陈利军,王畅.基于 DBSCAN 的地震电离层扰动异常数据检测方法[J].地震工程学报,2020,42(2):410-415.doi:10.3969/j.issn.1000－0844.2020.02.410

CHEN Lijun,WANG Chang.Detection Method for Seismic Ionospheric Disturbance Anomaly Data Based on DBSCAN[J].China Earthquake Engineering Journal,2020,42(2):410-415.doi:10.3969/j.issn.1000－0844.2020.02.410

基于 DBSCAN 的地震电离层扰动异常数据检测方法

陈利军[1,2], 王 畅[3]

(1. 中国人民解放军战略支援部队信息工程大学, 河南 郑州 450001;
2. 河南经贸职业学院 计算机工程学院, 河南 郑州 450018;
3. 华北水利水电大学 信息工程学院, 河南 郑州 450046)

摘要: 电离层扰动变化可以作为地震的短临前兆,为提升电离层扰动数据分析准确性,剔除异常数据,提出了基于 DBSCAN 的地震电离层扰动异常数据检测方法。首先利用 DBSCAN 方法的中心思想,分析电离层扰动数据,结合无线传感器网络数据的特点,将欧式距离设定为指标,对比不同电离层扰动数据并进行划分聚类,然后计算地震电离层扰动数据相似度。再利用 DBSCAN 方法提取相关的环境特征集,通过特征集实现地震电离层扰动异常数据检测。最后实验分析与传统方法相比,运行时间、误报率、漏报率以及检测率的对比结果。结果表明:所提方法的所需时间短,误报率和漏报率明显降低,检测率大大提高。

关键词: DBSCAN; 地震电离层扰动; 异常数据; 检测

中图分类号: TP391 **文献标志码:** A **文章编号:** 1000－0844(2020)02－0410－06

DOI:10.3969/j.issn.1000－0844.2020.02.410

Detection Method for Seismic Ionospheric Disturbance Anomaly Data Based on DBSCAN

CHEN Lijun[1,2], WANG Chang[3]

(*1.PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, Henan, China;*
2.Computer Engineering College, Henan Institute of Economics and Trade, Zhengzhou 450018, Henan, China;
3.Information Engineering College, North China University of Water Resources and Electric Power, Zhengzhou 450046, Henan, China)

Abstract: To improve the accuracy of the analysis of ionospheric disturbance data and eliminate anomalous data, this study proposes a detection method for seismic ionospheric disturbance anomaly data based on DBSCAN. According to the characteristics of wireless sensor network data, the Euclidean distance is set as the index, and different ionospheric disturbance data are compared and clustered. The similarities of the seismic ionospheric disturbance data are then calculated. The

收稿日期: 2019-12-01

基金项目: 河南省科技攻关项目(182102210329);河南省大数据双创基地项目(豫发改高技〔2017〕945 号)

第一作者简介: 陈利军(1978－),男,河南新乡人,硕士,副教授,研究方向:计算机。E-mail:xxgl126@126.com。

通信作者: 王 畅(1975－),女,河南周口人,硕士,副教授,研究方向:网络安全。

DBSCAN method is used to extract relevant environmental feature sets, through which the seismic ionospheric disturbance anomaly data can be detected. The run time, false positive rate, false negative rate, and detection rate of the proposed method are compared with those of a traditional method. Experimental results show that the proposed method requires minimal time, significantly reduces false positive and false negative rates, and improves detection rates.

Keywords: DBSCAN; seismic ionospheric disturbance; abnormal data; detection

0　引言

电离层数据的扰动变化可以作为地震的短临前兆[1-2]，因此，如何保证电离层扰动数据的准确可靠是需要解决的主要问题，其实质为如何剔除电离层扰动数据中的异常数据。目前，使用较为广泛的数据处理技术主要有 k-means 算法、模糊聚类算法等，以上方法都存在一定的不足，例如需要人为确定聚类数量、难以组建合理的隶属度函数、计算过程较为复杂等。

针对以上问题，本文提出了基于 DBSCAN[3] 的地震电离层扰动异常数据检测方法，聚类处理不同电离层扰动数据，计算地震电离层扰动数据相似度后利用 DBSCAN 方法提取相关的环境特征集，通过特征集实现地震电离层扰动异常数据检测。实验验证了所提方法能够快速、准确检测出地震电离层扰动异常数据。

1　地震电离层扰动异常数据分析

1.1　地震电离层扰动数据相似度划分

在实际的应用过程中，获取地震电离层扰动数据的测量值(http://www.usgs.gov)，上述的数据测量值所构建的向量代表现阶段环境的特点。在设定的时间段内，能够将时间划分成无数个时间片，将不同时间片中获取相对稳定的环境特征点集称之为环境特征[4-5]。将其作为研究的依据，并在其中加入数据挖掘的相关思想，将在具体环境中采集的特征点看作是需要调整的数据集。一般情况下，环境特征在较短的时间内不会发生较大的变化。所以，能够将无数的时间特征设定为时间聚类的簇，通过不同的环境特征获取相应的簇的特征。

DBSCAN 算法是一种较为典型的空间聚类方法[6]，具有以下几个方面的优势：

(1) 采集得到的数据整合速度快；

(2) 聚类速度相比其他方法较快；

(3) 能够有效抑制噪声；

(4) 实现任意空间的数据聚类；

空间中的随机数据集代表空间中的任意点，以下给出算法中相关参数：

(1) 空间中的随机点都是以 p 为中心点，以 ε 为半径的圆区域内包含的集合，ε 的取值一般情况需要由用户设定。

(2) 空间中不同数据集中最小个数用 M 表示，将该点称为核心点[7-8]。

(3) 两个点直接密度可达，使其满足相应的约束条件：

① 如果 p 在点 q 的领域中，则 $p \in N_s(q)$；

② 如果 q 为核心点，则 $|N_s(q)| \geqslant M$。

具体的结果如图 1 所示：

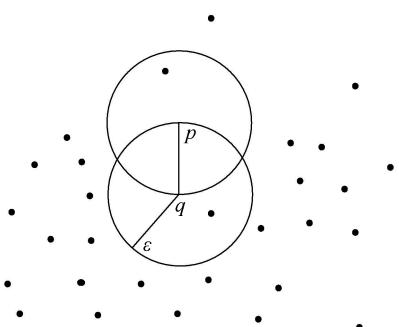

图 1　两点直接密度可达

Fig.1　Direct density reachable of two points

(4) 点 p 从点 q 密度可达，如果存在一个点链 $p_1, \cdots, p_n$，其中 $p_1 = q, p_n = n$ 并且从 $p_i (i = 1, \cdots, n-1)$ 直接密度可达，利用图 2 详细给出：

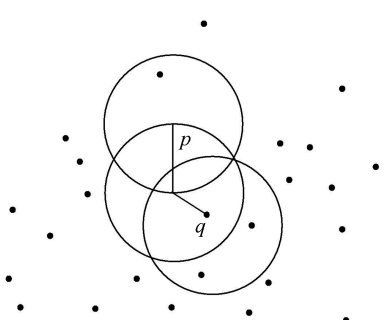

图 2　两点密度可达

Fig.2　Density reachable of two points

(5) 两点之间的密度是相互连接的[9]，如果存在点 o，使两点通过点 o 都是密度可达的，具体如

图 3所示：

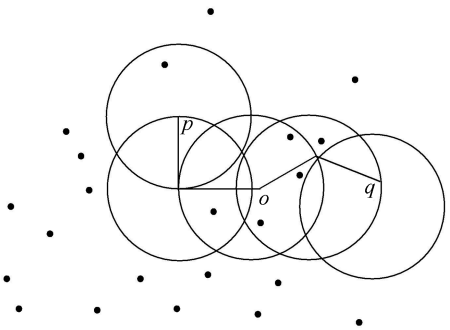

图 3 两点密度可连接

Fig.3 Density connectable of two points

1.2 地震电离层扰动异常数据检测

本文提出了基于 DBSCAN 的地震电离层扰动异常数据检测方法，具体可划分为两个步骤，分别为：训练阶段和检测阶段[10]。训练阶段可以实现得到环境特征集的目的，需要对相关的异常数据进行扰动训练。而测试阶段是在训练阶段之后进行的，可更加容易实现检测。

图 4 为整个算法的具体计算流程。

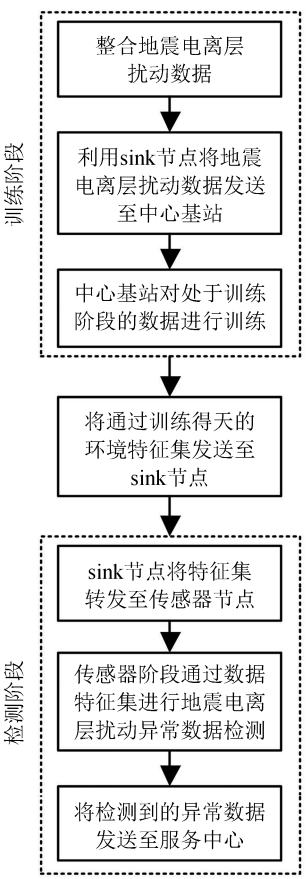

图 4 地震电离层扰动异常数据检测计算步骤

Fig.4 Steps for detection and calculation of seismic ionosphere disturbance anomaly data

（1）训练阶段：

传感器节点会通过相应的频率将地震电离层扰动数据发送至 S 节点，利用 S 节点向中心基站传送数据包 $D(n)$，持续到整个训练周期结束。各个数据包 $D(n)$ 主要是利用多个属性传感数据组成[11-12]，其中，m 代表属性的总数量。在中心基站中，使用 $DBSCAN$ 算法训练处于训练周期的数据，获取核心点集也就是环境特征集 $C(n)$，其中 n 代表核心点的数量。

（2）检测阶段：

节点每次都会采集到新的地震电离层扰动数据，将其与环境特征集 $C(n)$ 进行线性遍历对比，假设数据在中心点的区域内，则认为其是正常的数据；如果不在任意中心点的区域内，则认为其为异常数据，将最终结果发送至中心基站。

1.3 地震电离层扰动数据计算

针对地震电离层扰动数据的特殊性，利用突变系数进行数据划分。对时间序列进行全面观察与分析，能够获取如下相应结论：

如果设定时间序列为 $X=\{x_1,x_2,\cdots,x_n\}$，其中 x_i 代表在 X 时刻获取的数值，为实数，n 代表整个序列的长度。以下具体给出关于 X_i 的突变系数：

$$TB(x_i)=|(x_i/x_i-1)\times(x_i/x_i+1)| \tag{1}$$

通过分析式(1)可知，一个时间序列与它前后邻时序点比值相乘的绝对值就为它的突变系数。

由于整个时间序列中不能够缺少任何时间点的数值[13]，所以需要对不完整的数据集进行数据填充。其中，利用一定的值来填充数据集中缺失的数据将其称之为数据填充。

本文选取 KNN 算法进行计算，且设定欧式距离为整个方法的距离量度，则有：

$$d(i,j)=(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots|x_{ip}-x_{jp}|^2)^{\frac{1}{2}} \tag{2}$$

欧式距离是一种较为常用的距离定义，主要用来表示两个样本之间的相似度，以下给出具体的计算式：

$$d(x_i,x_j)=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2} \tag{3}$$

设定数据集为 $Y_X=\{x_i\mid i=1,2,\cdots,M\}$，其中 $x_i=(x_{i1},x_{i2},\cdots,x_{in})$ 中的 $x_{i1},x_{i2},\cdots,x_{in}$ 代表 x_i 的 n 个属性，如果两个样本之间的距离越小，则表明两个样本之间越相似，差异性越小[14]；反之，则说明两个样本之间存在较大的差异。

如果数据集 Y_X 中包含 k 个聚类子集 $X_1,\cdots,$

X_k，利用 $n_1,\cdots,n_k$ 代表不同子集中的样本总数；利用 $m_1,\cdots,m_k$ 代表不同子集的均值代表点。

误差平方与函数 E 是指不同样本点 p 到各个聚类中心 m_i 的距离平方和，具体计算过程为：

$$E=\sum_{i=1}^{k}\sum_{p\in X_i}\|p-m_i\|^2 \tag{4}$$

本文选取公式(4)作为 DBSCAN 方法性能的测度函数，在算法的聚类迭代阶段，E 的取值在逐渐减小，如果 E 的取值无法再变小时，则说明此时测度函数达到最优，整个迭代过程停止。

1.4　DBSCAN 的地震电离层扰动异常数据检测

本文选取 DBSCAN 作为地震电离层扰动异常数据检测方法的主要原因为：传感器节点获取的环境特征与传统的分布形式不同，具有计算范围相对较小、密度较大等特点。DBSCAN 方法是一种经典的数据聚类方法，在处理规模较大的数据时，该方法具有较强的伸缩性以及较高的效率。如果结果簇的分布形式相对密集，则说明两个聚类之间差异相对明显时获取的效果较好；反之，则表明获取的效果相对较差。通过分析可知，聚类结果分布的好坏与特征点的分布形式存在密切关系。

利用 DBSCAN 方法对地震电离层扰动异常数据进行检测，首先对相应的数据点进行分簇；然后计算不同数据点到聚类中心之间的欧式距离。如果计算获取的结果高于现阶段的平均距离[15]，则需要将数据放入疑似异常点集中；将各个不同的距离进行对比，如果获取的计算结果低于质心距离标准差，则代表该组数据为正常数据；反之，计算结果高于质心距离标准差，则代表该组数据为异常数据，并将该数据的具体坐标点输出。

给出详细的计算步骤为：

(1) 对样本集进行分簇；

(2) 输出异常数据坐标点，即：

$$\text{for}=\{\text{computer average distance of cluster}_{\text{avg}}\} \tag{5}$$

分别计算不同簇内数据点到质心之间的欧式距离，则有：

$$d_{\text{avg}}=\text{if}(d(x_i,m)\geqslant d_{\text{avg}}) \tag{6}$$

将数据点放入疑似异常数据点集中，则：

$$\text{put}(x_i)\text{into}=\{\text{computer average distance between } x_i \text{ and } m\} \tag{7}$$

计算传感数据 x_i 到所在簇质心 m 的方差：

$$v_d^*=\{\text{Consider } x_i \text{ as outlier point output its coordinates}\} \tag{8}$$

式中：n 代表传感数据样本的规模；$d(x_i,m)$ 代表样本数据 x_i 到所在簇心 m 在之间的距离。

综上所述，完成了地震电离层扰动异常数据检测。

2　仿真实验

2.1　实验环境

为了验证基于 DBSCAN 的地震电离层扰动异常数据检测方法的综合有效性，需要进行仿真实验。实验环境为 Windows7，处理器为 4 核 2.2 GHz IntelCore i7，内存为 8 GB 1333 MHz DDR3。实验选取在监测采集的 200 组数据(http://www.usgs.gov)作为测试样本，其中有 18 组数据为异常数据。

2.2　运行时间对比结果

选取 k-means 算法、模糊聚类算法两种传统方法作为对比方法进行仿真实验，以下对比不同方法的运行时间，具体的对比结果列于表 1。

表 1　不同方法运行时间对比结果

Table 1　Comparison of running time of different methods

测试样本数量/个	运行时间/s		
	所提方法	k-means 算法	模糊聚类算法
20	5	12	23
40	7	14	29
60	8	19	34
80	11	22	36
100	12	25	41
120	14	26	47

通过分析表 1 可知，随着测试样本数量的不断增加，不同方法的运行时间也在不断发生变化，其中所提方法的运行时间相比传统方法要明显低很多。其主要原因在于所提方法引用了 DBSCAN 方法进行异常数据检测，该方法具有较高的速率，所以有效降低了所提方法的运行时间。

2.3　误报率对比结果

表 2 为两种方法的数据误报率对别结果。

表 2　不同方法误报率对比结果

Table 2　Comparison results of false alarm rate of different methods

测试样本数量/个	数据误报率/%		
	所提方法	k-means 算法	模糊聚类算法
50	1.58	3.27	10.25
100	1.54	3.69	11.85
150	1.47	3.91	12.47
200	1.51	3.82	13.58
250	1.52	3.81	15.84
300	1.50	3.59	16.44

通过表 2 相关的实验数据可知，所提方法的误报率要远远低于传统方法。k-means 算法相比于模

糊聚类算法更为可靠。

2.4 检测率、漏报率对比结果

将测试样本分别输入传统方法、所提方法中进行仿真实验，获取的具体结果列于表 3、表 4。

表 3 不同方法漏报率对比结果

Table 3 Comparison results of missing alarm rate by different methods

聚类簇个数	所提方法/%	k-means算法	模糊聚类算法/%
2	2.58	2.32	5.69
3	2.11	3.57	5.78
4	2.51	2.95	5.84
5	1.85	4.62	5.99
6	1.54	3.28	6.25
7	1.66	3.19	6.17

表 4 不同方法检测率对比结果

Table 4 Comparison of detection rates by different methods

聚类簇个数	所提方法/%	k-means算法/%	模糊聚类算法/%
2	98.52	87.56	65.25
3	90.12	83.47	67.41
4	89.45	86.26	74.36
5	95.63	79.58	69.58
6	97.41	82.6	71.58
7	98.88	80.17	69.87

通过分析表 3、表 4 可知，在不同的训练簇数量下，相比传统方法，所提方法拥有更高的检测率以及较低的漏报率。

3 结语

针对传统方法存在的一系列问题，设计并提出一种新的检测方法。实验表明，所提方法相比较传统方法有效减少异常数据检测运行时间，误报率和漏报率大大降低，提升整个算法的检测率。

数据来源：

(1) 现阶段异常检测的方法有很多，但是本文只针对 DBSCAN 方法展开了深入研究，未来阶段需要将其他的聚类方法与测量建模相结合，进行深入研究。

(2) 本文选取地震电离层扰动数据作为实验数据，并没有将新的方法应用到复杂的数据中。所以研究的新方法是否能够成功应用于较为复杂的工程建模问题上，是未来阶段需要研究的主题。

参考文献(References)

[1] 李少波，孟伟，璩晶磊.基于密度的异常数据检测算法GSWCLOF[J].计算机工程与应用，2016，52(19)：7-11.
LI Shaobo，MENG Wei，QU Jinglei. GSWCLOF：Density-based Outlier Detection Algorithm on Data Stream[J].Computer Engineering and Applications，2016，52(19)：7-11.

[2] 周培培，丁庆海，罗海波，等.基于 DBSCAN 聚类算法的异常轨迹检测[J].红外与激光工程，2017，46(5)：230-237.
ZHOU Peipei，DING Qinghai，LUO Haibo，et al. Trajectory Outlier Detection Based on DBSCAN Clustering Algorithm[J]. Infrared and Laser Engineering，2017，46(5)：230-237.

[3] 田增山，王向勇，周牧，等.基于 DBSCAN 子空间匹配的蜂窝网室内指纹定位算法[J].电子与信息学报，2017，39(5)：1157-1163.
TIAN Zengshan，WANG Xiangyong，ZHOU Mu，et al. DBSCAN Based Subspace Matching for Indoor Cellular Network Fingerprint Positioning Algorithm[J].Journal of Electronics & Information Technology，2017，39(5)：1157-1163.

[4] 罗钦芳，丁国维，傅馨，等.基于“多层次分类”方法的异常 P2P 网贷借款识别[J].管理工程学报，2017，31(3)：201-209.
LUO Qinfang，DING Guowei，FU Xin，et al.Detecting Anomaly Ioans on P2P Lending Platform：Based on Hierarchical Classification Method[J].Journal of Industrial Engineering and Engineering Management，2017，31(3)：201-209.

[5] 李江苏，梁燕，王晓蕊.基于 POI 数据的郑东新区服务业空间聚类研究[J].地理研究，2018，37(1)：145-157.
LI Jiangsu，LIANG Yan，WANG Xiaorui.Spatial Clustering Analysis of Service Industries in Zhengdong New District Based on POI Data[J].Geographical Research，2018，37(1)：145-157.

[6] 杨先圣，姜磊，彭雄，等.基于大数据的异常检测方法研究[J].计算机工程与科学，2018，40(7)：1180-1186.
YANG Xiansheng，JIANG Lei，PENG Xiong，et al.A New Outlier Detection Method Based on Large Data[J].Computer Engineering and Science，2018，40(7)：1180-1186.

[7] 孙小磊，郑华，李晖，等.基于改进 DBSCAN 算法的变压器不良漏抗参数辨识[J].电力系统自动化，2017，41(9)：96-101.
SUN Xiaolei，ZHENG Hua，LI Hui，et al.Bad Data Identification for Leakage Reactance Parameters of Transformer Based on Improved DBSCAN Algorithm[J]. Automation of Electric Power Systems，2017，41(9)：96-101.

[8] 刘万军，秦济韬，曲海成.基于改进单类支持向量机的工业控制网络入侵检测方法[J].计算机应用，2018，38(5)：1360-1365，1371.
LIU Wanjun，QIN Jitao，QU Haicheng.Intrusion Detection Algorithm of Industrial Control Network Based on Improved One-class Support Vector Machine[J]. Journal of Computer Applications，2018，38(5)：1360-1365，1371.

[9] 王平，于祥春.基于密度聚类的能耗数据采集网关设计[J].电子技术应用，2018，44(6)：35-38，43.
WANG Ping，YU Xiangchun.Design of Power Information Collection Gateway Based on Density Clustering[J].Application of Electronic Technique，2018，44(6)：35-38，43.

[10] 江航，卢暾，顾寒苏，等.一种动态实时高校建筑能耗异常检测方法[J].计算机工程，2017，43(4)：15-20，27.
JIANG Hang，LUTun，GU Hansu，et al.A Dynamic and Real-time Outlier Detection Method forEnergy Consumption of Campus Building[J].Computer Engineering，2017，43(4)：15-20，27.

[11] 王志坚，王斌会.稳健改进的 AO 型异常点检测法在金融时序中的应用[J].数理统计与管理，2016，35(2)：369-380.
WANG Zhijian，WANG Binhui. The Application of Robust-improved AO Detect Method in Financial Time Series[J].Jouranl of Applied Statistics and Management，2016，35(2)：369-380.

[12] 沈琰辉，刘华文，徐晓丹，等.基于邻域离散度的异常点检测算法[J].计算机科学与探索，2016，10(12)：1763-1772.
SHEN Yanhui，LIU Huawen，XU Xiaodan，et al.Outlier Detection Algorithm Based on Dispersion of Neighbors[J].Journal of Frontiers of Computer Science & Technology，2016，10(12)：1763-1772.

[13] 许刚，王展，臧大伟，等.基于链路状态数据库的数据中心网络异常检测算法[J].计算机研究与发展，2018，55(4)：815-830.
XU Gang，WANG Zhan，ZANG Dawei，et al.Anomaly Detection Algorithm of Data Center Network Based on LSDB[J].Journal of Computer Research and Development，2018，55(4)：815-830.

[14] 胡石，李光辉，冯海林.基于 Top-k(σ)的无线传感器网络异常数据检测算法[J].南京大学学报(自然科学版)，2016，52(2)：261-269.
HU Shi，LI Guanghui，FENG Hailin. Top-k(σ)outlier Detection Algorithm for Wireless Sensor Networks[J].Journal of Nanjing University(Natural Sciences)，2016，52(2)：261-269.

[15] 胡琪，蒋朝惠，吕晓丹.基于进化算法的社会网络数据 k-匿名发布[J].贵州大学学报(自然科学版)，2016，33(1)：89-93.
HU Qi，JIANG Chaohui，LÜ Xiaodan. Evolutionary Algorithm for Social Network Data Publication Based on K-anonymity[J].Journal of Guizhou University(Natural Sciences)，2016，33(1)：89-93.