文章编号：1001-1749(2023)03-0389-10

因子组织方式对 SVM 滑坡危险性评价影响的研究

谢华伟[1]，陈建华[1]，甘先霞[1]，许开行[2]，赵 铮[3]
(1.成都理工大学 地球物理学院，成都 610059；
2.四川省华地建设工程有限责任公司，成都 610081；
3.中国科学院 地理科学与资源研究所，北京 100101)

摘 要： 滑坡是一种在世界范围内广泛分布的地质灾害，每年对经济和人民生命造成巨大的损失，开展滑坡危险性评价对滑坡防范和治理意义重大。关于哪种数据组织形式更适合进行滑坡危险性评价，以往的研究中鲜有提及，因此这里以支持向量机模型作为评价模型，以四川省芦山县作为研究区，基于因子分级与否两种数据形式进行滑坡危险性评价，采用验证集 ROC 曲线、全区评价结果分布特征、成功率曲线、直方图索引等指标进行结果对比，探索因子分级与否对评价结果的影响。模型训练时，两种数据构建的模型在验证集上的 *AUC* 值分别为 0.893(分级值)和 0.813(连续值)。将两个模型应用在全区危险性评价中，采用自然断点法将结果分为低、中、高三级，其中高危险性面积占比分别为 19.3%(分级值)和 20.4%(连续值)。采用成功率曲线和滑坡危险性索引直方图两个指标对该结果进行评估，分级数据 *AUC* 值为 0.80，连续数据 *AUC* 值为 0.76。直方图对比显示，分级值的结果直方图特征更加符合预期。综上所述，通过多指标对比，得出分级因子更适用于进行基于 SVM 的滑坡危险性评价的结论。

关键词： 数据组织形式；支持向量机；多重共线性分析；滑坡危险性评价；多指标对比
中图分类号： P 642.22 **文献标志码：** A **DOI：**10.3969/j.issn.1001-1749.2023.03.14

0 引言

滑坡灾害在世界范围内广泛发生，造成了难以估量的巨大损失，应用科学的方法对滑坡灾害进行评估，制作危险性分级区划图，有利于了解滑坡分布特征、发生规律，也有利于灾害防治及预警，降低其带来的损失。

区域滑坡危险性评价是包含地理信息系统，遥感、机器学习、统计学等多门学科知识的一种评价方法[1-3]，能够快速、大面积得到区域滑坡危险性分级图，为相关部门进行风险管理和防治规划提供辅助决策的信息。

近年来，国内、外诸多学者在滑坡危险性评价研究中提出了许多模型，例如层次分析法[4]，信息量法[2]，广义线性模型[5]、频率比[6-7]，以及基于机器学习[8]的 BP 神经网络[9]、多层感知机[10]、卷积神经网络[11]、随机森林[1]、逻辑回归[1]、支持向量机(Support Vector Machines，SVM)[12-13]等。其中 SVM 作为一种有效的二分类模型，能够在高维特征空间中较好的将变量划分到对应的类别中[14]，在解决小样本、高维度、非线性问题方面比其他学习方法更合理、更有效，被广泛应用在滑坡危险性评价中。

滑坡危险性评价是与滑坡影响因子紧密关联的

收稿日期：2022-01-24
基金项目：四川省科技计划项目(2019YFG0187)
第一作者：谢华伟(1997－)，男，硕士，研究方向为空间分析模型与方法，E-mail：mts8939@163.com。
通信作者：陈建华(1976－)，男，副教授，硕士生导师，主要从事空间模型分析研究，E-mail：chjh3@163.com。

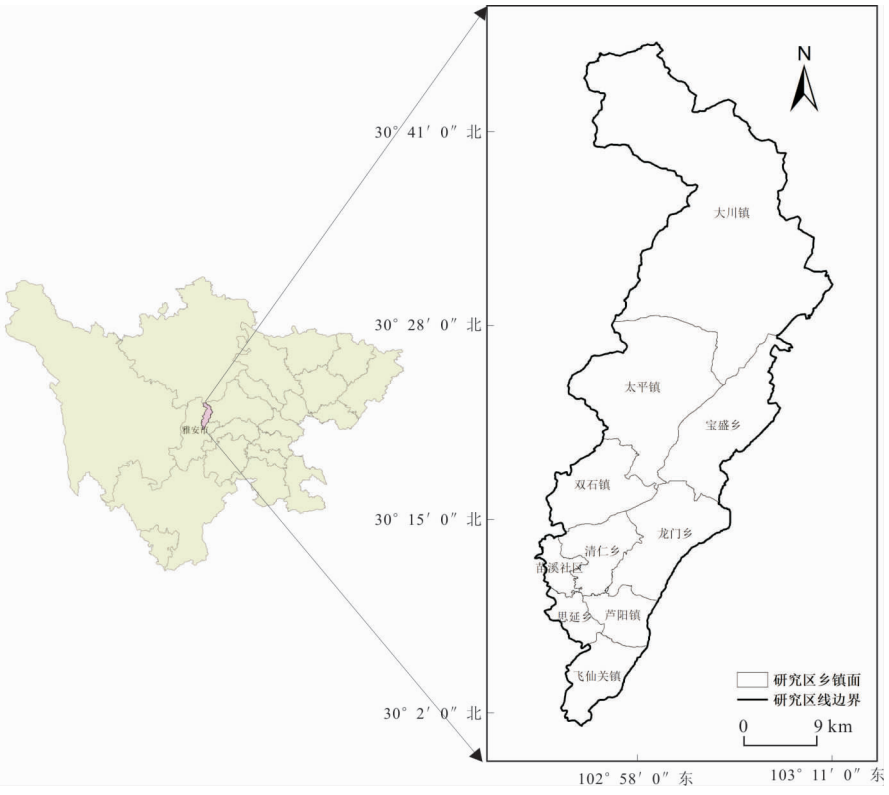

图 1 研究区地理位置

Fig. 1 Location of the study area

分析方法，在进行评价前需要根据研究区特征选择因子。目前所查阅文献中，有研究将因子数据统一划分为类别数据[15-17]，同时也有部分学者在研究中保留因子的原始值，这样的数据集通常同时包含了数值型因子和类别型因子。数值型因子是往往是连续的，因子数值大小具有实际的含义（如降雨量、坡度、坡向等）。而类别型因子每个类别表示研究区一种特征，在计算时往往用数字来表示不同类别，此时数字仅表明类型，不具有运算特征，在计算中可采用热编码等方式消除数值量纲影响。

上述两种数据组织方式所得到的因子数据具有不同的数据特征，基于该数据集的滑坡危险性评价结果应当也有差异，但是尚无人开展这方面的研究。因此笔者以四川省雅安市芦山县为研究区，选择该区域 12 个滑坡影响因子，以 SVM 为评价模型，对比分析不同数据组织方式对滑坡危险性评价结果的影响。

1 研究区及数据

1.1 研究区简介

芦山县位于四川盆地西部，雅安市东北部，总面积约 1 191 km^2，该区域位于龙门山地质构造带，地质构造发育，地形以丘陵、山地为主，南北高差大。芦山县属于中纬度内陆亚热带湿润气候，因南北地势高差悬殊，气候垂直变化显著。全县年均降水量为1 313.1 mm，自南向北递减，雨量多集中于七月、八月、九月份。该县 2013 年发生里氏 7.0 级地震，并在这几年频遭极端降雨，区域内滑坡等地质灾害频发。因此，选择该区域作为研究区，研究区地理位置如图 1 所示。

1.2 数据

滑坡危险性评价建模依赖于历史滑坡点和对应点位上的影响因子之间的非线性关系[17]。这些影响因子通常包括地层岩组、人类活动、气候特征、地形地貌等。笔者综合分析了研究区滑坡发生机制，从研究区地形地貌特征以及数据可获取性方面考虑，选择了地层岩组、距断层距离、高程、坡度、坡向、距水系距离、植被覆盖度、人口密度、距路网距离、土地利用、降雨、距地震点距离共 12 个因子。

上述因子数据中，坡度、坡向、高程数据使用 GIS 软件从 ASTER DEM（https://www.gscloud.cn/）中提取得到；植被覆盖度和土地利用从 Landsat8 卫星影像（http://earthexplorer.usgs.gov）中提取；距地震点距离中的地震点来源于中国地震台

网(https://news.ceic.ac.cn/)；断层和地层岩组来源于1:250 000宝兴幅地质图；降水数据来源于雅安市公共气象服务中心；人口数据来源于芦山县公安局；路网数据来源于芦山县交通运输局。

2　方法

2.1　支持向量机模型

SVM 模型是 Vapnik 基于统计学习理论提出来的模型，其基于最小化结构误差原则以及最优化理论寻找全局最优解决方案[19-20]。使用 SVM 进行二分类的基本思想是将输入的数据点映射到更高维的特征空间，找到最优超平面实现类别划分[20-22]。

支持向量机模型的超平面可以用公式(1)表示。

$$f(x)=\boldsymbol{\omega}^{\mathrm{T}}x+b \tag{1}$$

其中：$\boldsymbol{\omega}$ 为支持向量；b 为截距。当 $f(x)$ 等于 0 时，表明该点是超平面上的点，$f(x)>0$ 的点对应类别为 1(滑坡)，$f(x)<0$ 的点对应类别 −1(非滑坡)，为了正确划分两个类别，需要最大化两个类别之间的最小距离。数据点到超平面的间距表示如下：

$$|\boldsymbol{\omega}^{\mathrm{T}}x+b| \tag{2}$$

为了判断分类是否正确，引入函数间隔概念[21-22]：

$$\hat{\gamma}=y(\boldsymbol{\omega}^{\mathrm{T}}x+b) \tag{3}$$

其中：y 表示样本类别，当 y 与 $\boldsymbol{\omega}^{\mathrm{T}}x+b$ 符号一致时，表明分类正确，反之则是分类错误。基于式(1)至式(3)，求解最小间距即变为在数据集 (x_i,y_i) 下求解最小 $\hat{\gamma}$。为避免量纲影响，在实际求解时，要除以 $\boldsymbol{\omega}$ 的范数，表示为式(4)。

$$\gamma=\frac{\boldsymbol{\omega}^{\mathrm{T}}x+b}{\|\boldsymbol{\omega}\|} \tag{4}$$

为了求式(4)的绝对值，在公式两侧同时乘上类别值 y，则可以得到以下关系：

$$\hat{\gamma}=y\gamma=\frac{\hat{\gamma}}{\|\boldsymbol{\omega}\|} \tag{5}$$

因此我们要求类别点到超平面的最大间隔用公式(6)表示。

$$\max\frac{1}{\|\boldsymbol{\omega}\|} \tag{6}$$

求解公式(6)后即可得到对应样本集下的超平面，完成 SVM 构建。

2.2　数据预处理

初始收集到的数据包含了多种格式，例如人口数据初始格式为 excel 表格，遥感影像数据为 L1 级原始数据，地震点数据为 excel 表格等。为了进行后续模型构建以及对比分析，需要对数据进行相应处理。对于表格数据，通过其自带的坐标信息生成图层，并转换为栅格格式。对于其他数据，采用相应的数据处理软件处理后，统一设置坐标为 WGS84 坐标系，并将所有栅格数据统一为 30 m 分辨率，通过设置捕捉栅格确保不同因子的栅格对齐。

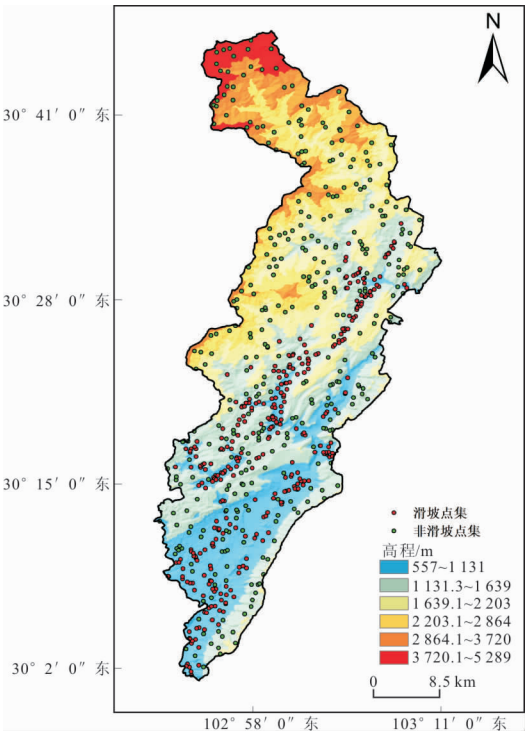

图 2　样本集

Fig. 2　Sample set

2.3　样本集生成

每个滑坡危险性评价模型都需要包含滑坡点点位、类别等信息的滑坡编录数据，用以研究滑坡发生机制并分析其与滑坡影响因子之间的关系[23]。笔者所使用的 2015 年滑坡点由芦山县自然资源与规划局提供，所获取到的滑坡点中包含了面积信息，但是没有滑坡的具体形态，因此将每条滑坡记录视为一个滑坡点。在研究区滑坡点 500 m 范围外随机生成等量非滑坡点，从而得到所用 706 个样本点。这里所使用的样本点数据如图 2 示，可以发现滑坡点主要集中分布在研究区低、中高程，而非滑坡点由于选择的随机性，在研究区均匀分布。

2.4　连续因子值提取

连续因子值是指进行模型训练时，使用对应点

位的原始因子值。在上述的 12 个因子数据值中，土地利用、地层岩组两个因子为类别数据，数据表示为 1、2、…、n，在此数字仅表示类别不具有数值含义。对于路网、水系、地震点、断层等几个因子，它们对滑坡的影响表现为在其周边一定范围内更容易发生滑坡，当超过一定范围时，对滑坡的影响减少[1, 12, 14, 24]。因此在本文中采用各个点到这几个因子的距离作为对应因子的值。

除上述几个需要计算距离的因子外，其余因子按照上文所述统一投影系统和统一分辨率的操作，将栅格数据分辨率统一为 30 m，然后将因子值提取到对应的每个评价单元中，再从全区评价单元导出样本集。

2.5 分类因子值提取

在目前已经完成的滑坡危险性评价研究中，有不少学者将各个因子数据进行分类，采用类别值进行模型训练[15, 17]。为了对比连续值和分类值两种类型对结果的影响，笔者对 10 种非类别数据进行分类，已经分类的数据保留原始值。

关于滑坡危险性评价中的因子分类，不同研究分类数量和分类标准不同，目前尚无统一标准。虽然有学者在文献中提到分为 8 类比较合适[25]，但是由于各个研究区地理特征差异，因此更多学者按照滑坡分布特征来进行分类[17]，笔者也采用这种思路进行分类，具体分类流程是：参考其他学者文献中的因子类别数对各个需要分类的因子进行初步分类，分类后根据各个类别中滑坡点分布情况进行类别调整，进而得到最终分类结果。以路网为例，将该因子按照到路网的距离远近初步分为 10 类，根据前人研究经验，在路网缓冲区内，随着到路网距离的增加，各个类别内滑坡点数量具有先增加后减少的趋势。而笔者对路网的初步分类中滑坡点具有上述特征，因此不做类别调整，其他因子分类与此类似。所选因子具体分类信息如表 1 所示，各个因子内滑坡呈一定规律分布。

3 模型构建与应用

3.1 因子分析

在进行滑坡危险性评价时，因子之间信息冗余会引入噪声影响模型性能，因此在进行模型训练之前，需要对因子进行筛选[26-27]。笔者使用方差膨胀系数（Variance Inflation Coefficient，VIF）来衡量因子之间的相关性。两种格式组织的数据通过多重共线性分析，得到的结果如表 1 所示。

根据前人经验所述，VIF 小于 10 h，可以认为因子之间是相互独立，不存在相关性，因此 12 个因子可以全部保留作为模型输入。

为了对比两种格式数据之间的差异，将两类因子的 VIF 值绘制成折线图（图 3），两种形式数据的 VIF 极值虽然有差异，但是极值对应的都是同一个因子，且两条曲线起伏大致相似，足以说明虽然改变了数据组织形式，但是因子数据对滑坡的影响没有随着数据组织格式变化而改变。

3.2 模型训练与分析

使用 SVM 作为评价模型计算两种数据对滑坡危险性评价结果的影响。为了得到具有高鲁棒性的模型，在进行模型构建之前，采用分层抽样法将上述样本集按照 8:1:1 的比例划分为训练集、测试集以及验证集三个部分。

表 1 两种数据组织格式下因子 VIF 对比

Tab. 1 Comparison of factor VIF in two data organization formats

因子(类别数)	分级 VIF	连续 VIF	因子(类别数)	分级 VIF	连续 VIF
坡向(8)	1.129	1.126	人口密度(9)	2.088	1.320
距地震点距离(10)	1.120	2.928	年均降雨量(9)	1.014	1.006
高程(6)	3.171	5.907	坡度(6)	1.731	1.672
距断层距离(10)	1.831	1.847	地层岩组(52)	1.121	1.159
土地利用类型(8)	1.270	1.327	植被覆盖度(9)	1.642	1.663
距路网距离(11)	1.058	2.543	距水系距离(11)	1.090	1.160

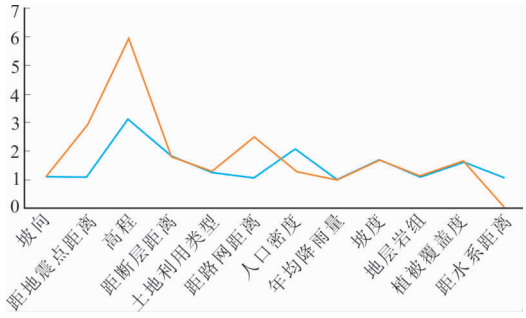

图 3　两种数据组织格式下因子 VIF 对比

Fig. 3　Comparison of factor VIF in two data organization formats

SVM 实现使用的是 sklearn 库，该库是基于 Python 的机器学习工具，可快速构建多种机器学习模型。构建 SVM 时，利用了径向基函数核，主要需要设置的参数是惩罚因子 C。采用交叉验证（Cross Validation，CV）[18] 与网格寻参相结合方式寻找最优参数 C。

交叉验证方法[17] 是机器学习模型中小样本情况下常用的训练模型方法。其核心思想是将样本等量划分为多份，每一次训练时，将其中一份作为测试数据，剩下的部分作为训练集进行模型参数训练。在本文中，以上述 80％的样本集作为训练数据进行交叉验证网格寻参。

曲线下面积（Area Under Curve，AUC），其值介于 0.5－1 之间[17]，能直观反映一个模型的好坏，使用其作为模型训练时的评估指标。此处曲线是指受试者特征曲线（Receiver Operating Characteristic，ROC），该曲线被广泛应用在各种模型评估中[2, 15, 18, 23]。

使用 CV 网格寻参时，容易存在过拟合现象，即模型在训练集上有较高的 *AUC* 值，而将模型推广应用时的效果却很差。为了解决这一问题，将交叉验证的每一次 *AUC* 值输出。以 *C* 值为横坐标，*AUC* 值为纵坐标，得到参数交叉验证曲线，选择测试集曲线拐点处的 *C* 值作为最终模型的 *C* 值，这样能够有效避免过拟合现象。这里用两种数据组织格式下的参数交叉验证图如图 4 所示。

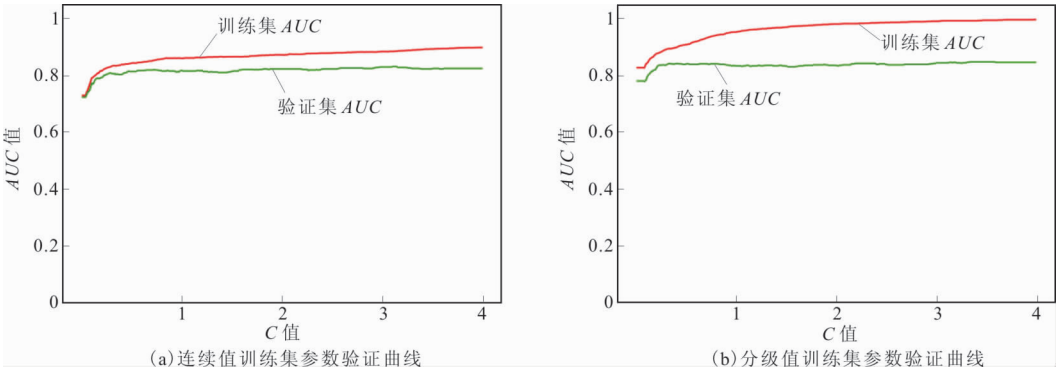

图 4　两种数据格式参数训练结果

Fig. 4　Two data format parameter training results

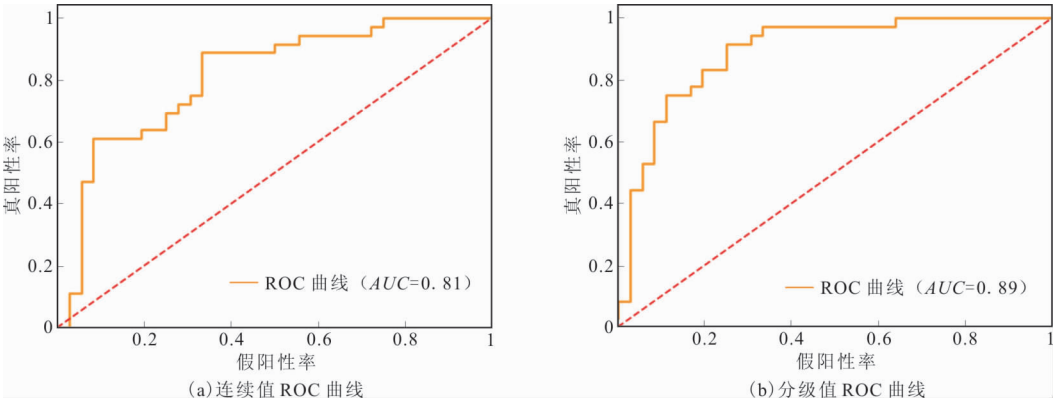

图 5　两个模型在验证集上 ROC 曲线

Fig. 5　ROC curves of the two models on the validation set

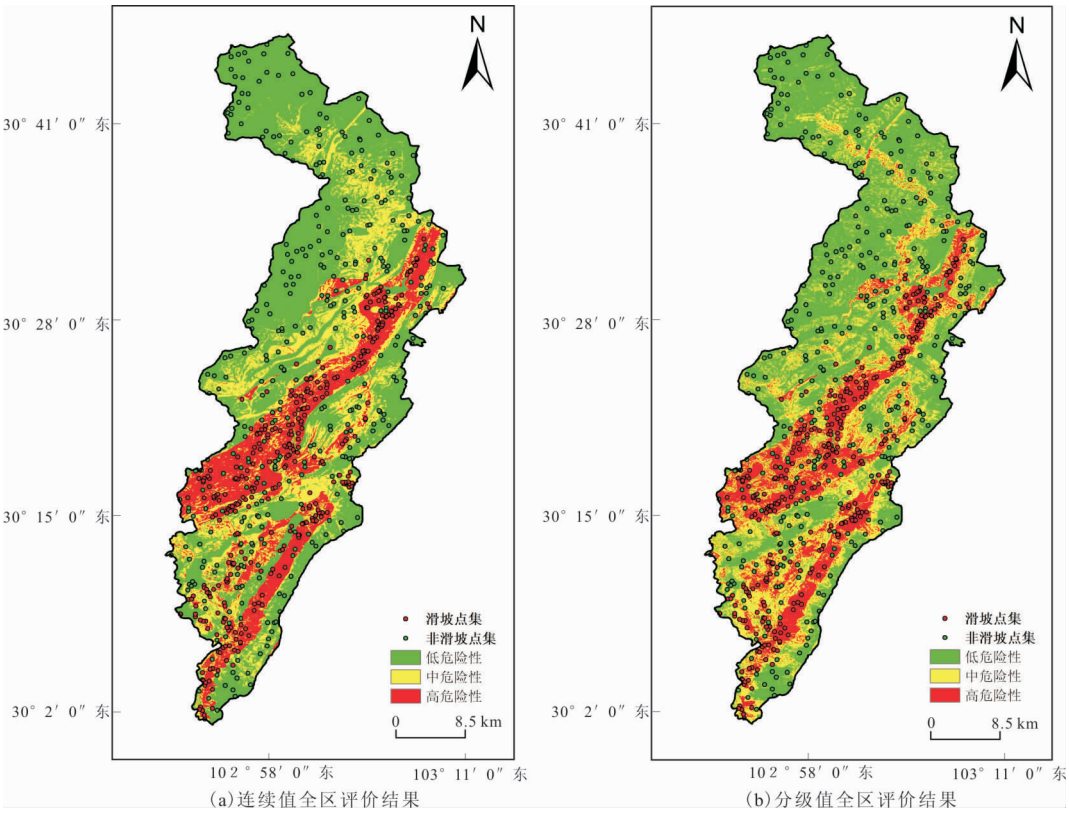

图 6 两种模型全区评价结果

Fig. 6 Evaluation results of the two models in the whole region

表 2 两个模型在验证集上指标对比

Tab. 2 Comparison of indicators between the two models on the validation set

模型	*AUC*	准确率	F1 得分
连续值模型	0.813	0.708	0.708
分级值模型	0.893	0.805	0.805

图 4(a)是连续值下的参数验证图，图 4(b)是分级值下的参数验证图。从图 4 中可以看出，训练集和测试集上两条曲线趋势接近一致。模型评估得到的最优参数 C 分别为 3.24(连续值)和 1.3(分级值)，最优参数对应的训练集上 *AUC* 值分别为 0.915、0.923。对应在图中位置，可以发现在这两个值附近，参数验证曲线有轻微起伏，且在这之后无明显变化，表明所选择值比较合理。

模型参数确定后，使用该参数重新构建模型，并将模型使用测试数据集进行重新训练，为了进一步验证模型有效性，将构建的模型应用在验证数据集上，使用 *AUC*、准确率[17]、F1 得分[28]几个指标进行效果评估。两个模型在验证集上的 ROC 曲线如图 5 所示。两个模型在验证集上的指标结果如表 2 所示。

由连续因子训练的模型在验证集上的 *AUC* 值为 0.813，而由分级因子训练的模型在验证集上的 *AUC* 值为 0.893。在滑坡危险性评价中，精度每提升一个百分点，对结果的改善都十分明显[25]。在上述的对比中可以发现，分级值模型的各个指标值都明显大于连续值模型的指标，这说明采用分级因子更加有利于构建基于 SVM 的区域滑坡危险性评价模型。

3.3 全区评价结果

进行不同格式数据组织，是为了确定更优的数据组织形式进而应用于区域滑坡危险性评价。因此，两种数据组织形式训练出来的 SVM 模型除了上述几个指标对比，还需要将模型应用到全区危险性评价中，综合对比两种模型的评价结果。

将全区 12 个因子按照上述两种形式组织后分别输入两个模型中，得到全区滑坡危险性索引。其值在 0—1 之间，表征滑坡发生的可能性，值越接近 1，则说明发生滑坡的可能性越大，反之则说明发生滑坡的可能性较小[17]。得到危险性索引后，将其按照一定指标分类，即可得到研究区滑坡危险性分级

表 3　两个模型全区评价分级结果对比

Tab. 3　Comparison of the evaluation results of the two models in the whole region

模型	危险性级别	各级格网数	面积占比/%	滑坡点数	滑坡点占比/%
连续值模型	低	720440	54.4	43	12.2
	中	333614	25.2	95	26.9
	高	269641	20.4	215	60.9
分级值模型	低	686099	51.8	25	7.1
	中	381835	28.8	90	25.5
	高	255760	19.3	238	67.4

图。自然断点法由于其与环境有逻辑一致性[14]，笔者选择其作为分级方法将研究区滑坡危险性索引分为：高、中、低三级（图 6）。

图 6(a)是连续值模型结果，图 6(b)是分级值结果。该结果表明，两个模型的结果整体相似，且都有明显的因子痕迹，这是由 SVM 内在机理所导致。高危险性主要集中在研究区南部和中部狭长地带，整体形态差异不大，中危险性围绕高危险性区域分布，而低危险性主要分布在人口稀少的山区。但它们细节上有较大差异，连续值的结果中高危险性区域分布比较集中，该区域内较少夹杂其他危险性级别。而分级值的结果中，高危险性区域虽然整体比较集中，但是其内部分布相对离散，夹杂了一些中危险性区域；对于中危险性区域，连续值的结果分布也相对比较集中，在北部除了中心地带，边缘地带较少有分布。分级值的结果分布离散，且分布范围较广。在实际自然环境中，滑坡分布通常比较离散，滑坡危险性评价是基于研究区自然本底因子得到的结果。通常情况下，潜在滑坡发生区域空间分布通常也是离散分布，同时，在前人研究成果中[1, 4, 12, 14, 17, 24, 29]，高危险性区域多呈离散状态分布，因此分级值的全区评价结果比连续值的评价结果更优。

4　结果分析

总结前人对区域滑坡危险性评价的研究可以发现，好的评价结果通常情况下各个危险性类别的面积由低到高呈递减趋势，危险性级别越高，所占研究区的面积越少，其内分布的滑坡点数越多。这里的两个模型得到的危险性评价分级图中，各级别面积以及其滑坡分布情况如表 3 所示。

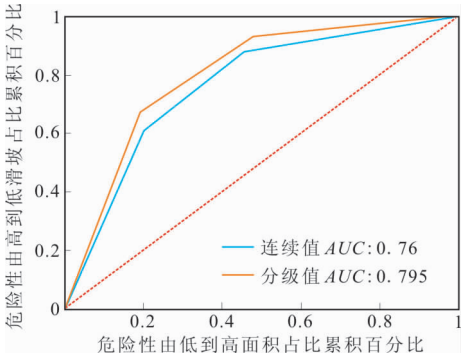

图 7　全区评价结果成功率曲线

Fig. 7　The success rate curve of regional evaluation results

两类数据训练出来的模型的评价结果中，各个级别危险性面积占比均由低危险性向高危险性呈递减趋势，高危险性面积占比分别为 19.3%（分级值）和 20.4%（连续值），其内分布了研究区绝大部分滑坡。结合对模型精度的分析，充分说明 SVM 在区域滑坡危险性评价中的可行性。

成功率曲线是一种以危险性由高到低面积累积百分比为横坐标，各个级别内滑坡点占比累积百分比为纵坐标绘制的曲线，常用于滑坡危险性分级区划评估[30]。笔者引入成功率曲线来对两种模型的评价结果进行评估（图 7），两种形式数据训练的模型在全区评价结果的 *AUC* 分别为 0.80 和 0.76，分级值模型的结果要明显优于连续值结果。

滑坡危险性索引的直方图也被广泛应用到滑坡危险性评价结果分析中[25]。在两个模型的索引直方图中（图 8），两个模型在低概率段的直方图特征比较相似，但是在中高概率值区域，连续值的评价单元数量先下降而后增加，分级值的数量总体比较均衡，更加符合我们的预期。

综上，通过成功率曲线对比和直方图对比，分级

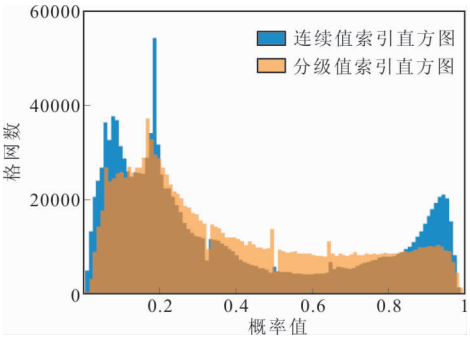

图 8 滑坡危险性索引直方图

Fig. 8 Histogram of landslide risk index

值模型全区评价结果的 *AUC* 明显高于连续值模型结果，两个模型的危险性索引直方图中分级因子直方图反映的数据特征更加符合我们预期，因此将因子数据进行分级处理更加适用于区域滑坡危险性评价。这是由于连续因子的值的范围变化较大，本文中仅考虑 10 个连续因子，其值的变化范围从 0.1 到 10 116，变化很大。虽然在模型训练之前进行相关处理，尽可能减小了数据量级带来的影响，但是模型仅基于 564 个样本处的因子进行训练，这部分因子的值不能充分表达全区数据特征，从而导致连续值模型各个指标相对较差。反之，分级形式不仅兼顾了各个因子对滑坡的影响，还减小数据量级差异，从而在模型训练时，样本能表达全区更多特征，得到更优结果。

5 结论

笔者对比了两种数据组织形式下区域滑坡危险性评价结果的差异。以四川省芦山县为研究区，构建评价单元，选择了该区域降雨、高程等 12 个影响因子，将因子组织为分级和连续两种形式。为避免因子信息冗余，对两类因子进行多重共线性分析，结果表明所选择因子相互独立。将样本集按照 8∶1∶1 的比例划分为训练集、测试集、验证集。采用 CV 网格寻参寻找最优参数 *C*。参数训练完成后，用最优参数重新构建模型，用验证集进行验证，采用 *AUC*、准确率等指标评估。然后基于两个模型进行芦山滑坡危险性评价，得到滑坡危险性索引，用自然断点法将结果分为低、中、高三级。统计每个级别面积占比和滑坡占比、绘制成功率曲线、绘制滑坡危险性索引直方图，综合这几个指标对比分析。

对比结果显示，在进行模型构建时，分级因子构建的模型在验证集上的 *AUC* 值、准确率等指标均高于连续值的结果。全区评价结果对比中，分级值模型结果的高危险性面积占比更低，其内分布的滑坡点数更多，且分级区划图中各个类别分布离散，更加符合现实情况；分级值模型成功率曲线的 *AUC* 值比连续值模型高；连续值模型的结果在危险性索引直方图中，高概率范围内评价单元数量剧增，与预期不符。而分级值模型中，大多数评价单元具有较低概率值，在概率较高部分变化比较小，评价单元数量也较少。

综上述，经过对比实验，可以确定在基于 SVM 的滑坡危险性评价过程中，采用因子分级的方式作为模型输入能得到更优评价结果。但是实验还存在以下不足之处：①分级值模型的结果中有明显的环形痕迹，该环形痕迹是地震因子某个类别的特征，这固然有 SVM 的内在机理因素，但是终究对结果产生了影响；②关于因子分级的标准问题，目前的诸多研究中，虽然有很多学者用到了分级因子作为模型输入，但是对于间距怎么划分、分多少类的问题还没有明确的标准，笔者也只是采用了其中比较常用的一种标准来划分；③笔者仅采用了 SVM 来进行对比，但是目前机器学习领域提出了许多其他的滑坡危险性评价模型，应该多使用几种机器学习方法对比，得到更可靠的结论。因此在后续的研究中，要对比更多的机器学习模型，同时要完善因子分级标准，使得分级更加合理。

参考文献：

[1] CHEN W, XIE X, WANG J, et al. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility [J]. Catena, 2017(15): 147－160.

[2] LI D, HUANG F, YAN L, et al. Landslide susceptibility prediction using particle－swarm－optimized multilayer perceptron: comparisons with multilayer－perceptron－only, BP neural network, and information value models [J]. Applied Sciences, 2019, 9(18): 1－18.

[3] POURGHASEMI H R, RAHMATI O. Prediction of the landslide susceptibility: Which algorithm, which precision? [J]. Catena, 2018(162): 177－192.

[4] ABEDINI M, GHASEMYAN B, REZAEI MOGADDAM M H. Landslide susceptibility mapping in Bijar city, Kurdistan Province, Iran: a comparative study by logistic regression and AHP models [J]. Environmental Earth Sciences, 2017, 76(8): 1－14.

[5] SHAHABI H, AHMAD B B, KHEZRI S. Evaluation and comparison of bivariate and multivariate statistical methods for landslide susceptibility mapping (case study: Zab basin) [J]. Arabian Journal of Geosciences, 2012, 6(10):3885－907.

[6] LI L, LAN H, GUO C, et al. A modified frequency ratio method for landslide susceptibility assessment [J]. Landslides, 2016, 14(2):727－741.

[7] ZHANG Y X, LAN H X, LI L P, et al. Optimizing the frequency ratio method for landslide susceptibility assessment: A case study of the Caiyuan Basin in the southeast mountainous area of China [J]. Journal of Mountain Science, 2020, 17(2):340－357.

[8] SAHA S, ARABAMERI A, SAHA A, et al. Prediction of landslide susceptibility in Rudraprayag, India using novel ensemble of conditional probability and boosted regression tree－based on cross－validation method [J]. Sci Total Environ, 2021(764):1－20.

[9] WU X, NIU R, REN F, et al. Landslide susceptibility mapping using rough sets and back－propagation neural networks in the Three Gorges, China [J]. Environmental Earth Sciences, 2013, 70(3):1307－18.

[10] DAO D V, JAAFARI A, BAYAT M, et al. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility [J]. Catena, 2020(188):1－13.

[11] FANG Z, WANG Y, PENG L, et al. Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping [J]. Computers & Geosciences, 2020(139):1－15.

[12] DURIC U, MARJANOVIC M, RADIC Z, et al. Machine learning based landslide assessment of the Belgrade metropolitan area: Pixel resolution effects and a cross－scaling concept [J]. Engineering Geology, 2019(256):23－38.

[13] KUMAR D, THAKUR M, DUBEY C S, et al. Landslide susceptibility mapping & prediction using support vector machine for Mandakini river basin, Garhwal Himalaya, India [J]. Geomorphology, 2017(295):115－125.

[14] DOU J, YUNUS A P, BUI D T, et al. Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan [J]. Landslides, 2019, 17(3):641－658.

[15] PHAM B T, JAAFARI A, PRAKASH I, et al. A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling [J]. Bulletin of Engineering Geology and the Environment, 2018, 78(4):2865－86.

[16] PANAHI M, GAYEN A, POURGHASEMI H R, et al. Spatial prediction of landslide susceptibility using hybrid support vector regression (SVR) and the adaptive neuro－fuzzy inference system (ANFIS) with various metaheuristic algorithms [J]. Sci Total Environ, 2020(741): 1－14.

[17] ABEDINI M, GHASEMIAN B, SHIRZADI A, et al. A comparative study of support vector machine and logistic model tree classifiers for shallow landslide susceptibility modeling [J]. Environmental Earth Sciences, 2019, 78(18): 1－15.

[18] SCHRATZ P, MUENCHOW J, ITURRITXA E, et al. Hyperparameter tuning and performance assessment of statistical and machine－learning algorithms using spatial data [J]. Ecological Modelling, 2019(406):109－120.

[19] VAPNIK V. The nature of statistical learning theory [M]. New York: Springer science & business media, 1999. 57－60.

[20] 胡安龙，王孔伟，李建林，等. 基于智能算法优化支持向量机模型的滑坡稳定性预测 [J]. 自然灾害学报，2016，25(05)：46－54.
HU A L, WANG K W, LI J L, et al. Landslide stability prediction based on intelligent algorithm optimization support vector machine model [J]. Journal of Natural Disasters, 2016, 25(05): 46－54. (In Chinese)

[21] 尹嘉鹏. 支持向量机核函数及关键参数选择研究 [D]. 哈尔滨：哈尔滨工业大学航天学院，2016：1－z.
YIN J P. Research on selection of kernel functions and key patameters in support vector machine [D]. Harbin: Harbin Institute of Technology School of Astronautics, 2016:1－z. (In Chinese)

[22] 张凯军，梁循. 一种改进的显性多核支持向量机 [J]. 自动化学报，2014，40(10):2288－94.
ZHANG K J, LIANG X. An improved explicit multi－core support vector machine [J]. Acta Automatica Sinica, 2014, 40(10): 2288－94. (In Chinese)

[23] JAAFARI A, NAJAFI A, POURGHASEMI H R, et al. GIS－based frequency ratio and index of entropy models for landslide susceptibility assessment in the Caspian forest, northern Iran [J]. International Journal of Environmental Science and Technology, 2014, 11(4):909－926.

[24] ARABAMERI A, PRADHAN B, REZAEI K, et al.

GIS－based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms [J]. Journal of Mountain Science, 2019, 16(3):595－618.

[25] HUANG F, CAO Z, GUO J, et al. Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping [J]. Catena, 2020, 191:1－14.

[26] 毛伊敏，陈华彬，李忠利，等. 不确定模糊 ID3 算法及其在滑坡危险性评价中应用研究 [J]. 江西理工大学学报，2017，38(05):92－98.
MAO Y M, CHEN H B, LI ZH L, et al. Uncertain fuzzy ID3 algorithm and its application in landslide risk assessment[J]. Journal of Jiangxi University of Science and Technology, 2017, 38(05): 92－98. (In Chinese)

[27] 孙继平，佘杰. 基于支持向量机的煤岩图像特征抽取与分类识别 [J]. 煤炭学报，2013，38(S2)：508－512.
SUN J P, SHE J. Feature extraction and classification and recognition of coal and rock images based on support vector machines [J]. Journal of China Coal Society, 2013, 38(S2):508－512. (In Chinese)

[28] TEHRANY M S, JONES S, SHABANI F, et al. A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi－source geospatial data [J]. Theoretical and Applied Climatology, 2018, 137(1－2): 637－653.

[29] GOROKHOVICH Y, VUSTIANIUK A. Implications of slope aspect for landslide risk assessment: A case study of Hurricane Maria in Puerto Rico in 2017 [J]. Geomorphology, 2021(391):1－10.

[30] 颜阁. 华池县滑坡易发性制图 [D]. 兰州：兰州大学地质科学与矿产资源学院，2016.
YAN G. Mapping of landslide susceptibility in Huachi county [D]. Lanzhou: Lanzhou University. School of Earth Sciences, 2016. (In Chinese)

Research on the influence of factor organization on the SVM results of regional landslide risk assessment

XIE Huawei[1], CHEN Jianhua[1], GAN Xianxia[1], XU Kaihang[2], ZHAO Zheng[3]

(1. Chengdu University of Technology, College of Geophysics, Chengdu 610059, China;
2. Sichuan Huadi Construction Engineering Co., Ltd., Chengdu 610081, China;
3. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China)

Abstract: A landslide is a kind of geological disaster widely distributed worldwide and causes considerable losses to the economy and people's lives yearly. Carrying out a landslide risk assessment is significant to landslide prevention and treatment. However, previous studies rarely mention which data organization is more suitable for landslide risk assessment. Therefore, The landslide risk evaluation is carried out support vector machine model as the evaluation model and Lushan County, Sichuan Province, as the research area, based on two factors: factor classification or not, and the results are compared with indicators such as the validation set ROC curve, the distribution characteristics of the evaluation results of the whole area, the success rate curve, and the histogram index to explore the impact of factor grading on the evaluation results. During model training, the AUC values of the models constructed by the two data types on the validation set were 0.893 (grading value) and 0.813 (continuous value), respectively. The two models were applied to the risk assessment of the whole district, and the natural breakpoint method was used to divide the results into three levels: low, medium, and high. Among them, high－risk areas accounted for 19.3% (grading value) and 20.4% (continuous value). Two indicators, the success rate curve, and landslide risk index histogram, were used to evaluate the results. The AUC value of grading data was 0.80, and the AUC value of continuous data was 0.76. The histogram comparison shows that the histogram characteristics of the graded values align more with expectations. This article compares multiple indicators and concludes that the classification factor is more suitable for SVM－based landslide risk assessment.

Keywords: data organization form; support vector machine; multi－collinearity analysis; landslide risk assessment; multi－index comparison