计算机科学
COMPUTER SCIENCE

基于流线距离聚类的海洋数据向量场可视化

王朕, 杨政威, 高顺起, 张磊

引用本文

王朕, 杨政威, 高顺起, 张磊.基于流线距离聚类的海洋数据向量场可视化[J]. 计算机科学, 2023, 50(6A): 220300284-7.

WANG Zhen, YANG Zhengwei, GAO Shunqi, ZHANG Lei. Visualization of Ocean Data Vector Field Based on Streamline Distance Clustering [J]. Computer Science, 2023, 50(6A): 220300284-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220300284

基于流线距离聚类的海洋数据向量场可视化

王 朕[1,2] 杨政威[1] 高顺起[1] 张 磊[1]
1 天津财经大学理工学院 天津 300222
2 自然资源部海洋信息技术创新中心 天津 300171

摘 要 流线可视化是海洋向量场可视化的重要研究对象,其中流线种子点的数量和位置设定是基础,而生成流线的准确聚类以及类内代表性流线的选取是消除冗余流线造成的视觉混乱和遮挡问题的关键。文中提出将PDM距离作为流线聚类的相似性度量值,在流线端点聚类的基础上再进行流线精细聚类,有效解决端点聚类结果不准确的问题,提升了流线聚类的准确性。通过排序聚类后类内流线对的PDM距离值,提取中线和边界线进行流线重绘,减少了流线遮挡和杂乱现象。针对基于距离的流线聚类计算量大的问题,提出了MDS算法以提升计算速度。此外,采取临界点检测算法减少了流线生成过程中耗时的漩涡生成,进一步提升了计算速度。使用中国沿海的海洋流场数据进行实验,验证了算法的有效性和优越性,流线绘制效果良好。

关键词:聚类;向量场可视化;流线距离
中图法分类号 TP391

Visualization of Ocean Data Vector Field Based on Streamline Distance Clustering

WANG Zhen[1,2],YANG Zhengwei[1],GAO Shunqi[1] and ZHANG Lei[1]
1 School of Science and Technology,Tianjin University of Finance and Economics,Tianjin 300222,China
2 Marine Information Technology Innovation Center,Ministry of Natural Resources,Tianjin 300171,China

Abstract Streamlines visualization is an important research object of ocean vector field visualization,in which the setting of number and location for streamlines seed points is the basis. The key to the problem,how to eliminate the visual confusion and occlusion caused by multiple streamlines,is the accurate clustering of generated streamlines and the selection of appropriate representative streamlines within the clustering. In this paper,the PDM distance is proposed to be the similarity measurement,then the fine streamline clustering is implemented after once endpoints streamline clustering. The proposed method effectively solves the problem of inaccurate endpoint clustering results and improves the accuracy of streamline clustering. After sort of all the PDM distance within each clustering,we extract midline and two boundary lines to redraw the streamlines,so as to reduce the phenomenon of visual confusion and occlusion caused by multiple streamlines. For the problem of huge amount of calculation,the MDS algorithm is proposed to reduce dimension and accelerate computing speed. In addition,in order to further accelerate the calculation speed,the critical point detection algorithm is adopted to reduce the time-consuming vortex generation during the process of streamline generation. The effectiveness and superiority of our proposed method are verified by using ocean flow field data from China coast,and the drawing effect of streamline is good.

Keywords Clustering,Vector field visualization,Streamline distance

1 引言

海洋数据具有多样性、复杂性、动态性的特点,导致海洋研究人员很难通过直接观察数据来发现海洋环境蕴含的规律和特性。由于这些特性造成的研究困难可以通过向量场可视化的方法进行改善,即可以将抽象的数据通过直观的图形展示出来,从而有助于对海洋数据进行更准确的分析。根据先验规律可以发现海洋数据按层分布,垂直方向的流速很小,近似为0,因此可以采用分层可视化将每一层按照二维数据进行处理。同样,对向量场可视化的研究一直也是科学可视化领域的研究热点。基于点图标的可视化方法[1]通过给每个采样点指定大小和方向的图标来描述向量场,这种方法实现简单但不能很好地表述向量场的特征,且当采样数据密集时会出现杂乱无章的现象;基于纹理的可视化方法[2]使用纹理的形式表现向量场,主要包括:线卷积积分(Line Integral Convolution,LIC)[3]、点噪声(SpotNoise)[4]和基于纹理图像的向量场可视化(Image Based Flow Visualization,IBFV)[5]。此类方法能完整地表述向量场的信息,但计算量巨大,且存在覆盖和遮挡的问题。基于流线的可视化方法[6]需要首先确定种子点的分布,然后通过流线积分算法生成流线。这种方法易于观察向量场的特征和规律,但可视化效果取决于种子点的分布设置和生成流线的取舍。

基金项目:自然资源部海洋信息技术创新中心开放基金(201906);国家自然科学基金(62172294)
This work was supported by the Open Fund of the Marine Information Technology Innovation Center of the Ministry of Natural Resources (201906) and National Natural Science Foundation of China(62172294).
通信作者:王朕(wangzhen@tjufe.edu.cn)

本文采用基于流线的可视化方法。首先，为了避免种子点的稀疏布局造成流场特征丢失的问题，基于向量场的规模均匀布置种子点，保证种子点的完整覆盖。其次，针对生成大量流线导致的显示杂乱以及遮挡等诸多问题，对流线数据进行聚类分析，将具有相似特征的流线划分为一类。流线聚类问题的关键是对相似性度量值的确定以及聚类方法的选取。文献[7]认为，任意两条流线的起始点间距和结束点间距都在某一阈值范围内时，即判断两条流线相似。这种基于流线端点的聚类，虽然端点间距在一定范围内可代表流线有相似的流动轨迹，但是没有考虑流线生成过程中周围速度变化过大导致流动轨迹发生突变的情形，这导致了某些聚类结果不准确。本文在端点聚类的基础上，提出以点密度模型 Point Density Model(PDM)距离作为相似性度量值再次进行精细聚类。由于 PDM 距离对位置和形状敏感，现已在脑纤维图的绘制方面有所应用[8-9]，它易于捕获缺失的纤维段，并且对纤维束的遮挡和变形有改善作用。除此之外，考虑 PDM 距离计算的时间复杂度是流线点数的平方，计算全距离矩阵时计算量大、耗时长，本文提出使用多维尺度变换(Multidimensional Scaling，MDS)[8]进行算法加速，对距离矩阵进行降维，即只需要计算部分距离矩阵。最后，对于聚类后的代表流线选取，本文采用选取类内中线和边界线的流线简化模式，在保留流场主要特征的同时有效地减少流场中流线的个数。通过与同类算法进行实验比较，结果表明本文提出的流线可视化方法效果良好。

本文的主要贡献如下：

(1)提出将 PDM 距离作为流线聚类的相似性度量值，在流线端点聚类的基础上再进行流线精细聚类，有效解决端点聚类结果不准确的问题，提升了流线聚类的准确性。

(2)提出使用 MDS 算法对流线间距离矩阵进行降维，提升计算速度。

(3)提出使用临界点检测算法，减少流线生成过程中耗时的漩涡，进一步提升了计算速度。

2 相关工作

近年来国内外学者对流线可视化进行了大量的研究，其中种子点的分布算法是研究的重点之一。Jobard 等[10]提出了一种产生均匀分布流线的方法，保证了局部流线的间距，但无法满足全体流线的均匀分布。Liu 等[11]提出了一种高级的流线均匀分布的算法，该算法采用双端队列进行距离测量并不断对流线进行检测，并依据拓扑信息对种子点放置进行优化，改善流线的连续性，但应用在三维流场中会产生遮挡问题。Merbaki 等[12]提出最远距离的种子放置方法，即选择现有流线最远的位置放置种子点生成流线，有利于长流线的生成，但会出现流线不连续的情况。Verma 等[13]提出了基于临界点的种子放置方法，可以较好地表达流场中的特征结构，但难以控制流场的密度，容易产生短流线。

流场简化方面的研究主要包括聚类算法和流线的相似性度量。Corouge 等[14]使用一条流线上的各点匹配另一条流线上距离最近的点，使用匹配的点对来定义相似性。Ding 等[15]将两流线间对应片段匹配的比值定义为流线的相似性。Brun 等[16]指出，若两条流线的起点和终点都在同一距离邻域内，则两流线相似。聚类方面的工作也有很多，Shene 等[17]提出了一种基于层次结构的聚类方法，并用于三维向量场可视化。Mcloughlin 等[18]提出一种基于流向特征量的聚类方法，该方法保证了用户的交互率，同时减少了耗时。Lu 等[19]基于迭代最临近点与 K 均值聚类提取流线的方法，该方法能有效地反映三维流场关键特性。Wang 等[7]基于原始流线端点位置进行聚类，各类再选取特定条数的流线来替代该类流线，使得在不丢失流场重要特征的前提下，避免了流线的杂乱，减少了相互遮蔽等现象。

3 基于流线距离聚类的向量场可视化

3.1 实现流程

本文基于流线的可视化方法主要包括 3 个部分，实现步骤如图 1 所示。

(1)生成流线，通过设置参数等信息完成对基本数据的可视化，同时可以根据需求设置临界点的位置来生成消除漩涡的流线。

(2)流线聚类，主要包括基于端点初步聚类和基于 PDM 距离的再次聚类，聚类过程中采用 MDS 算法对距离矩阵降维，提升计算速度。

(3)提取代表流线，通过排序流线对的 PDM 距离值提取中线和边界线。

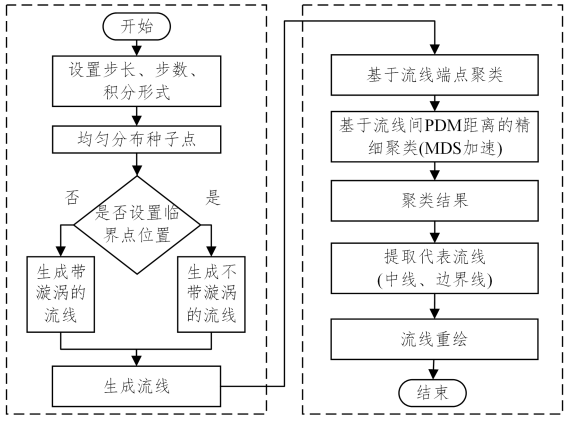


图 1 本文方法的实现流程

Fig. 1 Implementation process of the proposed method

3.2 流线生成

聚类结果的优劣与流线的生成精度有很大关系，为了保证计算流线的距离的准确性，使用四阶龙格库塔(Runge-Kutta)的方式来生成流线。对完整流线定义如下：任意向量场中的一点，若该点的向量不是 0 向量，则具有一条经过该点的流线，如果这条流线满足下列情形之一，则称该流线为经过该点的完整流线。

(1)起点和终点相连，即流线为闭合曲线。

(2)流线的端点到达向量场边界或流线端点处的向量值为 0。

首先均匀布置种子点，覆盖整个向量场，利用这些种子点生成一个完整的流线集合。将二维向量场按照规模分成 $m\times n$ 个正方形区域，找到区域的中心位置作为种子点生成一条完整流线。同时记录下每条流线的起始点和终止点坐标，以便对端点进行后续的操作与计算。

3.3 基于流线距离的聚类

本文的可视化方法是基于流线距离聚类实现的，主要包括

基于流线端点的欧氏距离聚类以及基于流线间 PDM 距离的聚类。最终目标是在损失较少流线特征的前提下提取少数代表流线的方式,实现海洋数据的向量场可视化。

3.3.1　基于流线端点的欧氏距离的聚类

设生成完整流线集共计 N 条,起点用 s 表示,终点用 e 表示,则起点集和终点集为:

$$P_N=\{s_1,s_2,\cdots,s_N\},Q_N=\{e_1,e_2,\cdots,e_N\}$$

首先对起点集合使用基于密度的 DBSCAN[20] 聚类算法进行聚类,通过将距离值小于设定阈值的流线聚为一类,可以将流线预先聚为若干类。在此基础上,对每一类的终点集合再次使用 DBSCAN 聚类,则每一类又可分为若干类,得到最终聚类结果,具体表示为:

$$R=f(F_{PQ}(f(P))) \tag{1}$$

其中,f 为聚类函数,F_{PQ} 为同一流线上起点集 P 到终点集 Q 的映射函数。

聚类结果中类内流线的起点间距和终点间距都在设定的阈值范围内,但是这个距离阈值通常难以确定,尤其是对全新的流场数据。阈值过小,会出现一条流线就聚为一类的情况,每类间的区分度不高,无法删除冗余流线;阈值过大,最终聚类的种类数会变少,很多无关的流线都会被强制聚为一类,导致后续代表流线的选取困难,聚类的效果不明显。此外,聚类结果存在端点相似而流线形状和距离相异而导致的错误聚类,此问题需要进一步解决。

3.3.2　基于流线间 PDM 距离的聚类

针对端点聚类存在的问题,本文将采取以下解决方案:仍以端点聚类进行初步分类,由于 DBSCAN 聚类算法对距离阈值敏感,根据其数据范围,距离阈值设为相邻种子点间距的 2 倍,避免出现一条流线聚为一类的情况。在此基础上,对每一类结果再采用流线间的 PDM 距离作为流线相似性度量值进行精细聚类。

首先,假设存在 3.3.1 节聚类结果的某流线集 $L=\{l_1,l_2,\cdots,l_n\}$,任取两条流线 X 和 Y,由 m 和 n 个点组成,记 $X=\{x_1,x_2,\cdots,x_m\}$,$Y=\{y_1,y_2,\cdots,y_n\}$,$1\leqslant i\leqslant m$,$1\leqslant j\leqslant n$,则 X,Y 的卷积值定义为:

$$\langle X,Y\rangle=\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}[K_\sigma(c_{x,i},c_{y,j})\tau_{x,i}]\cdot\tau_{y,j} \tag{2}$$

$$\|X\|^2=\sum_{i=1}^{m-1}\sum_{j=1}^{m-1}[K_\sigma(c_{x,i},c_{x,j})\tau_{x,i}]\cdot\tau_{x,j} \tag{3}$$

$$\|Y\|^2=\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}[K_\sigma(c_{y,i},c_{y,j})\tau_{y,i}]\cdot\tau_{y,j} \tag{4}$$

其中,$c_{x,i}=\frac{1}{2}(x_i+x_{i+1})$ 为 x_i 和 x_{i+1} 的中心位置,$\boldsymbol{\tau}_{x,i}=(x_{i+1}-x_i)$ 为 x_{i+1} 和 x_i 一个近似的切向量,K 为高斯核函数,σ 为 K 的可调参数。由此 PDM 距离定义为:

$$PDM(X,Y)=\sqrt{\|X\|^2+\|Y\|^2-2\langle X,Y\rangle} \tag{5}$$

依据式(5)可计算每一类内的任意一对流线间的 PDM 距离,其时间复杂度为 O(mn)。假设某一类中的流线条数为 k 条,则计算该类的 PDM 距离矩阵的时间复杂度为 O(k^2mn),空间复杂度为 O(k^2),对该距离矩阵进行聚类即得到聚类结果。

3.4　选取代表流线

聚类后,为消除冗余流线造成的视觉混乱和遮挡而选择代表性流线是流场可视化的重要的环节。一种简单的方法是选择质心(或平均)流线[21]作为每个聚类的代表。由于质心流线是通过人为地对同一聚类内的流线进行平均产生的,不能揭示真实的流动特征,因而较少被采用。在实际实验中更多地会选择离质心最近[22]的流线或者离质心最远[23]的流线作为类内的代表流线,因为它们通常可以描绘特定的流模式。本文以 PDM 距离为依据,选取离质心最近的流线(中线)和最远的流线(边界线)作为代表性流线。

由 3.3.2 节得到的 PDM 距离矩阵是一个对称矩阵,设距离矩阵为 $\boldsymbol{\Delta}$,则 $\Delta(i,j)$ 代表流线 i 到流线 j 的 PDM 距离,且有 $\Delta(i,i)=0$。注意到矩阵每一行的元素之和代表对应流线 i 到其他流线的总距离。由于中线到其他线之间的距离最短,因此可通过计算距离矩阵每行的元素之和后进行排序,数值最小的就是中线,而 $\Delta(i,j)$ 中距离最大的为一对边界线。本文选择中线以及边界线进行显示,可清晰地展示流场结构和特征,有效地减少流线遮挡和杂乱现象。

算法 1　流线选取算法

输入:PDM 距离矩阵 $\boldsymbol{\Delta}$

输出:代表流线集合 $r=\{l_{k_1},l_{k_1},l_{k_3}\}$

1. 根据距离矩阵的定义求其行和,第 i 行的和定义为 $S_i=\sum_{j=1}^{n}\Delta(i,j)$,其中 $\Delta(i,i)=0$。
2. 定义集合 $T=(S_1,S_2,\cdots,S_n)$,T_i 代表第 i 条流线到其他 n−1 条流线的 PDM 距离之和,对集合 T 中元素值进行排序,找到下标 $k_1=$ index(min(T)),$1\leqslant k_1\leqslant n$,得到中线流线 l_{k_1}。
3. 对 PDM 距离矩阵元素 Δ(i,j) 进行排序,找到下标 $(k_2,k_3)=$ index(max(Δ(i,j))),$1\leqslant k_2,k_3\leqslant n$,得到边界流线对 $\{l_{k_2},l_{k_3}\}$。

其中存储集合 T 的空间复杂度为 O(n),排序算法时间复杂度为 O($n\log n$)。因为类内所有流线中,中线到其他流线的距离最短,而两条边界线之间的距离最长,故选取的代表流线 $r=\{l_{k_1},l_{k_2},l_{k_3}\}$。

3.5　聚类评估

因为可视化领域中聚类效果的优劣可以通过主观的目视检查和比较获得,为了更加准确地评估聚类质量以及更加方便地进行聚类方法的横向比较,还可以使用定量的客观指标。本文部分借鉴了血流可视化的评价方法[24-25],提出以下两种聚类评估指标以度量流线可视化的效果。

(1)轮廓系数(Silhouette Coefficient,SC)[24-25]:

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{6}$$

其中,$a(i)$ 代表第 i 条流线到同类之间其他流线的平均距离,$b(i)$ 代表流线 i 到最近的其他类的所有流线的平均距离。所有样本轮廓系数的平均值为聚类结果的轮廓系数,它是类内聚力和分离的非线性组合测度,轮廓系数越大,说明类内聚力越高,不同类之间的分离越好。

(2)戴维森堡丁指数(Davies-Bouldin index,DBI)[26]:

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\overline{C_i}+\overline{C_j}}{\|w_i-w_j\|_2}\right) \tag{7}$$

其中,$\overline{C_i}+\overline{C_j}$ 代表任意两类的类内平均距离之和,$\|w_i-w_j\|_2$ 为两聚类中心的欧氏距离。DBI 越小表示聚类效果越好,因为生成类内距离短和类间距离长的算法将具有较低的 DBI。

4　算法加速

本文按照向量场全覆盖种子点进行数值积分计算生成流线,遇到漩涡时会产生大量不必要的计算。此外,计算 PDM

距离矩阵的时间复杂度过高。针对这两个问题，本文通过以下两种途径提出算法加速方案。

4.1 MDS 加速

3.3.2 节计算两条流线间 PDM 距离时，需要流线上所有的点都参与计算，运算时间较长。多维尺度变换(Multidimensional Scaling，MDS)[8]属于非线性降维[27]的一种形式，是一种传统的寻求保持数据点之间差异性的降维方法，它可以使得在原数据集合中相近的点仍然靠在一起，远离的点仍然远离。选择 MDS 算法对 PDM 进行降维能够在保持数据接近理论值的同时降低计算量。

算法 2 多维尺度变换算法

输入：原始流线全集 F

输出：流线映射二维点集 F′

1. 假设 F 中共有 n 条流线，从 F 中均匀抽取部分流线构成样本子集，假设抽取流线数为 a 条。为了保证样本子集能够体现流场特征，子集选取操作放在完整流线经过端点聚类之后，计算样本子集的距离矩阵 **A**，则 **A** 的大小为 a×a。
2. 计算子集和剩余流线集之间的 PDM 距离，得到距离矩阵 **B**，**B** 的大小为 a×(n−a)。
3. 计算近似距离矩阵 $\mathbf{D}=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^{\mathrm{T}} & \mathbf{C}\end{bmatrix}$，其中使用 Nystrom Extension[8] 的矩阵形式为 $\mathbf{B}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}^{-1}$，其中 **D** 是样本集 **A** 的特征向量构成的矩阵，可以求得 **D** 的特征向量的估计 $\overline{\mathbf{U}}$ 和 **D** 的估计 $\hat{\mathbf{D}}$，且 **C** 可由 $\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B}$ 近似。

$$\overline{\mathbf{U}}=\begin{bmatrix}\mathbf{U}\\ \mathbf{B}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}^{-1}\end{bmatrix} \tag{8}$$

$$\begin{aligned}\hat{\mathbf{D}}&=\overline{\mathbf{U}}\boldsymbol{\Lambda}\overline{\mathbf{U}}^{\mathrm{T}}=\begin{bmatrix}\mathbf{U}\\ \mathbf{B}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}^{-1}\end{bmatrix}\boldsymbol{\Lambda}\left[\mathbf{U}^{\mathrm{T}}\quad \boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{B}\right]\\ &=\begin{bmatrix}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} & \mathbf{B}\\ \mathbf{B}^{\mathrm{T}} & \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B}\end{bmatrix}\\ &=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^{\mathrm{T}} & \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B}\end{bmatrix}\end{aligned} \tag{9}$$

4. 取 **D** 中的最小的前 k 个特征向量并标准化即可得到最后的映射点集 F′，该点坐标集可以唯一对应到原始集合 F，并近似保留原流线间的距离值。

4.2 漩涡减少

在流线生成的过程中会出现大量漩涡，通过某个临界点之后流线会不断螺旋式地靠近漩涡中心。在不需要显示漩涡特性的可视化应用中会导致大量计算能力的不必要消耗和显示时间的延长。本文提出的解决方案是首先检测临界点，然后通过判断临界点的类型来确定是否存在漩涡，如果存在就尽早结束此类流线的生成。

基于二维向量场临界点理论[28]，根据 4 个顶角点的速度值，判断围成区域内有没有速度为零的临界点。如图 2 所示，记临界点坐标为 (x,y)，速度为 (u,v)，计算时可求 u 方向的零等值线，再求 v 方向的零等值线，如果两条等值线相较于该区域内部则认为包含临界点。四顶点组成的网格内临界点计算采用双线性插值算法，若存在临界点，具体位置通过求解下列方程组求得：

$$\begin{cases}u=u_0(x_2-x)(y_2-y)+u_1(x_2-x)(y-y_1)+\\ \quad u_2(x-x_1)(y_2-y)+u_3(x-x_1)(y-y_1)=0\\ v=v_0(x_2-x)(y_2-y)+v_1(x_2-x)(y-y_1)+\\ \quad v_2(x-x_1)(y_2-y)+v_3(x-x_1)(y-y_1)=0\end{cases} \tag{10}$$

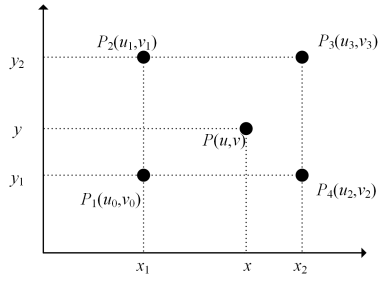


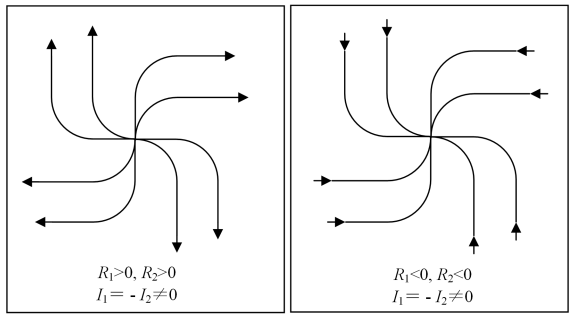
图 2 双线性插值判断临界点

Fig. 2 Bilinear interpolation judgment critical point

实际计算过程中，如果一个网格中的四个顶点向量分量有相同的符号(都大于 0 或都小于 0)，可以判断这个网格一定没有临界点。得到临界点位置之后，通过求其在 x 和 y 方向的偏导数得到雅可比矩阵 $\boldsymbol{J}$：

$$\boldsymbol{J}=\frac{\partial(u,v)}{\partial(x,y)}=\begin{bmatrix}\dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y}\\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y}\end{bmatrix} \tag{11}$$

进一步求得雅可比矩阵的特征值 $\lambda_1=R_1+I_1 i$ 和 $\lambda_2=R_2+I_2 i$(R 代表实部，I 代表虚部)。图 3 是排斥焦点与吸引焦点两种漩涡临界点的判别方法。



(a)排斥焦点 (b)吸引焦点

图 3 临界点分类与判定方法

Fig. 3 Classification and determination method of critical point

5 实验与讨论

实验环境为 Intel Core i7-10700 CPU @2.90 GHz，64 GB 内存，Nvidia GeForce RTX 3070 显卡，Windows10 64 位操作系统。实验数据选取香港科技大学海洋科学系取得的中国南海、东海、渤黄海近年来的二维流场数据[29]，数据格式为网络通用格式(Network Common Data Form，NetCDF)。由于海洋向量数据与地理位置有一定的关系，选用带有地形和纹理的环境进行向量场可视化的效果会更好。本文使用可视化工具(Visualization Toolkit，VTK)进行纹理贴图和可视化。

5.1 流线聚类实验

首先均匀布置种子点，采样间隔为 1 000，前后双向采用四阶龙格库塔积分在中国东海、南海、渤黄海流域所生成的完整流线，如图 4(a)－图 4(c)所示。参数设置步长为 2 000，步数为所能达到的最大步数，前后双向产生流线，并按照前向流线的末端点与后向流线的始端点进行合并。完整流线集并不会缺失关键特征，但流线显示杂乱，视觉效果差，无法提取有效信息帮助用户清晰直观地理解数据背后所隐藏的规律。

图 4(d)－图 4(f)给出了各海洋流线集通过端点聚类的结果，距离阈值为 4 500。图中将同一类型的流线标记为一种

颜色，可以发现大部分的流线有相似的流动趋势，但仍然存在少量聚类异常的情形。故进一步提取了南海中间偏西南部和西部两处异常区域，使用黑色椭圆区域对其进行标注，放大后如图 4(g)、图 4(h)所示。在图 4(h)中可以发现该类流线上有一个凸起的部分，流动趋势发生了偏离。因此仅使用端点聚类存在着异常或不准确的情形，但是由于其端点间距离小于给定阈值而被归为一类。图 4(g)中间偏右边凸起的部分同理。针对上述问题，选择使用 PDM 距离作为流线相似性准则对其进行精细聚类。聚类效果如图 4(i)、图 4(j)所示，流线将凸出的部分聚为一类，不在凸出部分的聚为另一类，聚类结果准确有效。

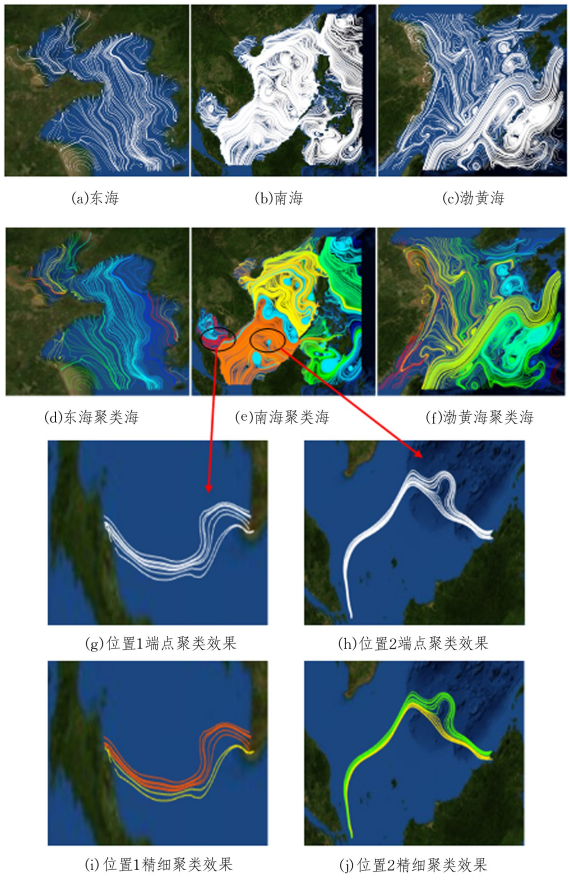


图 4　基于流线聚类的可视化

Fig. 4　Visualization based on streamline clustering

为了验证 PDM 中参数的取值是否达到最优效果，选取了南海数据集上流线条数从 20～260 的类进行了比较，σ 的取值范围为 10^{-9}～10^{-5}，当 σ 取值为 10^{-8} 时轮廓系数停止显著变化，如图 5 中红线所示。因此，选取 10^{-8} 为 PDM 的 σ 参数。

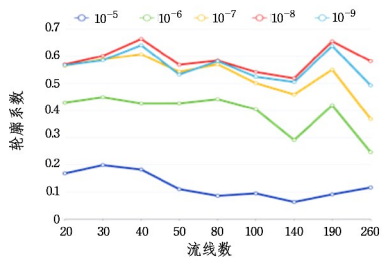


图 5　参数测定(电子版为彩图)

Fig. 5　Parameter determination

为了验证本文提出的 PDM 距离是否在流线聚类中发挥良好的效果，选取了文献[30]中常用于流线聚类的相似性度量值和聚类算法进行比较实验。选择其中推荐的 3 种聚类算法：DBSCAN 聚类、K 均值(K-means)聚类[30]和凝聚层次聚类(Agglomerative Hierarchical Clustering，AHC)[30]，距离准则为 single-linkage。相似性度量值选取 MCP(Mean-of-closest-point)距离[24-25]和 Hausdorff 距离[30]。在中国南海、东海、渤黄海数据集上计算轮廓系数和戴维森堡丁指数并进行比较。其中，K-means 聚类算法不可直接对距离矩阵聚类，因此将距离矩阵使用主成分分析(Principal Components Analysis，PCA)[30]的方法降维后进行聚类。如图 6 和图 7 所示，AHC 聚类算法＋PDM 距离相似性度量值的组合在两个指标上有着最好的评价效果。同时观察到聚类算法 Kmeans＋PCA 和 AHC 相比 DBSCAN 表现更好。因为 DBSCAN 是基于密度聚类的算法，它会不恰当地将一些重要的有几何特征的积分曲线视为异常值，且由于这些积分曲线形状不同，与其他流线的距离非常大，故会导致较差的指标数值。Kmeans＋PCA 和 AHC 获得较高的数值是由于数据集产生的簇形状为球、凸形状。

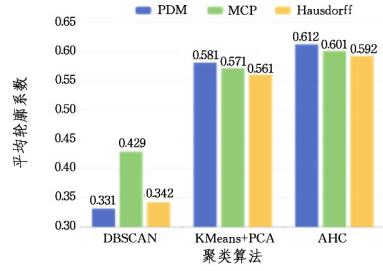


图 6　轮廓系数

Fig. 6　Silhouette coefficient

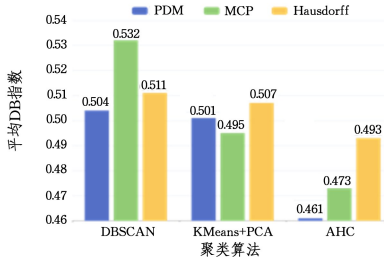


图 7　戴维森堡丁指数

Fig. 7　Davies-Bouldin index

将完整流线集精细聚类之后选取代表流线进行流场简化，本文选取流线的中线和边界线作为代表流线。如图 8 所示，将流线聚类结果中大于等于 3 条的类进行中线和两条边界线的提取。其中紫色流线代表中线，绿色流线为边界线，红色流线是产生漩涡的流线，而针对类中流线少于 3 条的情况，将其颜色映射为蓝色。

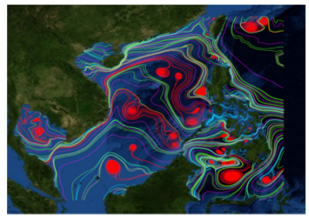


图 8　流线简化(电子版为彩图)

Fig. 8　Streamline simplification

5.2 算法加速实验

加速实验中的流线集所使用的是南海数据集，总流线为1730条。表1中第一列为按照端点聚类后提取10%，20%，30%的流线条数，分别为174，346，520条。随着流线数量的翻倍增长会导致运算时间的数倍增加。因此，不直接对流线进行PDM距离计算，而是先基于端点聚类之后再进行计算。另外PDM距离矩阵经MDS降维之后，进行流线精细聚类的运算时间只需原来的11%～14%，且与使用PDM聚类结果的差异在1%～2%。

表1 PDM全距离矩阵与MDS降维矩阵的比较

Table 1 Comparison between PDM full distance matrix and MDS dimension reduction matrix

流线数	PDM运行时间/s	MDS运行时间/s	聚类差异/%
174	399.2	45.6	1.724
346	1753.9	199.3	1.445
520	3569.7	497.4	1.222

漩涡减少的实验效果如图9所示，左边为在南海数据集上生成的漩涡示意图，右边是去除漩涡后的示意图。在生成流线时，需要设定漩涡半径 r，判断当生成流线进入以临界点坐标 P 为圆心、r 为半径的圆形区域内即停止积分，达到消除漩涡、减少流线生成时间的目的。当然，有时在观测海洋数据时也需要了解漩涡情况，因此可以根据实际需求进行显示。为了更好地表达流场特征，可视化结果中选择牺牲部分计算时间来对漩涡进行表达。

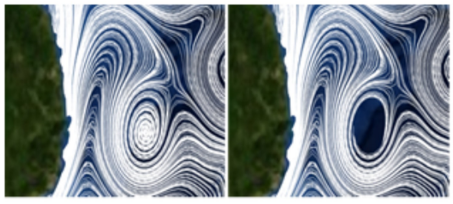


图9 漩涡减少对比图

Fig. 9 Vortex reduction comparison

结束语 高质量的向量场可视化技术有助于用户理解大规模海洋数据背后所隐藏的规律，提高研究工作效率。本文提出PDM距离作为聚类相似性度量值，对向量场的流线进行精细聚类，可准确高效地将向量场分为若干类。在分类基础之上，通过将PDM距离值排序，选取每一类的中线和两条边界线作为代表流线进行显示，关键信息保留完整且可视化效果良好。同样也可以按照文献[31]中的方法，选取聚类结果中感兴趣的一条流线作为目标流线，在流线中查找与其最相似(与目标流线的PDM距离最短)的若干条流线进行可视化展示。此外，本文通过MDS算法对数据进行降维处理，加快计算过程，以及对临界点进行检测进而减少流线生成过程中的漩涡产生，使得可视化计算时间大幅缩短。通过在中国沿海的海洋数据集上的测试，应用本文方法可以形成高质量的海洋向量场可视化产品。尽管使用降维算法加快了计算速度，但是遇到更大数量规模时，运算速度仍然较慢。下一步的工作重点是如何更好地减少运算时间以及与其他的可视化方法融合以取得更好的效果。

参 考 文 献

[1] FAN Y, WU X Q, ZHANG J K, et al. Research and realization of flow field dynamic visualization based on geometric shader [J]. Computer Engineering and Applications, 2019, 55(9): 157-161.

[2] DU X, LIU H, TSENG H W, et al. A vector field texture generation method without convolution calculation [J]. Symmetry, 2020, 12(5): 724-746.

[3] CABRAL B, LEEDOM L C. Imaging vector fields using line integral convolution [C]// Proceedings of the 20st Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH, 1993: 263-270.

[4] VAN WIJK J J. Spot noise texture synthesis for data visualization [C]// Proceedings of Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH. New York: ACM Press, 1991: 309-318.

[5] VAN WIJK J J. Image based flow visualization [J]. ACM Transactions on Graphics, 2002, 21(3): 745-754.

[6] JI X F, ZHANG F, WANG Z Y, et al. Research of dynamic streamline visualization of ocean flow field with real-time construction of coloring model [J]. Journal of Zhejiang University(Science Edition), 2020, 47(1): 45-51.

[7] WANG S R, WANG G P. Clustering based 2d vector field Visualization [J]. Journal of Computer-Aided Design and Computer Graphics, 2014, 26(10): 1593-1602.

[8] SILESS V, MEDINA S, VAROQUAX G, et al. A comparison of metrics and algorithms for fiber clustering [C]// Proceedings of the 2013 International Workshop on Pattern Recognition in Neuroimaging. 2013: 190-193.

[9] AUZIAS G, J GLAUNÈS, COLLIOT O, et al. DISCO: a coherent diffeomorphic framework for brain registration under exhaustive sulcal constraints [C]// Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention. Springer-Verlag, 2009: 730-738.

[10] JOBARD B, LEFER W. Creating evenly-spaced streamlines of arbitrary density [C]// Proceedings of Visualization in Scientific Computing. Heidelberg: Springer, 1997: 43-55.

[11] LIU Z, MOORHEAD R, GRONER J. An Advanced Evenly-Spaced Streamline Placement Algorithm [J]. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5): 965-972.

[12] MEBARKI A, ALLIEZ P, DEVILLERS O. Farthest point seeding for efficient placement of streamlines [C]// Proceedings of IEEE Visualization. Washington D C: IEEE Computer Society Press, 2005: 479-486.

[13] VERMA V, KAO D, PANG A. A flow-guided streamline seeding strategy [C]// Proceedings of the Conference on Visualization. Los Alamitos: IEEE Computer Society Press, 2000: 163-170.

[14] COROUGE I, GOUTTARD S, GERIG G. Towards a shape model of white matter fiber bundles using diffusion tensor MRI [C]// Proceedings of 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro. 2004: 344-347.

[15] DING Z H, GORE J C, ANDERSON A W. Case study: recon-

struction,visualization and quantification of neuronal fiber pathways[C]// Proceedings of IEEE Visualization. 2001:453-456.

[16] BRUN A,PARK H,KNUTSSON H,et al. Coloring of DT-MRI Fiber Traces Using Laplacian Eigenmaps[C]// Proceedings of Lecture Notes in Computer Science. Springer-Verlag,2003:518-529.

[17] SHENE C K,WANG C L,YU H F,et al. Hierarchical streamline bundles[J]. IEEE Transaction on Visualization and Computer Graphics,2012,18(8):1353-1367.

[18] MCLOUGHLIN T,JONES M W,LARAMEE R S,et al. Similarity Measures for Enhancing Interactive Streamline Seeding [J]. IEEE Transactions on Visualization and Computer Graphics,2013,19(8):1342-1353.

[19] LU D Y,ZHU D M,WANG Z Q. Streamline Selection Algorithm for Three-Dimensional Flow Fields[J]. Journal of Computer-Aided Design and Computer Graphics,2013,25(5):666-673.

[20] ESTER M,KRIEGEL H P,SANDER J,et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI,1996:226-231.

[21] MCLOUGHLIN T,JONES M W,LARAMEE R S,et al. Similarity measures for enhancing interactive streamline seeding[J]. IEEE Transactions on Visualization and Computer Graphics,2013,19(8):1342-1353.

[22] CHEN C K,YAN S,YU H F,et al. An illustrative visualization framework for 3d vector fields[C]// Proceedings of Computer Graph. Forum,2011:1941-1951.

[23] DONNELL L J O,WESTIN C F,GOLBY A J. Tract-based morphometry for white matter group analysis[J]. NeuroImage,2009,45(3):832-844.

[24] OELTZE S,LEHMANN D J,THEISEL H,et al. Evaluation of streamline clustering techniques for blood flow data[R]. Otto-von-Guericke-Universität Magdeburg,Technical Report,2012.

[25] OELTZE S,LEHMANN D. J,KUHN A,et al. Blood flow clustering and applications in virtual stenting of intracranial aneurysms[J]. IEEE Transactions on Visualization and Computer Graphics,2014,20(5):686-701.

[26] DAVIES D L,BOULDIN D W. A cluster separation measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1979,1(2):224-227.

[27] DU J,WANG X,HU L J. Technology of nonlinear dimension reduction and its application in visualization[J]. Journal of Donghua University(Natural Science),2020,46(4):675-680.

[28] HELMAN J L,HESSELINK L. Visualizing vector field topology in fluid flows[J]. IEEE Computer Graphics and Applications,1991,11(3):36-46.

[29] GAN J,LIU Z,LIANG L. Numerical modeling of intrinsically and extrinsically forced seasonal circulation in the China Seas:A kinematic study[J]. Journal of Geophysical Research:Oceans,2016,121(7):4697-4715.

[30] SHI L Y,LARAMEE R,CHEN G N. Integral curve clustering and simplification for flow visualization:a comparative evaluation[J]. IEEE Transactions on Visualization and Computer Graphics,2021,27(3):1967-1985.

[31] XIONG G Z,HUANG Z B,DAI Z T,et al. A data driven characteristically filtering method for 3d flow field[J]. Journal of Beijing University of Posts and Telecommunications, 2019,42(6):91-97.

WANG Zhen, born in 1981, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include multi-dimensional reconstruction and visualization for application in computed tomography and computer vision.