实数据集采用中国 Argo 实时资料中心[1)] 获取的 2017 年印度洋 Argo 浮标剖面数据，数据格式为 .dat。数据文件中包含仪器型号、数据模式、放置日期、编号等设备基础信息，以及测量周期、压力、温度、盐度等测量数据信息。

5.2.1　数据预处理

直接获取到的 Argo 浮标原始数据中包含众多与异常检测无关的冗余属性数据，因此不能直接将未处理的数据文件直接导入，需要对其进行属性筛选以及控制单一变量等预处理。

考虑到 Argo 浮标数据存在空间分布、季节和年季变化特征，同一年内四季的测量数据具有显著差异性。为此，本实验选取盐度与压力进行验证，二者具有关联性并且可预计。为控制“单一变量”，选取同一网格内部（−35.614，62.896）、同一浮标（标号 1900050）、同一时间（10:00 AM）、一年内（2017）的 365 个观测数据进行实验。

表 3　Argo 浮标数据

Table 3　Argo buoy data

	数量	占比/%
簇 1	90	24.66
簇 2	91	24.93
簇 3	92	25.21
簇 4	92	25.21
总量	365	100

5.2.2　实验结果

3 种算法的运行对比结果如表 4 所列，在真实数据情况下，相对于对比算法的检测结果，改进算法异常数据的检测率最高，误检率最低，偏差值最小，聚类稳定性更优。因此，改进算法表现出良好的检测性能。

表 4　真实数据的实验结果

Table 4　Experimental results of real data

实验参数（均值）	异常检测算法		
	基于 K 均值	基于最大距离	改进算法
检测率/%	64.70	71.24	85.91
偏差值/%	18.48	38.09	5.47
误检率/%	5.61	9.00	1.77
异常数据单位误差平方和	7.69	3972.91	43.51

结束语　基于 K 均值算法、异常检测、海洋 Argo 浮标数据的研究，本文提出了一种改进算法。该算法以全局考量数据的分布动态为前提，首先基于聚类中心最大距离原则与迭代划分、区块化相结合原则，加入邻近度思想选取初始聚类中心；然后基于可完全反映数据总体情况的自定义异常度判定标准，进行异常数据检测。实验证明，基于 K 均值的海洋数据异常检测算法相比于对比算法更具优势。大数据时代数据量呈指数增长，在海洋数据异常检测领域使用大规模数据计算引擎处理更高维度、大宗数据集将是我们下一步努力研究的重点。

参 考 文 献

[1]　LIU Z H，WU X F，XU J P，et al. Argoocean observations in

1) http://www.argo.org.cn/

China for 15 years [J]. Progressin Geoscience, 2016, 31(5): 445-460. (in Chinese)
刘增宏,吴晓芬,许建平,等. 中国 Argo 海洋观测十五年[J]. 地球科学进展,2016,31(5):445-460.

[2] DING J, WANG L, SHEN D, et al. An Anomaly Detection System on Big Data[J]. Natural Science Journal of Hainan University, 2015, 33(1): 24-27.

[3] WANG H Z, ZHANG R, WANG G H, et al. Quality Control Technology of temperature and Salt profile observation data of Argo buoy [J]. Journal of Geophysics, 2012, 55(2): 577-588. (in Chinese)
王辉赞,张韧,王桂华,等. Argo 浮标温盐剖面观测资料的质量控制技术[J]. 地球物理学报,2012,55(2):577-588.

[4] SHAOLEI L U, HONG L I, LIU Z. Improvement of Argo salinity data delayed-mode quality control Method[J]. Journal of Pla University of Science& Technology, 2014, 15(6): 598-606.

[5] TZORTZIS G, LIKAS A, TZORTZIS G. The MinMaxk-Means clustering algorithm[J]. Pattern Recognition, 2014, 47(7): 2505-2516.

[6] CHEN G P, WANG W P, HUANG J, et al. Improved initial clustering center selection method for k-means algorithm [J]. Journal of Chinese Computer Systems, 2012, 33(6): 1320-1323.

[7] XING C Z, GU H. K-means algorithm for optimizing initial clustering centers based on average density [J]. Computer Engineering and Application, 2014, 50(20): 135-138. (in Chinese)
邢长征,谷浩. 基于平均密度优化初始聚类中心的 k-means 算法[J]. 计算机工程与应用,2014,50(20):135-138.

[8] CELEBI M E, KINGRAVI H A, VELA P A. A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm[J]. Expert Systems with Applications, 2013, 40(1): 200-210.

[9] HAN Z J. Adaptive K-means initialization method based on data denseness [J]. Computer Applications and Software, 2014, 31(2): 182-187. (in Chinese)
韩最蛟. 基于数据密集性的自适应 K 均值初始化方法[J]. 计算机应用与软件,2014,31(2):182-187.

[10] ZUO J, CHEN Z M. Anomaly detection algorithm based on improved K-means clustering [J]. Computer Science, 2016, 43(8): 258-261. (in Chinese)
左进,陈泽茂. 基于改进 K 均值聚类的异常检测算法[J]. 计算机科学,2016,43(8):258-261.

[11] CHEN G P, WANG W P, HUANG J. An improved K-means algorithm for initial clustering Center selection [J]. Minicomputer System, 2012, 33(6): 170-173. (in Chinese)
陈光平,王文鹏,黄俊. 一种改进初始聚类中心选择的 K-means 算法[J]. 小型微型计算机系统,2012,33(6):170-173.

[12] HAN C, YUAN Y S, MEI T, et al. Outlier Detection algorithm based on K-means [J]. Computer Engineering and Application, 2017, 53(3): 58-63. (in Chinese)
韩崇,袁颖珊,梅焘,等. 基于 K-means 的数据流离群点检测算法[J]. 计算机工程与应用,2017,53(3):58-63.

[13] SAMRIN R, VASUMATHI D. Hybrid Weighted K-Means Clustering and Artificial Neural Network for an Anomaly-Based Network Intrusion Detection System[J]. Journal of Intelligent Systems, 2016, 27(2): 135-147.

[14] SHEN G. Improved k-means initialization method based on data density [J]. Computer Engineering & Applications, 2014, 51(11): 139-144.

[15] TZORTZIS G, LIKAS A, TZORTZIS G. The MinMax k-Meansclustering algorithm [J]. Pattern Recognition, 2014, 47(7): 2505-2516.