

深度聚类算法在 SuperDARN 雷达目标回波分类中的应用*

孔 星 刘二小 陈烽聚 乔 磊

(杭州电子科技大学通信工程学院 杭州 310018)

摘 要 SuperDARN 雷达目标回波中通常包含多种类型散射的回波, 例如电离层不规则体回波、地面/海面散射回波、极区中层夏季回波以及流星余迹回波等. 利用 SuperDARN 采集的电离层回波制作的电离层对流图对于空间天气研究具有重要意义. SuperDARN 接收到的电离层回波通常会与地面海面的散射回波混淆, 从而造成绘制的电离层对流图不准确, 因此对于 SuperDARN 目标回波进行聚类分析具有重要意义. 本文首次将基于自动编码器网络的图嵌入深度聚类算法应用于 SuperDARN 目标回波数据, 有效地对 SuperDARN 回波数据进行了分类. 此外, 还将该模型与传统算法和机器学习聚类算法进行了比较. 该模型在样本数据中的应用表明, 深度聚类算法能够捕捉到回波数据的深层结构特征, 提高了回波聚类的准确性.

关键词 SuperDARN 雷达, 深度聚类, 自动编码器模型, *K*-means 聚类

中图分类号 P352

Application of Deep Clustering Algorithm in Target Echo Classification of SuperDARN Radar

KONG Xing LIU Erxiao CHEN Fengju QIAO Lei

(College of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018)

Abstract SuperDARN radar target echoes usually contain echoes of various types of scattering, such as ionospheric irregularities, ground/sea scatter echoes, polar mesosphere summer echoes and meteor trail echoes. The ionospheric echoes collected by SuperDARN radar are used to map the ionospheric convection to study the large-scale dynamics of the magnetosphere-ionospheric system, which is of great significance for space weather observation and exploration. In general, the scattered ionospheric echoes received by SuperDARN are often mixed with the scattered echoes from the ground or the sea, resulting in inaccurate ionospheric convection maps. Therefore, cluster analysis of SuperDARN target echoes is of great significance. Traditionally, ground/sea scattering is determined by the lower threshold of the combination of velocity and spectral width, but the scope of use of this method is limited, and there is ambi-

* 国家重点研发计划项目 (2018YFC1407304, 2018YFC1407300, 2022YFC2807205) 和国家自然科学基金项目 (41974185) 共同资助

2023-11-23 收到原稿, 2023-12-21 收到修定稿

©The Author(s) 2024. This is an open access article under the CC-BY 4.0 License
(<https://creativecommons.org/licenses/by/4.0/>)

guity for the echoes in the mid-latitude region. In order to avoid the influence of latitude conditions and reduce the omission of useful information in the data, multiple data features of the radar target echo are collected as much as possible, such as line-of-sight Doppler velocity, the spectral width, backscatter power and the elevation angle of arrival. In this paper, the graph embedding deep clustering algorithm based on autoencoder network is applied to SuperDARN target echo data for the first time, and SuperDARN echo data is effectively classified. In addition, two different types of machine learning clustering algorithms are introduced to compare with this model. The deep clustering model, traditional classification algorithm and machine learning clustering algorithm are applied to the same echo data set, and the clustering results of different clustering algorithms are compared. The application of different clustering models on sample data sets and the evaluation of clustering indexes show that the deep clustering algorithm can capture the deep structural features of the echo data, effectively compress and reduce the dimensionality of the high-dimensional data set, make full use of the useful information in the data set, and improve the precision of the target echo data clustering of SuperDARN radar.

Key words SuperDARN radar, Deep clustering, Automatic encoder model, K -means clustering

0 引言

电离层等离子体对流是空间天气的重要特征之一, 通常在空间天气建模及预报时, 需要明确由日地耦合过程引起的大尺度对流运动, 即电离层对流, 通过观测极区等离子体对流检测来自行星际空间的能量流入, 进而实现空间天气预报^[1]. 电离层对流在空间天气研究中占有重要的地位, 因此可靠的电离层对流图对于进一步研究空间天气是具有重要意义的.

对电离层对流进行连续监测的最全面的方法之一是使用超级双子极光雷达观测网 (SuperDARN)^[2,3]. SuperDARN 是一个由超过 40 台高频相干散射雷达组成的国际科学观测雷达网络, 其在高频带传输由多个脉冲组成的信号序列, 并从返回信号的自相关函数 (ACF) 中推导出信噪比、多普勒速度和谱宽等信息. SuperDARN 全球对流图是通过将 SuperDARN 雷达的视距速度拟合到以球面谐波表示的静电势函数的展开而产生的. 制图技术主要包括两个阶段: (1) 视距速度数据的网格化; (2) 网格化数据的最小二乘拟合, 其中的网格化过程是表示将数据预处理成更有利于拟合的形式^[4]. 制图首先需要去除雷达收集到的数据中的地面散射数据, 保留电离层散射数据. 因此绘制准确的电离层对流图, 干净纯粹的电离层回波数据是必不可少的.

在很多情况下, 雷达能够探测到的目标回波位置在距离较大的情况下会产生很大的误差, 从而造成回波来源混淆. 实际情况中的 SuperDARN 目标回波中

包含电离层回波、地面/海面回波、极区中层夏季回波以及流星余迹回波等, 其中地面/海面散射回波数据可用于估算电离层的临界频率, 并提供从发射机到接收机数千公里范围内的电离层状态, 用于遥感、目标检测和跟踪以及提供海上导航和海浪预报等; 极区中层夏季回波 (PMSE) 是夏季在较低电子温度下出现在高纬度中层高度范围的很强的雷达回波, 可用于观测和研究极区中层物理机制^[5]; 流星余迹回波可以用于流星余迹通信实验以及空气动力学研究; 电离层回波用于绘制电离层对流图, 是研究空间天气很重要的依据. 因此, 不同类别的回波都拥有自己的独特用途, 对于混合回波进行有效的聚类分类是非常有必要的.

以往利用 SuperDARN 研究电离层对流均使用的是电离层不规则体回波, 该回波的提取和分类主要是基于回波多普勒速度及谱宽特征. 关于回波多普勒特征的提取, 先前的研究描述了获取多普勒速度和谱宽的标准方法^[6], 其中速度来自多普勒频率, 而多普勒频率由自相关函数 (ACF) 相位的线性最小二乘拟合确定, 谱宽由功率与高斯洛伦兹谱的拟合得出. 但是这种标准方法的前提是假设 SuperDARN 所收集到的后向散射回波信号是单一来源, 若将推导出的回波数据直接用于电离层不规则体分析是会出现很大误差的. 实际上, SuperDARN 雷达采集到的回波数据包含不止一个目标来源, 其中最常见的情况是电离层源叠加到地面回波源上^[7]. Schmidt^[8] 提出了一种多信号分类方法 (MUSIC), Barthes 等^[7] 证实该方法

适用于 SuperDARN 目标回波数据, 并将其应用于分离不同速度和谱宽的目标回波, 提高了电离层对流测量的质量. 然而 MUSIC 算法对目标回波来源的分类能力会受到频谱宽度的限制. 随着谱宽的增加, 分离难度也会越来越大, 需要大量的附加计算, 使得该方法更加地复杂. Chisham 和 Pinnock^[9] 开发了一种基于高频传播模式的经验识别方法, 可以从回波的 Range-Gate 和 Velocity 分布中区分地面散射回波和电离层散射回波, 并且消除其他不需要的后向散射的干扰回波. 虽然基于 Range-Gate 和 Velocity 的滤波能够去除一些杂波的污染, 但从长远来看, 需要添加更多其他物理因素的叠加约束, 例如功率、谱宽、速度误差以及谱宽误差等信息. Ponomarenko^[10] 等利用 Blanchard-Bake 算法成功检测到混合回波中叠加的大部分地面散射回波, 并将其与其他回波分离, 但即便如此, 混合回波中仍含有未及时分离的目标回波, 导致后续研究遗漏掉大量的有用信息. Ponomarenko 等^[11] 开发并测试了一种分离电离层和地面散射分量的双分量拟合算法. 该方法能够成功地反卷积电离层散射和地面/海面散射模式, 并从混合回波中获得可靠的电离层速度估计. Blanchard 等^[12] 则制订了更加具体、更加严格的区分地面散射和电离层散射信号的标准, 也是目前 SuperDARN 雷达目标回波使用的传统分类方法, 即

$$|v_D| \leq 30 - \frac{1}{3} \times |w_s|. \quad (1)$$

其中, $|v_D|$ 为多普勒速度, 单位 $\text{m} \cdot \text{s}^{-1}$; $|w_s|$ 为谱宽. 该公式能够显著降低地面散射信号被误分类为电离层散射信号的概率. 值得注意的是, 该方法偏向于将低速电离层散射归类为地面散射, 这样也会造成一定程度的混淆, 虽然该公式法的分类更加的简单直观, 可解释性强, 但基于公式的分类法在处理复杂和非线性问题时存在局限性. 因为公式法的表达能力有限, 无法灵活地处理复杂的数据模式和数据关系, 因此不确定是否有回波数据遗漏的情况发生, 或者由于天气等外部条件约束, 会导致出现数据误判, 并且基于公式的分类法不适用于大量数据, 大数据中的样本点集数量庞大, 会使得公式计算的复杂度过高而导致效率低下. Ribeiro 等^[13] 提出将深度优先搜索模型应用于 SuperDARN 数据的时间或空间维度上, 根据速度比将数据归类为地面回波或者电离层回波. 但该方法利

用的是窗函数过滤器, 这样做会使得时空分辨率降低. Burrell 等^[14] 针对 SuperDARN 回波提出了基于仰角的虚拟高度经验模型. 该技术需要可靠的仰角测量, 但目前该技术提供的仰角测量数据仍然存在偏差, 且模型只考虑单波束和时间的空间变化, 有多处的条件限制, 可用的范围较小.

实际上, 对于混合且未标记的数据, 有一种被称为聚类的机器学习算法, 目的是根据数据样本的潜在特征对其进行分类. 聚类是一种典型的无监督学习算法, 该算法从混淆的无标签数据中寻找具有共同内在结构特征的数据, 根据其内部相似度将混合样本数据分成多个类别, 为数据打上标签. 在传统的机器学习聚类算法中, 根据分类标准的不同有多种类型的聚类方法, 包括 K 均值聚类、mean shift 聚类、DBSCAN 聚类以及高斯混合聚类等. SuperDARN 雷达所收集的后向散射回波也同样属于无标签的混杂数据, 回波中包含有多个不同信号来源的回波, 因此聚类的机器学习算法也同样适用于 SuperDARN 聚类回波数据. 为了减少数据有用信息的遗漏以及降低数据点误判的概率, 需要尽可能多地收集数据的特征数量, 构建高维度特征数据集, 合理利用数据集的多个特征信息来做更加准确的分类, 提高数据分类的准确性. 面对高维数据集, 传统的机器学习算法无法获得理想的聚类结果, 为了改善传统机器学习算法的弊端, 本文采用基于深度学习的聚类算法 (简称深度聚类), 该类算法可以处理高维数据集, 并且该类算法已广泛应用于计算机视觉^[15]、语音识别^[16]、自然语言处理^[17] 和图像分割^[18,19] 等. 深度聚类算法将深度学习与聚类算法相结合, 挖掘庞大数据集中最显著的特征, 已成为无监督学习的研究热点. 本文从信号回波数据本身聚类分类的角度对 SuperDARN 混合回波数据进行分类, 从内部数据结构特征出发, 忽略外部因素, 例如季节环境, 天气状况等条件的约束, 首次尝试将深度聚类算法应用于 SuperDARN 目标回波聚类的研究中.

1 数据与预处理

1.1 数据集

所使用的数据集来自 2014 年 3 月 3 日 SuperDARN 雷达网中的昭和站 (SYE) 雷达以及 2014 年 12 月 4 日的中山站雷达数据 (ZHO)^[20]. 基于 Super-

表 1 数据集
Table 1 Data-Set

Radar	Day	Features
SYE (Data-Set1)	2014-03-03	一天的时间(s), Range-Gate, 回波强度, 多普勒速度, 多普勒谱宽, 雷达自动处理 2π 模糊度 (使相位处于 $-\pi \sim \pi$ 之间), 仰角, 速度误差, 谱宽误差, 初始功率 p_0 , 频率
ZHO (Data-Set2)	2014-12-04	一天的时间(s), Range-Gate, 回波强度, 多普勒速度, 多普勒谱宽, 雷达自动处理 2π 模糊度 (使相位处于 $-\pi \sim \pi$ 之间), 仰角, 速度误差, 谱宽误差, 初始功率 p_0 , 频率

ADRN 雷达的后向散射回波, 通过拟合雷达回波的自相关函数计算回波的功率、视线方向多普勒速度和谱宽信息. 这种拟合是通过 FITACF 程序进行的, FITACF 主要有两个版本, 本文使用的是雷达软件包 RST4.7 中默认的 FITACF2.5^[21] 算法. 具体数据集见表 1.

基于昭和站以及中山站雷达采集到的数据, 为了降低出现数据误判的概率, 减少导致误判有用信息所产生的遗漏, 提取回波中的 11 个数据特征用于数据处理, 使其构成一个二维数组形式的结构化数据集, 作为样本集.

1.2 数据预处理

使用到的算法是基于无监督数据的, 数据集中不含有类型标签, 因此数据集不划分为训练数据集以及测试数据集. 由于从雷达收集到的回波数据中存在有 9.99, 99.99, 999.99 这样的值, 还存在有 0 值以及与 inf 类似的无效数据, 不利于后续数据有效特征的提取, 因此将数据集中的该类无效样本进行删除处理. 同时对二维格式样本数据集设置小数点后保留 4 位小数, 并且使其数组格式规范化, 以利于后续聚类后标签的赋予.

2 算法

2.1 构建图相似度矩阵

基于二维数组形式的数据集, 选用余弦相似度衡量数据点之间的相似性, 构建图相似度矩阵作为自动编码器模型的输入.

图是一种用来表示对象之间关系的数据结构, 其由顶点和连接顶点的边共同组成. 数据集中的数据点由顶点表示, 将顶点连接起来的是边 (见图 1).

在图 1 中, $G(V, E)$ 表示构建图结构, 其中 $V = \{v_1, v_2, \dots, v_m\}$ 为顶点集, 即数据集; E 为边的集合. 在对数据集计算图相似度之前, 需要先对不规则的数据集进行数据归一化处理.

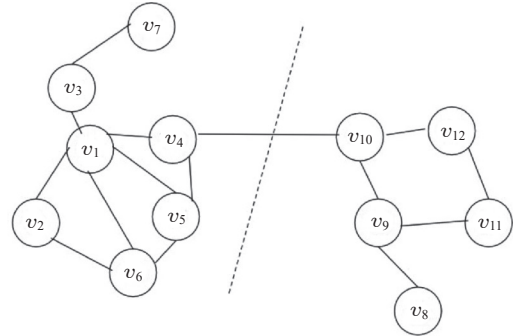


图 1 构建图结构

Fig. 1 Graph structure

2.1.1 归一化处理

由于数据集集中的 11 个不同特征的量纲不同, 因此对数据集集中的每一个非零特征进行归一化, 其公式如下:

$$u_{\text{norm}} = \frac{u}{|u|}. \quad (2)$$

式中 u_{norm} 代表数据集集中的每一个非零特征向量.

2.1.2 计算余弦距离

依据归一化后的数据构建图结构, 用余弦距离 $d_{i,j}$ 表示权重, 公式如下:

$$d_{i,j} = 1 - \frac{v_i \cdot v_j}{\|v_i\|_2 \|v_j\|_2}. \quad (3)$$

多个 $d_{i,j}$ 共同构成非方阵形式的邻接矩阵 $\mathbf{X}$.

2.1.3 构建方型矩阵

将非方型的邻接矩阵 $\mathbf{X}$ 转换成方型邻接矩阵 $\mathbf{W}$, 有

$$\mathbf{W} = \frac{\mathbf{X} \times \mathbf{X}(-1)}{2}. \quad (4)$$

2.1.4 计算拉普拉斯游走

计算自动编码器的输入, 即拉普拉斯的随机游走, 用于图形数据的聚类任务.

首先计算度矩阵 $\mathbf{D}$, 这里 $\mathbf{D}$ 为一个对角矩阵, 其中每个元素 $D(i, i)$ 表示顶点的度, 即与顶点相连的边

的权重的和, 其计算公式如下:

$$D(i, i) = \sum W(i, j). \quad (5)$$

计算拉普拉斯的随机游走, 有

$$L_n = I - D^{-1}W. \quad (6)$$

这里, L_n 为归一化拉普拉斯的随机游走, 即图相似度矩阵, 作为模型的输入.

2.2 K 均值聚类模型

K 均值聚类是一种经典的机器学习聚类算法, 其将数据集分为 K 个不同的簇, 每个簇中的样本都具有相似的特征. 在聚类每次迭代中, K 均值聚类通过计算每个数据点与当前质心的距离, 并将其分配到最近的质心所属的类别中, 根据该类别中的数据点更新质心的位置. 质心更新公式如下:

$$C_k = \frac{1}{|S_k|} \sum_{x \in S_k} x. \quad (7)$$

式中, C_k 为第 k 个类别的质心, S_k 为属于第 k 个类别的数据点的集合.

K 均值聚类的目标是 minimized 所有数据点与其所属类别质心之间的总距离. 通过最小化平方误差准则实现 (E), 即

$$E = \sum_{k=1}^K \sum_{x \in S_k} \|x - C_k\|^2. \quad (8)$$

式中, k 为类别的数量, S_k 为属于第 k 个类别的数据点的集合, C_k 为第 k 个类别的质心.

该算法通过不断迭代, 将样本点归类到离其最近的聚类中心, 直到收敛.

2.3 谱聚类模型

谱聚类是基于图论的聚类算法, 利用 2.1 节构建的拉普拉斯矩阵, 对其进行特征值分解、特征向量选择以及归一化. 特征值分解公式如下:

$$L \cdot v = \gamma \cdot v. \quad (9)$$

其中, L 为拉普拉斯矩阵, v 为特征向量, γ 为对应的特征值.

取 K 个特征向量作为输入数据, 代入上述公式, 将复数特征向量转换为实数. 最后进行归一化, 采用 L2 归一化方式. L2 归一化方式是指将每个特征向量的范数 (即向量的长度) 缩放为 1. 具体地, 对于一个特征向量 x , L2 归一化的计算公式如下:

$$\|X\| = \sqrt{x[0]^2 + x[1]^2 + \cdots + x[n-1]^2}. \quad (10)$$

获取到输入样本 X , 输入进 K 均值聚类, 完成谱聚类.

2.4 深度聚类模型

深度聚类是一种结合了深度学习网络与聚类算法的方法, 其利用深度学习特征表示来进行聚类指派^[22].

选用深度学习中的自动编码器模型用于压缩数据编码, 提取高维数据集中的明显特征并将输入的高维数据集转换为低维的编码数据. 自动编码器延续前馈神经网络的学习方法, 利用反向传播算法训练, 寻找数据最显著的特征, 学习数据的潜在表示形式.

自动编码器^[23] 是一种专门为无监督数据的表示而设计的神经网络, 该网络包含有输入层、隐藏层和输出层, 通常输入层与输出层维数相同, 而隐藏层的维度比输入层数据的维度小, 这有助于发现数据最显著的特征. 自动编码器网络由两部分组成, 一部分是编码器, 其函数为 $f_\theta = h(x, \theta)$, 将原始数据 x 映射到潜在表示 f_θ ; 另一部分是解码器, $g_\varphi = g(z, \varphi)$, 产生重构数据 g_φ , 其中 θ 和 φ 分别表示编码器和解码器的参数. 将自动编码器的编码器网络层数设置为 4 层, 解码器对称设置, 共 7 层. 网络的输入是由上述经数据预处理之后得到的余弦相似度矩阵. 隐藏层被设置为比输入层拥有更少的神经元, 在保证有用信息不流失的情况下使用较少的特征 (神经元) 表示输入数据, 从而达到降维和压缩的目的. 算法整体结构如图 2 所示, 深度自动编码器的输入是经过计算的归一化图的余弦相似度. 在编码器的输入和解码器的重构输出尽可能相似的前提下, 对余弦相似度进行压缩编码, 得到降维后的嵌入数据.

编码器的每一层都有一个阈值, 节点之间通过权值进行连接. 权值、阈值、输入均通过激活函数计算, 并利用输出与输入之间的误差进行反向传播, 对各层节点的阈值、权值等参数进行校正, 直至误差在可接受范围内. 在深度聚类算法中, 选择 Adam 算法进行模型优化, 并根据数据迭代更新目标损失函数变量. 与自编码器相比, 深度自编码器 (DEC) 增加了深度, 提高了特征学习能力. 局部保留约束的应用保留了原始数据的局部结构属性, 提高了聚类性能. 在本文使用的模型中, 自定义设置深度自动编码器网络的层数和参数, 定义输入数据的批量数, 并继续优化, 直到自动编码器输出与输入尽可能地相似, 得到压缩后的嵌入数据.

压缩后的数据拥有数据集本身最显著的特征, 然

后将嵌入数据输入进 K 均值聚类, 设置聚类簇数 K , 完成深度聚类, 简称 AE- K 均值算法。

3 分析与讨论

以 2014 年 12 月 SuperDARN 雷达昭和站以及中山站^[24] 观测到的回波数据为样本数据集, 对模型进行评估。鉴于先前研究中, 通常认为在 SuperDARN 雷达目标回波中最常见的情况, 即地面/海面回波和电离层散射回波的叠加, 因此人为设置聚类簇数 K 为 2, 即将混合回波默认分为两类, 一类是地面/海

面回波, 另一类是电离层回波。选用 2014 年 3 月昭和站雷达收集到的部分数据作为 Data-Set1, 将三种模型的聚类结果进行可视化, 对比传统公式法分类结果如图 3 所示。

图 3(a) 给出了昭和站雷达在功率、速度和谱宽中的分布。其中速度子图中的灰色数据点表示地面/海面散射回波, 除灰色标注以外的为电离层散射回波。从图中可以看出, 绿色方框框住的 Range-Gate 约在 20~40 处、时间约在 09:00 UT—11:00 UT 的数据集按照传统公式算法, 依据速度和谱宽分布特点被划分为了地面散射回波; 被橙色方框框住的数据,

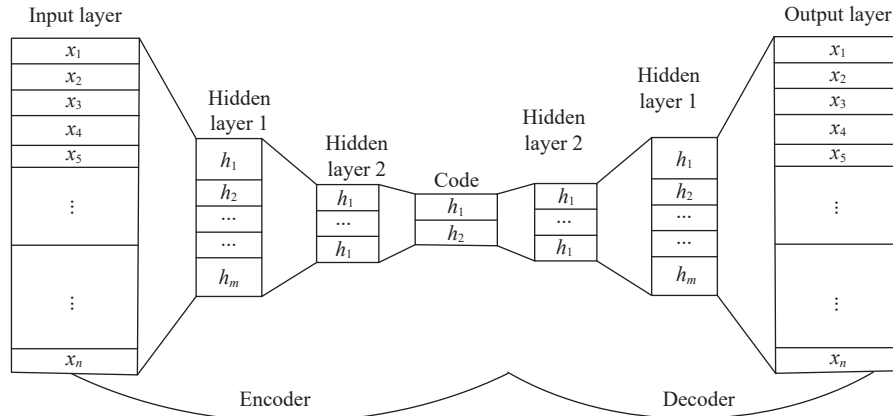


图 2 自动编码器网络结构

Fig. 2 Autoencoder network structure

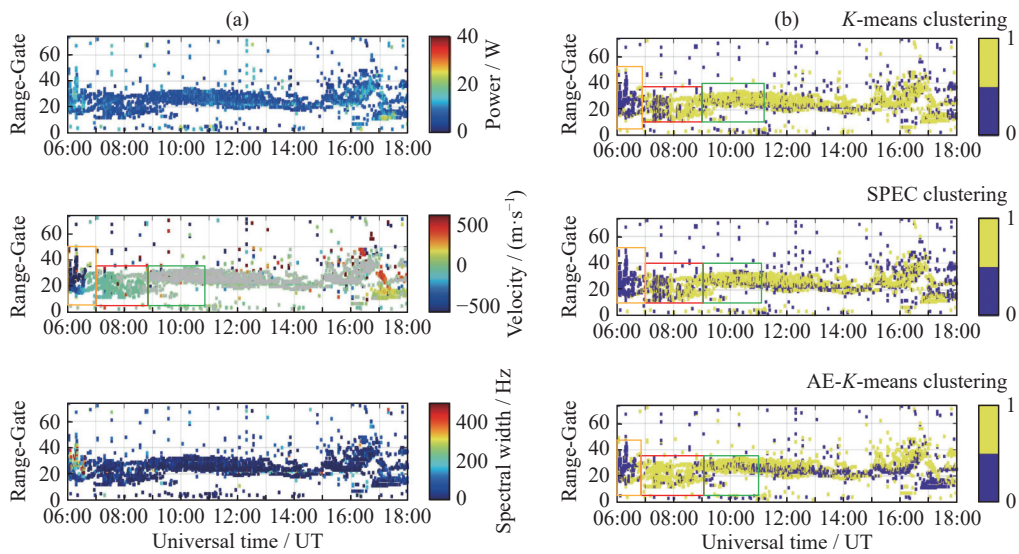


图 3 传统方法分类昭和站雷达数据 (a) 与三种聚类算法聚类昭和站雷达数据 (b) 结果可视化

Fig. 3 Visualization of data results from SYE Station classified by traditional methods (a), and three clustering algorithms (b)

Range-Gate 约在 $0 \sim 50$ 处, 时间约在 06:00 UT—07:00 UT 处, 有高速的特征, 传统分类算法将其划分到了电离层回波, 与绿色方框内的数据类型不一致; 而红色方框框住的数据即 Range-Gate 约在 $20 \sim 40$ 、时间约在 07:00 UT—09:00 UT 处的数据集, 观察其在速度和谱宽上的特点与绿色方框内的数据特征相似, 速度较低, 在几十 $\text{m}\cdot\text{s}^{-1}$ 的范围内, 谱宽较窄, 因此根据传统公式法两部分数据应该是同一类别. 而在图 3(a) 中, 传统 SuperDARN 雷达目标回波分类算法将红色方框内的数据和橙色方框内的数据归类为电离层回波, 与绿色方框内数据属于不同类别. 观察图 3(b) 中的聚类结果数据分类, 其中的 0, 1 分别代表两种不同的类别, 对比图 3(a) 中的数据类型, 可以将 0 类视为电离层回波, 1 类视为地面/海面回波. 观察 K 均值 Clustering 子图、Spec Clustering 子图, 比较红色方框内的数据, 都有不同程度的部分回波被划分至 0 类, 而 AE- K -means 聚类算法将绿色方框和红色方框内的几乎所有数据划分到同一类别即 1 类, 并且将橙色方框内的数据划分至 0 类.

基于上述聚类结果, 对图 3 所示的方框部分数据进行特征分析, 绘制特征分布直方图来验证两部分属性是否相似, 从而验证本文聚类算法的优越性, 三部分数据的特征分布直方图如图 4 所示.

从图 4(a)(b) 数据的分布直方图来看, 红色和绿色方框内的两部分数据集具有相同的低速和低谱宽, 速度基本都分布在每秒几十米内, 谱宽分布也相似, 并且在 Power 特征上也具有相同的分布特点, 而图 4(c) 给出的橙色方框内数据的特征分布直方图表现出了高速的特点, 与 (a)(b) 在速度分布中有很大的差别, 因此图 4 验证了上述的讨论, 即红和绿色方框两部分数据应该被划分为同一类型的回波, 结合数据特点以及传统算法分类结果应为地面散射回波, 而橙色方框的数据应该被划分至与红绿方框内的数据不同类型的回波, 结合高速特点以及传统分类结果应为电离层散射回波. 因此图 4(b) 聚类结果中的 AE- K 均值算法可以有效准确地对 SuperDARN 回波数据进行全面聚类, 相比于传统算法以及另外两种机器学习聚类方法, AE- K 均值能够有效减少有用数据的漏判.

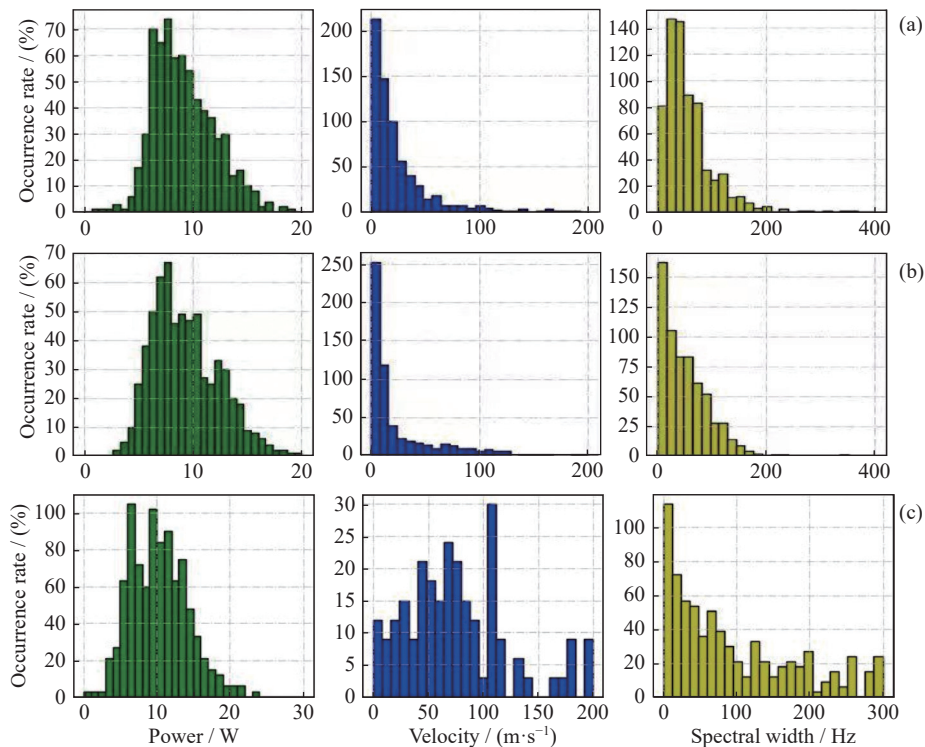


图 4 (a) 图 3(a) 红色方框内数据特征分布直方图, (b) 图 3(a) 绿色框内数据特征分布直方图, (c) 图 3(a) 橙色框内的数据特征分布直方图

Fig. 4 Histogram of feature distribution for data in red box (a), data in green box (b) and data in orange box (c) from Fig. 3 (a)

同样, 将其应用于 SuperDARN 中山站雷达, 提取中山站雷达同年 12 月的部分数据作为 Data-Set2, 将其输入进聚类算法, 生成聚类标签和聚类可视化结果如图 5 所示。

从图 5(a) 可以看出, 绿色方框的数据是传统算法分类后得到的地面/海面回波, 被标注为灰色, 剩下的部分数据包括红色方框以及橙色方框内的数据被划分为电离层回波。仔细对比红色方框的数据点, 在速度和谱宽图上, 拥有低速低谱宽的特点, 所以应该归类为地面/海面回波。分析图 5(b), 经三种聚类算法聚类后得到的分类结果, 对比传统分类算法的分类结果, 0 代表地面/海面散射回波, 1 代表电离层散射回波, 从图中发现仅有 AE-K 均值聚类将红色方框内的回波与绿色方框内的回波划分至同一类别, 并且与橙色方框内的回波不同类型。与上述昭和站雷达数据分析思路类似, 为了验证红色方框回波与绿色方框回波以及橙色方框回波特征特点, 绘制了三部分数据的特征分布直方图 (见图 6)。

从图 6(a)(b) 可以看出, 两部分数据同样具有类似的低速低谱宽特点, 符合地面/海面回波特征, 图 6(c) 表示的橙色方框内回波具有速度高的特征, 符合电离层散射回波特点, 从而验证了本文结果。总体而言, AE-K-means 算法具有聚类的有效性以及优越性, 对于 SuperDARN 雷达目标回波的聚类更加合理。

从聚类评价指标来看, 选取 Silhouette Coefficient (轮廓系数, R_{SI}) 以及 Calinski-Harabasz (判别指数, R_{CH}) 指数作为评价本文选取的三种聚类方法的指标, R_{SI} 为衡量样本的聚类紧密度和分离度的指标, R_{SI} 值的计算公式为

$$R_{SI}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \quad (11)$$

其中, $a(i)$ 为每个样本 i 与其他样本的平均距离, $b(i)$ 为样本 i 与最近其他簇中样本的平均距离, 对于所有样本的轮廓系数, 计算平均值, 即为整个数据集的轮廓系数。 R_{CH} 是通过计算聚类内部离散度和聚类间离散度指标来评估聚类的紧密度, 指数值越大表示聚类效果越好。 R_{CH} 值的计算公式为

$$R_{CH} = \frac{\sum(d_{S-B(c)})}{\sum(d_{S-W(c)})} \cdot \frac{n-k}{k-1}. \quad (12)$$

其中, n 为样本数, k 为簇的数量, $d_{S-W(c)}$ 为簇内样本的平均距离, $d_{S-B(c)}$ 为簇中心与整个数据集中心的距离。

基于 Data-Set1, 分别运行三种模型聚类数据, 计算得到的 R_{SI} 值和 R_{CH} 值见表 2。从表 2 可知, AE-K 均值算法的 R_{SI} 系数和 R_{CH} 指数要高于另外两种机器学习算法, 即说明 AE-K 均值聚类算法的聚类效果优于其他两种算法。

同时, 为证明聚类时人为设置 K 为 2 的准确性, 采用不同算法对 Data-Set1 分别计算了输入 K 值为

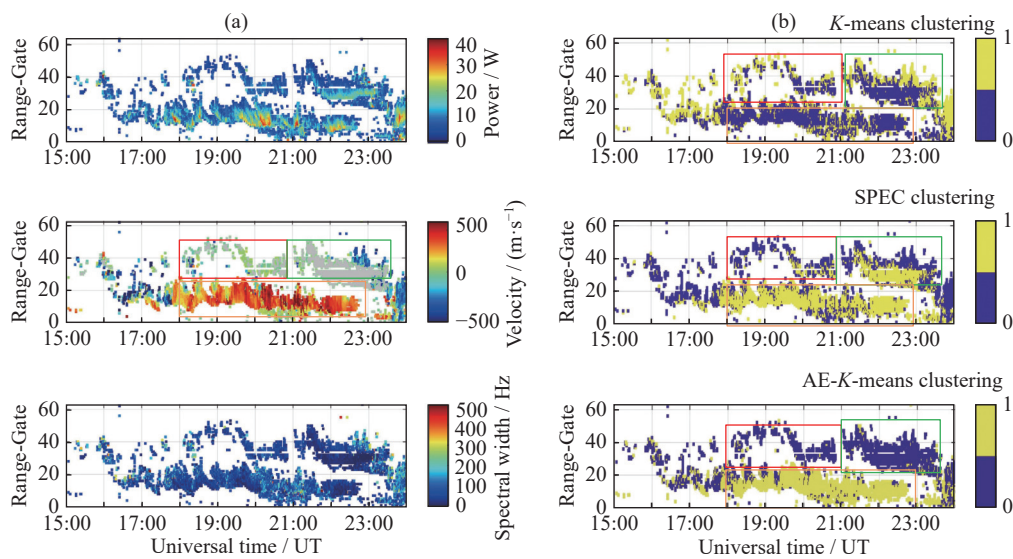


图 5 传统方法分类中山站雷达数据 (a) 与三种聚类算法聚类中山站雷达数据 (b) 结果可视化

Fig. 5 Visualization of data results from ZHO Station classified by traditional methods (a), and three clustering algorithms (b)

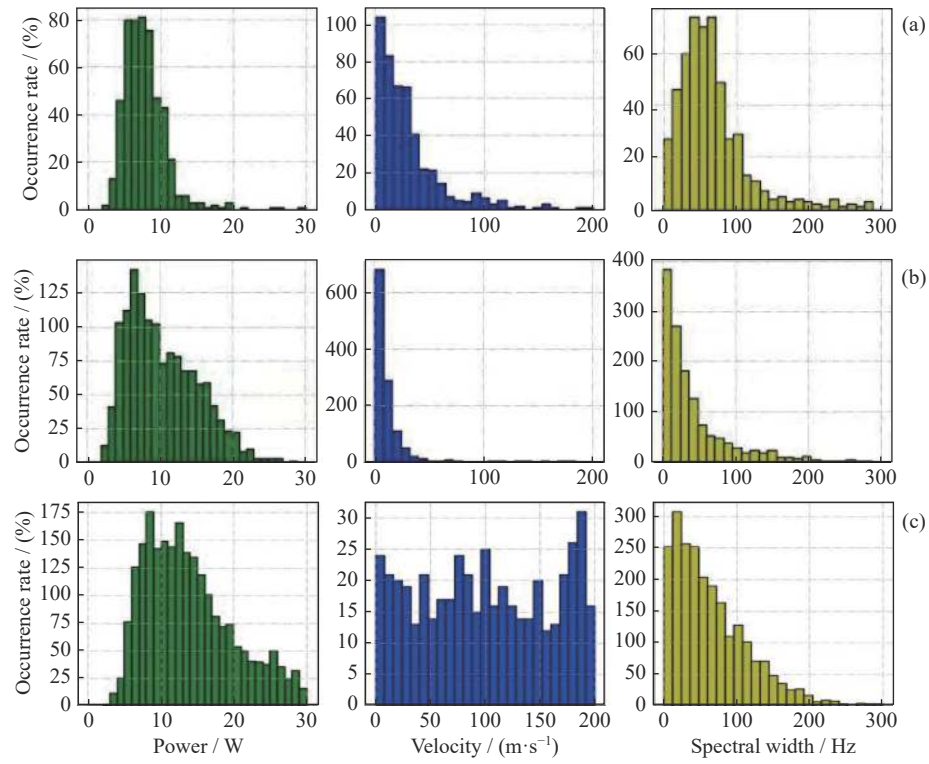


图 6 (a) 图 5(a) 红色方框内数据特征分布直方图, (b) 图 5(a) 绿色框内的数据特征分布直方图, (c) 图 5(a) 橙色框内的数据特征分布直方图

Fig. 6 Histogram of feature distribution for data in red box (a), data in green box (b) and data in orange box (c) from Fig.5(a)

表 2 三种聚类算法得到的 R_{SI} 值和 R_{CH} 值
Table 2 R_{SI} and R_{CH} values obtained by three clustering algorithms

Score	R_{SI}	R_{CH}
K -means	44.70	7449.96
SPEC	39.43	6376.27
AE- K -means	60.94	24421.17

表 3 不同算法设置不同 K 值聚类数据所得的 R_{SI} 值
Table 3 Different algorithms set the R_{SI} values obtained by clustering data with different K values

Class	2	3	4	5
K -means	44.70	43.65	45.26	37.92
SPEC	39.43	39.71	30.59	30.79
AE- K -means	60.94	52.70	55.97	54.37

2~5 的聚类 R_{SI} 值 (见表 3), 其对应折线的数值变化如图 7 所示.

从表 3 以及图 7 可以看出, 当 K 值取 2 时所对应的 R_{SI} 值是最大的, 因此证明先前设置 K 值为 2 是

合理有效的.

为进一步验证 AE- K 均值算法的稳定性, 基于 Data-Set1, 多次执行 AE- K 均值聚类算法, 统计观察该算法聚类后的两种类别的样本数量见表 4.

从表 4 可以看出, AE- K 均值聚类算法在第一次、第二次、第三次以及第四次和第五次运行时, 该算法不同聚类结果所拥有的样本数量相差不大, 证明了算法的稳定性.

以往的研究以及传统的 SuperDARN 回波分类中, 往往只考虑数据的二维或三维特征, 即最显著的速度和谱宽特征. 虽然速度和谱宽特征的确携带有最为重要的信息, 但其他特征也有包含影响分类的有用信息, 例如有些数据的误判和漏判现象即说明了以往算法的缺陷, 导致分类后的部分数据结果不够准确. 因此, 除上述重要的谱宽和速度特征外, 本文相对于传统算法还适当地提取了数据集中的 Range-Gate, Power, Time 等 11 个特征, 利用深度聚类算法处理高维数据的优越性对 SD 雷达回波进行数据分类处理. 根据 AE- K 均值算法聚类后的数据还可以分析其

特征值的重要贡献程度, 利用图 3(b) 的聚类结果数据, 绘制各特征值的贡献度, 对比传统算法分类后的数据不同特征贡献度结果如图 8 所示.

从图 8 可以清晰看出, 不同特征对聚类结果的贡献程度不同, 图 8(a) 给出了传统分类方法分类 SuperDARN 雷达目标回波数据时不同特征对分类结果的贡献度, 不可忽略的是有关速度和谱宽的信息尤为重要. 图 8(b) 给出了 AE- K 均值算法聚类回波数据时不同特征对聚类结果的贡献度, 其中除了速度和谱宽特征外, 图 8(b) 还表现出雷达数据处理时的 2π 模糊度以及仰角数据等信息也同样对聚类结果的影响很大. 为了聚类结果的准确性, 要尽可能地减少忽视这些隐藏的有用信息. 为使得聚类更加全面, 需要尽可能地收集数据中的有效信息, 适当选择多个数据特征来进行数据处理.

4 结论

SuperDARN 雷达所收集到的后向散射回波多是电离层回波与地面/海面回波的叠加, 绘制准确的电

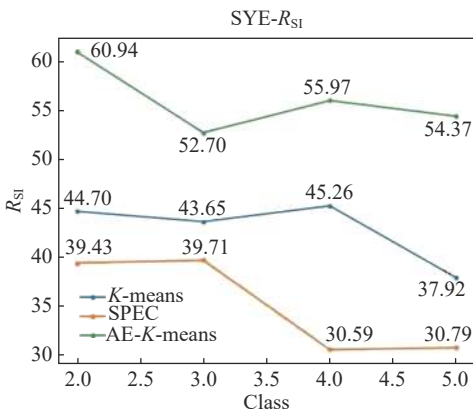


图 7 K 取 2~5 时的 R_{SI} 值变化

Fig. 7 Changes of R_{SI} value as K varies from 2 to 5

表 4 AE- K -means 算法的稳定性

Table 4 Stability of AE- K -means algorithm

频次	样本数量
1	3543, 6633
2	3798, 6378
3	3774, 6402
4	3531, 6645
5	3951, 6225

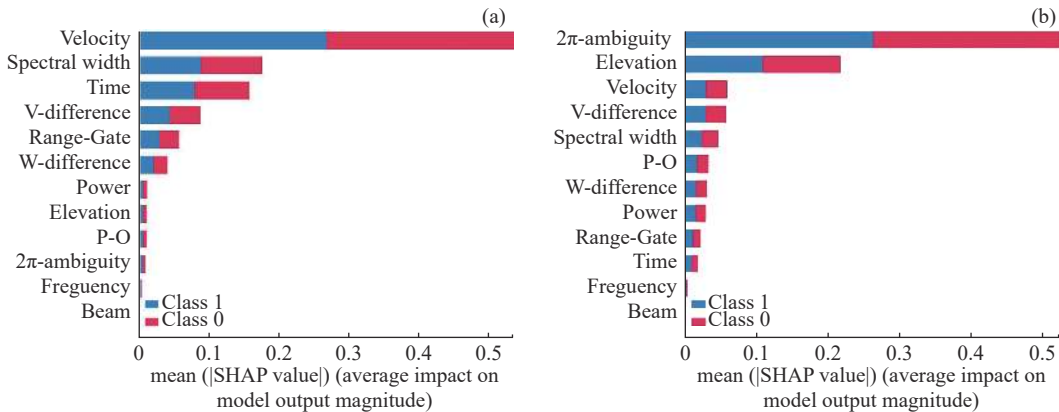


图 8 (a) 传统分类模型有贡献特征的 SHAP, (b) AE- K 均值模型有贡献特征的 SHAP

Fig. 8 Traditional method (a) and AE- K -means model (b) contribute to SHAP diagram interpretation of features

离层对流图需要干净准确的电离层回波数据, 针对如何获取到干净的电离层回波问题, 采用三种模型来对混合回波进行有效聚类, 分别是 K 均值聚类、谱聚类、AE- K 均值聚类, 同时将基于传统公式法分类得到的数据集做效果对比。

分别基于昭和站雷达以及中山站雷达从物理角度观察数据特征, 直观地分析 AE- K 均值聚类算法的优越性, 基于昭和站雷达的部分数据, 从 R_{SI} 和 R_{CH} 值聚类指标上评估 K 均值聚类、谱聚类以及 AE- K 均值聚类, 结果表明 AE- K 均值聚类相比于其他两种机器学习算法以及传统公式法更加有效, 对 SuperDARN 雷达目标回波的分类更加合理, 且性能明显优于另外两种模型。同时该深度模型在以后的混合回波多分类问题中也能提供一定的参考。

根据三种聚类模型在 SuperDARN 雷达回波数据的分类结果, 可以得到以下结论。

(1) 根据无监督聚类标准 R_{SI} 和 R_{CH} 来评估 K 均值聚类、谱聚类、AE- K 均值聚类, 其中 K -means 聚类和谱聚类的 R_{SI} 和 R_{CH} 值分别为 44.70, 7449.96 和 39.43, 6376.27, 而 AE- K -means 深度聚类的 R_{SI} 值和 R_{CH} 值是 60.94 和 24421.17, 显著高于前两种模型。

(2) 根据三种聚类模型的结果可视化图与传统公式法的分类结果图对比, 绘制分类后的数据特征直方图, 证明传统公式法以及另外两种机器学习的聚类算法都有一定程度的数据漏判及误判情况, 而 AE- K 均值聚类算法在聚类上的效果更好, 聚类分类更加全面。

(3) 总体而言, AE- K 均值聚类模型针对高维度的 SuperDARN 雷达混合目标回波数据集的聚类分类是更加有效且全面的, 深度学习网络能够学习到数据集集中的潜在隐藏结构, 减少有用信息的遗漏, 总体聚类分类效果更好。

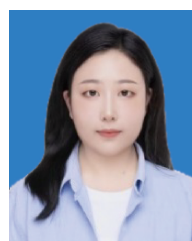
空间天气研究中会存在有许多不同的物理环境, 下一步工作将是在磁暴、亚爆和空间天气事件等不同条件下改进深度聚类算法。同时认为 SuperDARN 雷达目标回波中不仅仅有地面/海面回波和电离层回波, 同时还可能包含有极区中层夏季回波和流星余迹回波等, 鉴于所用数据集中可能其他回波所占比重较小, 因此忽略了其他回波的存在。在今后的研究中, 将致力于选用更复杂的混合回波做更多簇数的聚类, 提取不同种类回波, 使得每种回波都能发挥自己的用途。

参考文献

- [1] KUSTOV A V, PAPITASHVILI V O, SOFKO G J, *et al.* Dayside ionospheric plasma convection, electric fields, and field-aligned currents derived from the SuperDARN radar observations and predicted by the IZMEM model[J]. *Journal of Geophysical Research: Space Physics*, 1997, **102**(A11): 24057-24067
- [2] MARCUCCI M F, COCO I, MASSETTI S, *et al.* Echo occurrence in the southern polar ionosphere for the SuperDARN Dome C East and Dome C North radars[J]. *Polar Science*, 2021, **28**: 100684
- [3] GREENWALD R A. History of the Super Dual Auroral Radar Network (SuperDARN)-I: pre-SuperDARN developments in high frequency radar technology for ionospheric research and selected scientific results[J]. *History of Geo- and Space Sciences*, 2021, **12**(1): 77-93
- [4] CHISHAM G, PINNOCK M. Assessing the contamination of SuperDARN global convection maps by non-F-region backscatter[J]. *Annales Geophysicae*, 2002, **20**(1): 13-28
- [5] ULLAH S, LI H L, RAUF A, *et al.* Effect of ions on conductivity and permittivity in the Polar Mesosphere Summer Echoes region[J]. *Earth and Planetary Physics*, 2021, **5**(2): 196-204
- [6] VILLAIN J P, GREENWALD R A, BAKER K B, *et al.* HF radar observations of E region plasma irregularities produced by oblique electron streaming[J]. *Journal of Geophysical Research: Space Physics*, 1987, **92**(A11): 12327-12342
- [7] BARTHES L, ANDRÉ R, CERISIER J C, *et al.* Separation of multiple echoes using a high-resolution spectral analysis for SuperDARN HF radars[J]. *Radio Science*, 1998, **33**(4): 1005-1017
- [8] SCHMIDT R. Multiple emitter location and signal parameter estimation[J]. *IEEE Transactions on Antennas and Propagation*, 1986, **34**(3): 276-280
- [9] GREENWALD R A, BAKER K B, DUDENEY J R, *et al.* Darn/superdarn: a global view of the dynamics of high-latitude convection[J]. *Space Science Reviews*, 1995, **71**(71): 761-796
- [10] PONOMARENKO P V, WATERS C L, MENK F W. Factors determining spectral width of HF echoes from high latitudes[J]. *Annales Geophysicae*, 2007, **25**(3): 675-687
- [11] PONOMARENKO P V, WATERS C L, MENK F W. Effects of mixed scatter on SuperDARN convection maps[J]. *Annales Geophysicae*, 2008, **26**(6): 1517-1523
- [12] BLANCHARD G T, SUNDEEN S, BAKER K B. Probabilistic identification of high-frequency radar backscatter from the ground and ionosphere based on spectral characteristics[J]. *Radio Science*, 2009, **44**(5): RSS012
- [13] RIBEIRO A J, RUOHONIEMI J M, BAKER J B H, *et al.* A new approach for identifying ionospheric backscatter in

- midlatitude SuperDARN HF radar observations[J]. *Radio Science*, 2011, **46**(4): RS4011
- [14] BURRELL A G, MILAN S E, PERRY G W, *et al.* Automatically determining the origin direction and propagation mode of high-frequency radar backscatter[J]. *Radio Science*, 2015, **50**(12): 1225-1245
- [15] CHANG J F, LU Y, XUE P, *et al.* Iterative clustering pruning for convolutional neural networks[J]. *Knowledge-Based Systems*, 2023, **265**: 110386
- [16] CHUI K T. Driver stress recognition for smart transportation: applying multiobjective genetic algorithm for improving fuzzy c-means clustering with reduced time and model complexity[J]. *Sustainable Computing: Informatics and Systems*, 2022, **35**: 100668
- [17] DUTTA R, DAS N, MAJUMDER M, *et al.* Aspect based sentiment analysis using multi-criteria decision-making and deep learning under COVID-19 pandemic in India[J]. *CAAI Transactions on Intelligence Technology*, 2023, **8**(1): 219-234
- [18] WAGNER F H, DALAGNOL R, SÁNCHEZ A H, *et al.* K-textures, a self-supervised hard clustering deep learning algorithm for satellite image segmentation[J]. *Frontiers in Environmental Science*, 2022, **10**: 946729
- [19] ANITHA T, LATHA G P L, PRASAD M S. A proficient adaptive K-means based brain tumor segmentation and detection using deep learning scheme with PSO[J]. *Journal of Computational Science & Intelligent Technologies*, 2020, **1**(3): 9-14
- [20] LIU E X, HU H Q, LIU R Y, *et al.* Diurnal variation of the HF radar echoes at Zhongshan Station and the influence of geomagnetic activity[J]. *Chinese Journal of Geophysics*, 2012, **55**(9): 3066-3076 (刘二小, 胡红桥, 刘瑞源, 等. 中山站高频雷达回波的日变化特征及地磁活动的影响[J]. *地球物理学报*, 2012, **55**(9): 3066-3076)
- [21] RIBEIRO A J, RUOHONIEMI J M, PONOMARENKO P V, *et al.* A comparison of SuperDARN ACF fitting methods[J]. *Radio Science*, 2013, **48**(3): 274-282
- [22] KIM Y, HEILAND J. Convolutional autoencoders, clustering and POD for low-dimensional parametrization of navier-stokes equations[OL]. arXiv preprint arXiv: 2302.01278, 2023
- [23] HINTON G E, ZEMEL R S. Autoencoders, minimum description length and Helmholtz free energy[C]//Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver: Morgan Kaufmann Publishers Inc. , 1993
- [24] HU H Q, LIU E X, LIU R Y, *et al.* Statistical characteristics of ionospheric backscatter observed by SuperDARN Zhongshan radar in Antarctica[J]. *Advances in Polar Science*, 2013, **24**(1): 19-31

作者简介



孔 星 女, 1998 年出生于山东省济宁市, 现为杭州电子科技大学研究生, 主要研究方向是 SuperDARN 雷达目标回波数据处理.

E-mail: kongxing@hdu.edu.cn



刘二小 (通信作者) 男, 1984 年出生于内蒙古包头, 现为杭州电子科技大学通信工程学院副教授, 主要研究方向是高频雷达数据处理、极区空间环境特性分析及深度学习应用等.

E-mail: liuerxiao@hdu.edu.cn