引文格式：车一鸣 ,邓维熙 ,冯飞军 ,等. 耦合机器学习与连通性修正的河流选取模型[J]. 地理空间信息, 2025, 23(9): 22-26. doi:10. 3969/j.issn. 1672-4623.2025.09.006

耦合机器学习与连通性修正的河流选取模型

车一鸣[1]，邓维熙[2*]，冯飞军[1]，张静静[1]，卢文渊[2]，任东宇[2]
（1. 宁夏回族自治区测绘地理信息院，宁夏 银川 750002；2. 自然资源部第三地理信息制图院，四川 成都 610100）

摘　要：传统河流选取方法在数据量较大时易出现效率较低等问题，因此以河南、宁夏河流数据为例，利用支持向量机（SVM）模型对河流要素进行初选取，并引入改进的Dijkstra算法构建水系路径；再对初选取结果进行连通性校正，实现河流要素的自动选取，保证河流选取的准确性与高效性。结果表明，SVM模型能有效抽稀河流要素，效率高于随机森林算法且对高保留等级的河流判断准确；改进的Dijkstra算法能解决初选取结果中河流断流等问题，且在数据量较大时能显著减少水系路径构建时间。耦合机器学习与连通性修正的方法在一定程度上能保持河流的流通性与完整性，从而满足不同尺度下制图需求。

关键词：SVM；Dijkstra算法；河流选取；线要素
中图分类号：P283　　**文献标志码：**B　　**文章编号：**1672-4623（2025）09-0022-05

River Selection Model Coupled with Machine Learning and Connectivity Correction

CHE Yiming[1], DENG Weixi[2], FENG Feijun[1], ZHANG Jingjing[1], LU Wenyuan[2], REN Dongyu[2]
(1. The Surveying, Mapping and Geoinformatics Institute of Ningxia Hui Autonomous Region, Yinchuan 750002, China;
2. The Third Geoinformation Mapping Institute of Ministry of Natural Resources, Chengdu 610100, China)

Abstract: The traditional river selection methods are prone to problems such as low efficiency when handling large datasets. Therefore, taking river data from Henan Province and Ningxia Hui Autonomous Region for example, we used support vector machine (SVM) model to initially select river elements, and introduced improved Dijkstra algorithm to construct the river network paths. Then, we proceeded connectivity correction on the initially selected results to achieve the automatic selection of river elements. The results show that the SVM model effectively simplifies river elements, outperforming the random forest algorithm in terms of efficiency and accuracy in identifying high-priority rivers. The improved Dijkstra algorithm addresses issues such as river discontinuities in the initial selection and significantly reduces the time required for constructing river network paths in large datasets. The integration of machine learning and connectivity correction maintains the continuity and integrity of rivers to a certain extent, meeting the cartographic needs at different scales.

Key words: SVM, Dijkstra algorithm, river selection, line element

随着科技的不断发展，地理信息数据的应用面不断扩大，传统基础比例尺地图数据库因更新周期长已难以满足用户需求，因此需要在其基础上探索快速、准确的多比例尺数据库的建设方法[1-4]。水系数据作为一种地理信息数据，既能展示当前地理空间格局的分布，又能为规划布局提供参考[5]；同时由于我国水资源分布不均，掌握不同比例尺下的水系分布，可为兴建水库、引渠灌溉等提供更科学的指导[6]。河流选取作为制图综合中的重要研究方向，其本质是保留选取等级高的河流，并在保证河流空间连续性的基础上，减轻河流的密集程度[7]。基于知识规则[8]、特征指标[9]的传统要素选取方法虽然建立了较完善的专家知识体系，但在实际应用中仍面临两个关键问题：①人工特征的选取过度依赖于操作者的知识经验；②基于知识规则的选取在特征权重分配和空间关联方面存在一定局限性，可能导致结果仅为局部最优。随着大数据技术的不断更新，越来越多的学者[10-12]开始探索基于机器学习的河流选取方法，尝试挖掘矢量数据及其周边的空间特征，以便制作样本进行训练。同时，部分研究表明支持向量机（SVM）算法非常适合小样本训练方式，在保证准确率的同时，能大幅减少工作量[13]。然而基于机器学习的河流选取方法仍存在一定的局限性，如对计算机硬件要求高、选取后存在破碎等[14]，因此需要与其他算法相结合，在确保效率的基础上控制要素的连通性。构建线要素的Stroke是一种能保持连通性和提高效率的办法[15]，但将其应用于自

项目来源：宁夏自然科学基金资助项目（2023AAC03572）；宁夏优秀人才支持计划资助项目。
第一作者简介：车一鸣（1988—），硕士，主要从事基础测绘与地图制图工作，E-mail：446824914@qq.com。
通信作者：邓维熙（1999—），硕士，主要从事地图综合与软件开发工作，E-mail：448235152@qq.com。

然河流中，需要提前对河流进行简化，增加了工作量。Dijkstra算法[16]是一种基于广度优先思想的算法，以图为基础，以求取两个图节点之间最短路径为目的，不仅能缩短运行时间，还能确保河流要素构建出连通的水系路径，具有普遍最优性[17]。在其基础上引入启发函数和评价函数，还能进一步提高效率，且适用于大数据量的背景中。综上所述，本文以SVM模型为核心对河流要素进行初选取，从而确保选取的准确性；再以改进的Dijkstra算法为核心构建河流连通性修正模型，从而确保河流选取的连通性与高效性；最后集成二者构建河流选取模型，并以河南、宁夏河流数据验证方法的可行性。

1　研究方法

本文的技术路线主要包括河流初选取和连通性修正两个模型，河流初选取综合考虑河流要素的几何特征与拓扑特征，采用SVM模型对河流进行训练迭代，从而实现初选取；连通性修正则是为了尽量减少河流中断现象，采用Dijkstra算法构建水系路径，并根据河网具体情况启用A*算法，赋予河流连通属性；在初选取的基础上综合考虑其连通性，从而得到综合后的矢量数据，实现河流自动选取。

1.1　特征库构建

参考河流自动选取的相关研究，本文选择以下特征构建机器学习特征库：①长度，是河流最基础的属性特征；②直曲比，是河段起点与终点之间的直线距离与实际长度的比值，用于衡量河段弯曲程度；③汇水面积，是河流某一出口断面上游流域分水线所围区域的总面积，面积越大表明该河流汇集的地表水资源越多，其在水文网络中的重要性越高，本文通过叠加DEM与河网数据得到河流的汇水面积[18]；④河流等级，本文采用Strahler分级法对河流进行编码分级，直接发源于河源的河流为1级，同级的两条河流交汇形成的河流的等级比原来增加1级，不同等级的两条河流交汇形成的河流等级等于原来河流中的等级较高者（图1）；⑤下一级支流数，是指与该条河流直接相连的支流数量，对于同等级河流，值越大，综合选取时的级别越高。

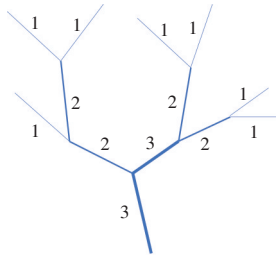


图1　Strahler河流分级示意图

1.2　基于SVM模型的河流初选取

为保证实验结论具有普适性，在河流初选取前，对所有河流进行实交断线，确保实验数据充足，且剔除过长河流对整体选取结果的影响。SVM模型的核心思想是构造一个分类超平面，尽可能分离不同类别的样本，且分类间隔最大。其分类函数定义为：

$$f(x)=\mathrm{sgn}(\boldsymbol{W}\times x+b) \tag{1}$$

$$\mathrm{sgn}(x)=\begin{cases}-1 & x<0\\ 0 & x=0\\ 1 & x>0\end{cases} \tag{2}$$

式中，$\boldsymbol{W}$为超平面法向量；b为偏置；sgn为逻辑函数。

对于每条河段，首先在特征库中找到其对应的特征，并将所有特征抽象为特征向量$\boldsymbol{X}=[x_1, x_2, x_3, \cdots, x_i]$，$x_i$为上一节中选取的特征指标；再通过标准化方法将特征值归一化至同一尺度，以便为SVM模型提供训练样本。惩罚因子C反映SVM模型在构造最优分类平面时对噪声数据的容忍程度。较大的C倾向于减少训练误差，但可能导致过拟合；较小的C可提高泛化能力，但可能增加训练误差。核函数参数γ决定了核函数的径向作用范围，较大的γ更关注局部特征，但可能导致过拟合；较小的γ则更关注全局特征，但可能导致欠拟合。因此，本文采用网格法确定C和γ的最优值，对C和γ分别给定M、N个值，构造得到$M\times N$种(C,γ)参数组合，并分别评价它们的性能，以此找到最佳参数。通过实验，最终确定$C=1$、$\gamma=0.8$。

1.3　基于改进Dijkstra算法的高效水系路径构建

1.3.1　融合启发式搜索的水系路径构建

水系路径构建的核心是Dijkstra算法。它是一种针对有权图的广度优先算法，以遍历所有图节点获取最短路径为目的，应用于河流选取时能从每个河源节点（起点）不断向外遍历延伸直至河口节点（终点），通过权重对比更新当前路径，从而获取单源最优路径[19]。在利用Dijkstra算法构建最优路径时，引入A*算法（启发式搜索）加速搜索，以便进一步提高水系路径构建速度。构建水系路径的具体步骤为：①从河源节点a_0开始向外遍历与其共边的所有节点，并为对应边赋初始权重$w_{ak}=L_{0k}$，其中L_{0k}为a_0与a_k相连边的长度；②获取权重w最小的边所连接的节点a_k，记录其相关信息，更新路径长度，并以节点a_k为新的河源节点向外遍历与其共边的节点，通过不断向外遍历，构建a_0对应的水系路径l_0；③当河网较

大时启用A*算法，引入欧氏距离作为启发式函数 $h(x)$ 计算当前节点 a_k 到下一目标节点的代价以及评估函数 $f(x)$ 计算节点的优先级：

$$\begin{cases} f(x)=h(x)+g(x) \\ h(x)=\sqrt{(x_{k+1}-x_k)^2+(y_{k+1}-y_k)^2} \\ g(x)=g(x_0)+w(0,k) \end{cases} \tag{3}$$

式中，$g(x)$ 为实际代价，当位于河源节点时 $g(x)=0$，其他节点则趋于无穷；$w(0,k)$ 为由 a_0 到 a_k 的权重，若计算得到的 $g(x)$ 小于实际 $g(x)$，则更新 $g(x)$ 为当前计算值，并设置节点 x 为优先。

水系路径构建完毕后，从当前河网中剔除构建水系路径的所有要素，若剔除后的河网小于触发A*算法阈值，则重新回归至Dijstra算法进行水系路径构建，从而确保高效率。

1.3.2 顾及约束条件的水系路径重构

构建水系路径时，通常存在多个河源节点（a_0、a_1、…）共用同一节点 a_k 的情况，因此为确保每条水系路径的连通性以及水系路径之间的独立性，设置3个约束条件：

1）在遍历过程中，若当前节点 a_k 已被其他节点遍历且用于构建水系路径 l_0 的节点，则通过对比当前路径到节点 a_k 的总权重与节点 a_k 所属水系路径中河源节点到节点 a_k 的总权重来确定节点 a_k 的归属。

$$l_0=l_{0k}+l_{kn} \tag{4}$$

$$w_{\text{total}}=\sum_1^k w_k \tag{5}$$

$$\begin{cases} w_{\text{total}-0k}<w_{\text{total}-1k}, & l_0=l_{1k}+l_{kn} \\ w_{\text{total}-0k}\geqslant w_{\text{total}-1k}, & l_0=l_0 \end{cases} \tag{6}$$

式中，l_0 为从 a_0 开始构建的水系路径；l_{0k} 为由 a_0 到 a_k 构建的水系路径；$w_{\text{total}-0k}$ 为 a_0 到 a_k 的总权重。

2）在更新水系路径过程中，若 l_0 中的节点 a_k 被更新至 l_1 中，则将 l_0 的终点更新为节点 a_k 的上一节点 a_{k-1}，若节点 a_{k-1} 为河源节点则删除水系路径 l_0，再将 a_{k+1}、a_{k+2} 等重新加入遍历列表中，等待被连接为新的水系路径。

3）相关节点重新加入遍历列表后，重新判断当前河网大小，若超过阈值则启用A*算法。具体工作流程见图2。

2 实验结果与分析

2.1 实验数据

为更好地展示本文方法在不同地区的适用性，选取宁夏回族自治区固原市和河南省作为研究区，其中

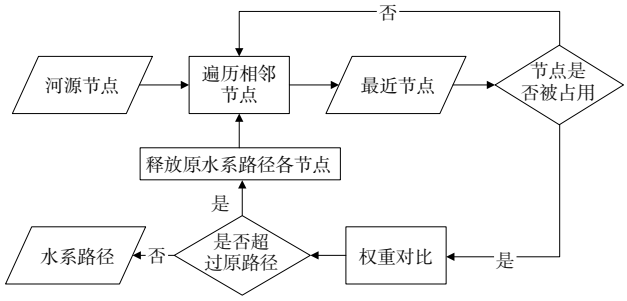
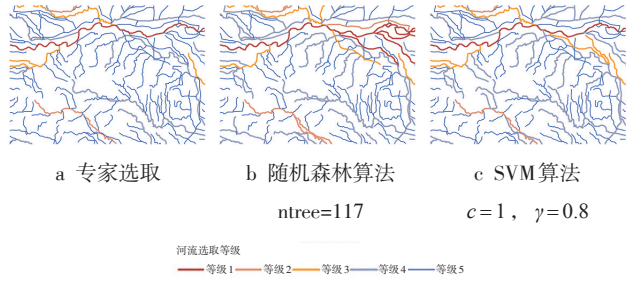


图2 水系路径重构示意图

固原市河网结构呈明显的树状，共有4 314条河流要素，来源于宁夏地图服务保障体系建设项目构建的全区1∶50 000地理底图数据库；河南省河网结构呈网状，共有34 631条河流要素，来源于天地图1∶25 000矢量数据库。训练过程中上述河流要素均按8∶1∶1的比例随机划分训练集、验证集与测试集，同时对所有河流要素进行实交断线，尽可能减少过长河流对实验选取结果的影响，并剔除了双线河、人工沟渠等，确保测试数据的纯净性。

2.2 基于机器学习的河流自动分级对比分析

为展示不同机器学习算法对河流提取的效果，在实验前委托专家和数据生产人员对郑州市河流数据（共计672条）进行手动分级，作为基础依据（图3a），实验选择稳定且经过验证的随机森林与SVM算法进行对比分析，各自最优参数下的结果见图3b、3c，可以看出，两种算法的分级结果与专家选取结果总体差异较小，但仍存在瑕疵，随机森林算法中等级1的河流明显更多，且出现了明显的河流包裹现象，因此在自动选取时必然会出现局部区域密度大的问题；而SVM算法中等级4的河流较多，虽整体连贯性更优，但在自动选取时可能会出现整体密度大的情况。



a 专家选取　b 随机森林算法 ntree=117　c SVM算法 $c=1$，$\gamma=0.8$

图3 河流选取等级分级

为了从定量的角度进行评价，以专家选取结果为基础，本文将等级1~3的河流划分为高保留级河流，等级4的河流划分为中保留级河流，等级5的河流划分为低保留级河流，以一致率、运行时间为评价标准，具体结果见表1，可以看出，两种算法的总体一致率相差无几，但随机森林算法对高保留等级河流的误判严重，将导致后续的河流删除和重构出现严重的

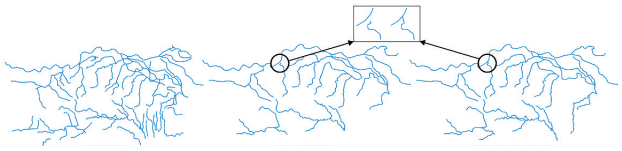
河网结构错误；两种算法的运行时间存在明显差异，随机森林算法的耗时几乎为SVM算法的两倍。河流选取时的河流数据量为当前实验的数十倍甚至上百倍，因此本文选择SVM算法进行后续研究。

表1　河流分级评价表

方法	高保留级河流（一致河流）/条	中保留级河流（一致河流）/条	低保留级河流（一致河流）/条	总体一致率/%	运行时间/min
专家选取	97	146	429	–	60.0
随机森林算法	138（97）	173（140）	361（356）	88.2	1.1
SVM算法	109（97）	197（141）	366（359）	88.8	0.6

2.3　耦合SVM与连通性校正的河流自动选取分析

基于SVM模型对河流要素进行初选取是本文实现河流自动选取的核心之一。综合考虑河流要素的几何特征和上下游关系，对河流要素进行了取舍，以河南省郑州市局部测试数据为例，本文方法的选取结果见图4，可以看出，SVM模型能在抽稀河流要素的基础上维持河网的基本形状，但如图4b圆圈区域所示，选取后的河流出现中断，结合图3c分析可知，尽管该段河流等级较高，但长度过短，导致在选取过程中被赋予较低的评分，从而被删除，这也表明虽然机器学习算法在等级分级时能取得较好的结果，但在选取时仍会存在一定的错误，因此需对选取结果进行连通性处理，从而确保结果的相对准确。改进的Dijkstra算法能将单一河流连接为水系路径，通过引入3个限制条件来约束水系路径的连接，从而不断修正水系路径的连接情况，具体结果见图4c，通过连通性修正后的河流选取结果不仅能重新连接已删除的要素，在同一制图比例尺下还能让整体的稀疏程度更加均匀。



a 原始数据　b SVM初选取结果　c 连通性校正后选取结果

图4　河流选取结果

以河南省1∶25 000河流矢量数据为例，不同水系路径构建算法的运行时间见图5，当数据量较大时，A*算法能明显提高水系路径的构建效率，而数据量较小时，Dijkstra算法与A*算法的耗时几乎相同。

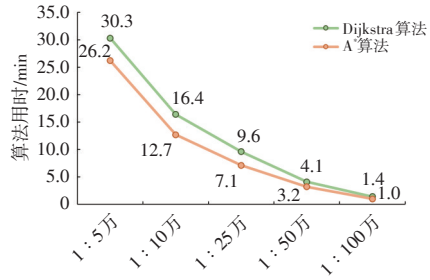
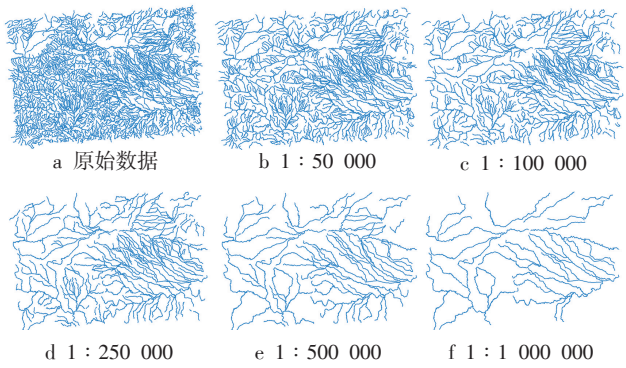


图5　连通性算法用时对比

耦合机器学习与连通性校正的河流自动选取方法在河南省的应用结果见图6。结果表明，本文方法能适应不同制图比例尺，且各比例尺下稀疏程度差异明显，从河流主体结构来看，黄河、淮河、洛河、丹江等重要河流及其重要支流都得以保留，且无中断缺失现象。本文从定量的角度对比了本文方法与专家选取结果的差异（表2），在1∶500 000比例尺中，两种方法的差异最大，达到8.9%，而在1∶100 000和1∶250 000比例尺中，两种方法选取的河流数量基本一致；但从时间角度来看，专家用时基本是本文方法



a 原始数据　b 1∶50 000　c 1∶100 000

d 1∶250 000　e 1∶500 000　f 1∶1 000 000

图6　不同比例尺下河南省河流选取结果

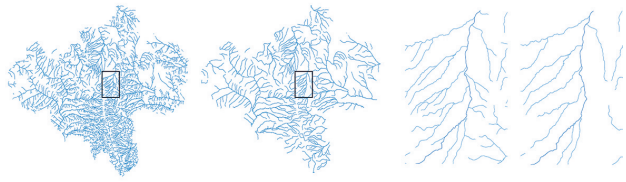
表2　河南省河流选取对比

比例尺	专家选取/条	删除率/%	算法选取/条	删除率/%	选取偏差/%	专家用时/h	算法用时/min
1∶50 000	28 681	82.82	27 234	78.64	−5.05	24.0	32.0
1∶100 000	19 327	55.81	20 279	58.56	4.69	13.0	14.2
1∶250 000	11 936	34.47	11 439	33.03	4.16	7.6	8.3
1∶500 000	4 598	13.28	4 188	12.09	−8.90	3.2	3.9
1∶1 000 000	1 730	5.00	1 608	4.64	−7.05	0.8	1.4

注：专家选取参照GB/T 12343.1−2008、GB/T 12343.2−2008、GB/T 12343.3−2009，其中1∶50 000、1∶100 000酌情舍去图面1 cm以下河流，其余比例尺舍去图面5 mm以下河流；专家用时以专家口述为准，部分用时按工作时间进行大致换算。

的30倍以上，体现了本文方法的高效率性。

为了更好地展示本文方法的实际应用场景，以固原市河流水系进行测试，选取结果见图7，分析可知，1∶100 000比例尺下固原市河系中河流密度显著减少，很多长度较短、等级较低的河流被删除，清水河、葫芦河等主要黄河支流被完整保留，能基本保持河网形状，选取结果能在一定程度上保持河流的流通性与完整性。



a 原始数据　　b 1∶100 000　　c 清水河流域选取结果

图7 河流选取结果对比

固原市各县区河流选取前后要素数量统计见表3，各地区河流删除率均达到35%以上，模型在不同地区的河流数据上均展现出良好的泛化能力，表明模型具有一定的普适性。

表3 固原市和各县区河流选取结果前后对比

区域名称	要素数量/条（选取前）	要素数量/条（选取后）	删除率/%
固原市	4 314	1 899	44.02
原州区	921	590	35.94
西吉县	1 184	703	40.63
泾源县	599	316	47.25
隆德县	778	393	49.49
彭阳县	887	455	48.70

注：由于部分河流跨县区，因此各县区要素数量之和不等于固原市总体。

3 结 语

本文融合河流要素几何、拓扑等特征，以SVM算法为核心构建河流初选取模型，并引入改进的Dijkstra算法与3个限制因子对初选取结果进行连通性校正，最终对二者进行耦合，实现河流要素的自动选取。在利用机器学习算法对河流进行初选取时，随机森林算法与SVM算法均能取得较好的结果，总体一致率能达到88%以上。从河流保留分级来看，SVM算法能有效实现高保留等级河流的准确提取；从运行时间来看，SVM算法能显著节约时间。耦合机器学习与连通性校正的河流自动选取方法能在网状、树状结构的河流数据中进行较准确的自动选取，不仅能在减少河流密度的基础上有效保持河流连通性，还对于不同区域的数据选取展现出良好的泛化能力，且能大幅提升效率。因此，本文方法有望在河流数据自动化处理和地图综合工作中发挥重要作用，满足不同比例尺下的数据管理与成图需求。

本文存在的不足表现为：①本文方法是基于单线河设计的，可能不适用于双线河数据；②由于SVM算法的局限性，测试结果中仍存在水系分级不合理的情况；③在机器学习中特征库的构建十分重要，本文方法选取的特征仅根据相关研究和前期实验确定，人为影响较大。在后续研究中，将进一步改善河流初选取算法，利用深度学习发掘水系要素的深层特征，并进行特征优选确定最优特征组合，尽可能地减轻人工设计特征的干扰；同时尝试将双线河与单线河进行统一综合，以增强其连续性，以期实现更好的水系综合效果。

参考文献

[1] Fan Y T, Yang J Y, Zhang C, et al. A Event-based Change Detection Method of Cadastral Database Incremental Updating [J]. Mathematical and Computer Modelling, 2010,51(11):1 343-1 350

[2] 熊顺,杜清运,马超,等. 面向地图综合的VGI道路网重要度层次分解选取方法研究[J]. 地球信息科学学报,2024,26(1):135-143

[3] 廖志强. 基于多比例尺地图数据库的智能制图体系探索与实践[J]. 测绘与空间地理信息,2024,47(10):54-58

[4] 张亮,闵天,胡菡,等. 1∶2 000到1∶10 000数字线划图缩编技术研究[J]. 地理空间信息,2024,22(9):117-120

[5] 王文宁,闫浩文,禄小敏,等. 地图上的河系自动综合研究综述[J]. 地理与地理信息科学,2021,37(5):1-8

[6] 常利伟,程磊,陈进,等. 水库和大坝地理信息数据集的对比及其应用研究进展[J]. 中国农村水利水电,2023(6):1-11

[7] 武芳,巩现勇,杜佳威. 地图制图综合回顾与前望[J]. 测绘学报,2017,46(10):1 645-1 664

[8] 武芳,谭笑,翟仁健,等. 基于知识表达和推理的河网自动选取[J]. 辽宁工程技术大学学报,2007,26(2):183-186

[9] 陆向珍,康二梅. 几何特征重要度算法支持下的大规模线状要素地图综合方法[J]. 测绘通报,2024(6):176-181

[10] Knura M. Learning from Vector Data: Enhancing Vector-based Shape Encoding and Shape Classification for Map Generalization Purposes[J]. Cartography and Geographic Information Science,2024,51(1):146-167

[11] Sen A, Gokgoz T. An Experimental Approach for Selection/Elimination in Stream Network Generalization Using Support Vector Machines[J]. Geocarto International, 2015, 30(3):311-329

（下转第84页）