逻辑回归和人工神经网络模型在滑坡灾害空间预测中的应用

刘艺梁,殷坤龙,刘 斌
(中国地质大学(武汉)工程学院,武汉 430074)

摘要:以三峡坝区到巴东段为研究区,选择坡度、坡向、软弱夹层、水系影响范围和土地利用状况等 9 项评价指标,分别采用逻辑回归和人工神经网络(ANN)模型,在 ArcGIS 平台上进行滑坡灾害危险性预测。此外,应用受试者工作特征曲线(receiver operating characteristic curve, ROC 曲线)分析方法对两种模型的预测结果进行对比,分析结果表明滑坡危险性预测区划图和实际的滑坡发育情况基本吻合,逻辑回归模型和 ANN 模型的 ROC 曲线下面积 AUC 值分别为 0.806 和 0.799,两种模型的预测结果可以相互验证,且逻辑回归模型的预测精度相对较高。

关键词:三峡库区; Logistic 回归; 人工神经网络; GIS; ROC 曲线
中图分类号:P642.22 **文献标识码**:A **文章编号**:1000-3665(2010)05-0092-05

滑坡灾害是我国山区最重要的自然灾害之一,据国土资源部 290 个县市地质灾害调查结果显示,滑坡在地质灾害点中所占比例最大达 51%[1]。因此,加强滑坡危险性评估研究、进行滑坡灾害的危险性区划制图对于减灾防灾、提高滑坡灾害的预防能力具有重要的基础意义。

在过去的二十年里,经验模型、信息模型、统计预测模型、模式识别模型(专家系统、神经网络法)等在滑坡灾害评估研究得到了广泛的应用。刘传正、毕华兴[2~3]等将数量化理论与地理信息系统紧密结合,对区域滑坡进行了空间预测和危险等级划分。张桂荣[4]等采用半定量和定量两种方法对区域滑坡进行了预测,并通过两种方法的预测结果,对比分析了滑坡的形成和各影响因素的关系,进行了定性的描述分析。吴益平[5]等将信息量模型、信息 - 物元模型、信息 - 神经网络模型应用于滑坡危险性预测中,并对 3 种模型的预测结果进行了对比分析,指出了每种模型的优劣,但没有对模型的预测结果进行定量分析。

本文结合三峡坝区到巴东段的滑坡发生特征,从 ArcGIS 空间数据库中选取 9 项评价指标,采用逻辑回归模型和人工神经网络模型(ANN)对研究区的滑坡灾害进行空间预测,然后将这两种模型的预测值在 ArcGIS 中转为栅格文件,得到研究区的滑坡危险性预测区划图,最后运用 ROC 曲线分析方法对两种模型的预测结果进行了定量分析。

1 研究区背景

研究区横跨秭归和巴东两县。秭归县位于鄂西褶皱山地,地势西南高东北低,平均海拔 1000m 以上,相对高差一般在 500 ~ 1300m 之间,多年平均降水量 1493.2mm。巴东县位于湖北省西部,地处长江三峡中段西陵峡与神农溪口之间的过渡地带,山地高程 85 ~ 550m,属中低山峡谷区,多年平均降水量 1100.7mm。

秭归县地层发育齐全,自元古界至第四系均有出露,构造形迹主要以 NNE—SN、EW、NE 向断裂为主;巴东县出露地层主要有三叠系下统嘉陵江组(T_1j)、中统巴东组(T_2b)及第四系(Q),构造形迹以 EW 向的褶皱、断裂和 SN、NW、NE 向断裂为主。

研究区内秭归县共有 732 处地质灾害点或地质灾害隐患点,灾害类型以滑坡、崩塌为主;巴东县库岸变形破坏以滑坡为主(142 个),占库岸灾害总数(196 个)的 72.4%,体积占总方量的 93.2%。研究区沿江主要滑坡灾害分布见图 1。

2 数据

基于栅格数据应用逻辑回归模型和人工神经网络模型进行滑坡灾害空间预测分析时,将研究区划分为 70 908 个栅格,比例尺为 1∶5万,栅格像素选取的规格

收稿日期:2009-11-26; 修订日期:2010-02-01
基金项目:国家自然科学基金资助项目(40872176)
作者简介:刘艺梁(1985-),男,硕士研究生,主要从事地质灾害预测预报方面的研究。
E-mail:liuyiliangcug@gmail.com

为 100m × 100m。根据长江三峡坝区到巴东段滑坡发生的实际情况，考虑的预测因素有坡度、坡向、山脊山谷、岩性、坡体结构、软弱夹层、断层影响范围、水系影响范围、土地利用，这些预测因素是通过 ArcGIS9.3 中的 ARC/INFO 进行空间数据库的计算或提取完成的。栅格数据的其它来源主要有 DEM 数据和矢量数据。从 DEM 数据可以转化取得如坡度、坡向、高程、山影等地形地貌数据。而根据其作用已经分区的矢量数据，如岩性、岩体结构、植被、土地利用等，可以直接根据属性转化为栅格数据；其它矢量数据例如断层、褶皱、水系等，需要 Buffer（缓冲）处理等空间处理后才能确定其影响范围，然后才能转为栅格数据。

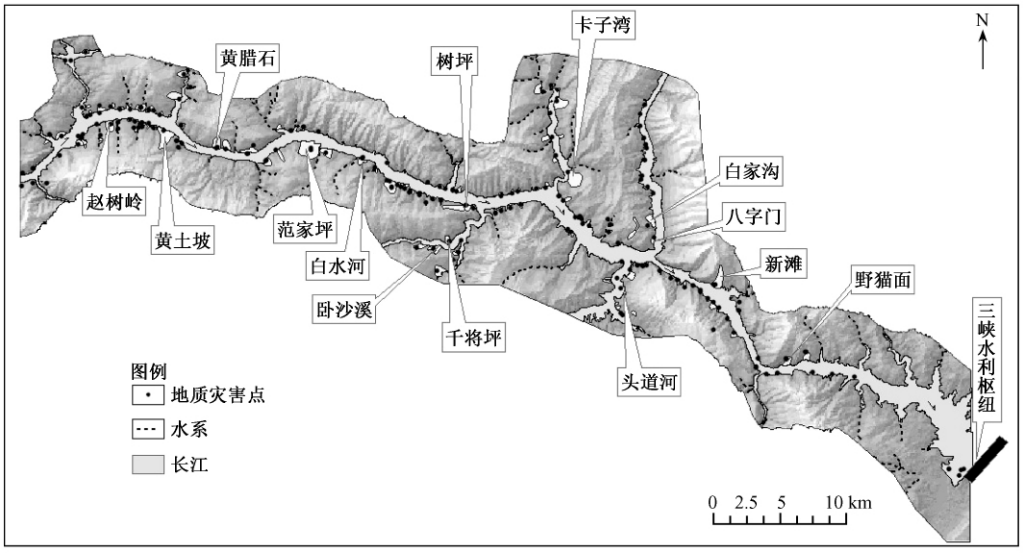

图 1　研究区沿江主要滑坡灾害分布图

Fig. 1　Study area's main landslide location map

3　逻辑回归分析模型

逻辑回归分析主要是在一个因变量和多个自变量之间形成多元回归关系，从而预测任何一块区域某一事件的发生概率。逻辑回归的优势在于进行统计分析时，自变量可以是连续的，也可以是离散的，也没有必要满足正态分布。而一般的多元统计分析模型中，变量必须满足正态分布。

在逻辑回归分析中，因变量 Y 是一个二分类变量，其取值 $Y=1$ 和 $Y=0$，分别代表发生过滑坡和未发生滑坡。影响 Y 取值的 n 个自变量分别为 $X_1, X_2, \cdots, X_n$，在 n 个自变量作用下滑坡发生的条件概率为 $P=P(Y=1 \mid X_1, X_2, \cdots, X_n)$，则 logistic 回归模型可表示为：

$$z_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_n X_{in} \quad (1)$$

$$P_i = \frac{1}{1 + \exp(-z_i)} \quad (2)$$

式中：z_i——中间变量参数；

a_0——回归常数；

a_i——第 j 个变量的回归系数（$i, j = 1, 2, \cdots, n$）；

X_{ij}——第 i 号单元中第 j 个变量的取值，存在滑坡取 1，否则取 0；

P_i——第 i 号单元内滑坡发生概率的回归预测值（$i = 1, 2, \cdots, n$）。

逻辑回归分析最重要的是将变量转换为二进制数据。本文对于连续型变量的数据（坡度、坡向、坡体结构、断层影响范围、水系影响范围），从 ArcGIS 中读取相应的数据，绘制连续型变量的频率分布直方图，然后根据直方图的频率分布，选取几个突变点将连续型变量分为几个区间，如果滑坡存在于这个区间，就取值 1，否则取 0。而离散型的分类数据（山谷山脊、岩性、软弱夹层、土地利用）则根据分类数据以二进制表示，即存在取 1，否则取 0。逻辑回归分析中考虑 26 个连续变量和 15 个离散变量，总共 41 个独立的变量，因素状态见表 1。然后在 SPSS15.0 中计算预测因素不同状态的回归系数，接着将逻辑回归方程代入 Matlab7.0 编制的程序，得到整个研究区滑坡发生的概率值，最后将每个栅格单元的概率值在 ArcGIS 中转化为栅格文件，根据自然断点法，生成滑坡危险性预测区划图（图 2）。由图 2 可知，滑坡灾害危险性预测区划图和实际的滑坡灾害分布图基本一致。

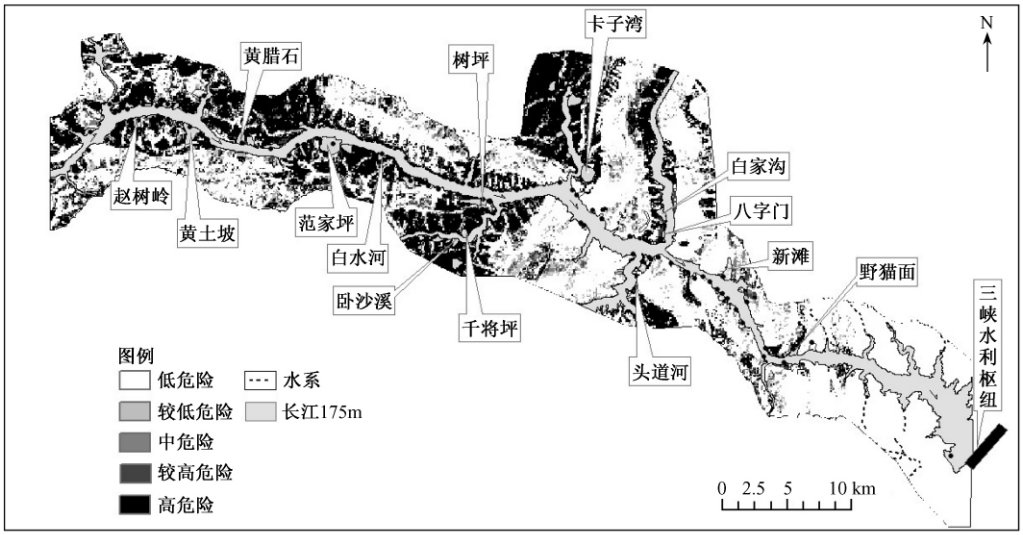

图 2 逻辑回归模型滑坡危险性预测区划图

Fig. 2 Landslide hazard zoning map produced with the logistic regression model

表 1 因素状态表

Table 1 Factor state

因素	状态	因素	状态
坡度(°)	0 ~ 6	坡体结构(°)	0 ~ 45
	6 ~ 25		45 ~ 120
	25 ~ 40		120 ~ 160
	40 ~ 90		160 ~ 180
坡向(°)	0 ~ 35	软弱夹层	有
	35 ~ 90		无
	90 ~ 160	断层影响范围(m)	< 500
	160 ~ 234		500 ~ 1000
	234 ~ 306		1000 ~ 1500
	306 ~ 342		1500 ~ 2000
	342 ~ 360		> 2000
山谷山脊	山谷	水系影响范围(m)	< 300
	山脊		300 ~ 600
	其它		600 ~ 900
岩性	花岗岩,闪长岩		900 ~ 1200
	石灰岩,灰岩		> 1200
	泥质灰岩,泥灰岩	土地利用	树木
	页岩		灌木
	砂岩夹煤层		小草
	黄土		建筑区

4 人工神经网络分析模型

人工神经网络模型是在对人类大脑神经系统及功能的模仿的基础上建立的推理模型,具有很强的非线性映射能力,其中应用最广泛的是误差反向传播网络即 BP 神经网络。

传统的 BP 算法在训练过程速度比较缓慢,而且容易陷入局部最小值。为了提高训练速率和神经网络模型的稳定性,许多学者都认为方法的改进关键在于动量系数和学习率的选取[6~7]。因此,动量系数和学习率的选取在 BP 神经网络算法的改进上是非常重要的。根据训练调试,文中动量系数取 0.9,学习率取 0.05。

神经网络结构确定的关键是隐含层层数和隐含层神经元个数。实践表明,隐含层数目的增加可以提高 BP 网络的非线性映射能力,但是隐含层超过一定值,网络性能反而会降低。而单隐层的 BP 网络可以逼近一个任意的连续非线性函数。因此,这里采用单隐层的 BP 网络,隐含层的神经元个数直接影响着网络的非线性能力。考虑到输入层和输出层的数目以及训练数据的数目,根据科尔莫哥洛夫(Kolmogorov)定理,采取"9 – 19 – 1"的神经网络结构[8]。

最后,采用 Matlab7.0 来训练、测试神经网络模型,训练函数选用 traingdm,传递函数选用 tansig[9]。BP 神经网络的性能函数采用 mse,以保证达到最小的均方误差值(MSE)。本文的神经网络模型中的训练和测试数据最后的 MSE 值分别为 0.00208 和 0.0217.待神经网络的训练目标达到后,将整个研究区的数据代入神经网络模型来评价研究区的稳定性。然后将得到的每个单元的预测值在 ArcGIS 中转为栅格文件,根据自然断点法,生成滑坡危险性预测区划图(图 3)。由图 3 可知,滑坡灾害危险性预测区划图和实际的滑坡灾害分布图基本一致。

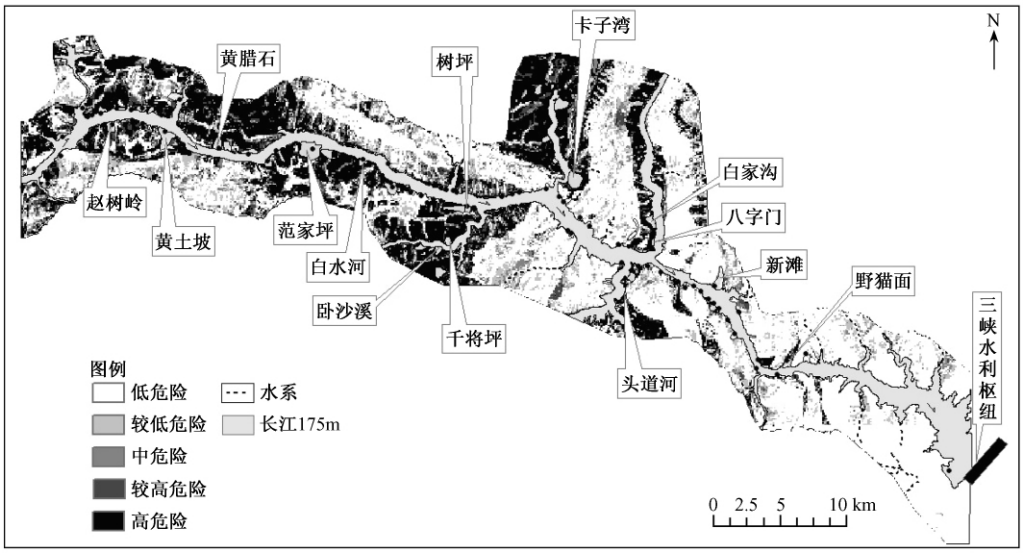

图 3　人工神经网络模型滑坡危险性预测区划图

Fig. 3　Landslide hazard zoning map produced with the artificial neural network model

5　ROC 曲线分析

受试者工作特征曲线(receiver operating characteristic curve，ROC 曲线)分析方法最初是应用于雷达信号接收能力的评价，后广泛应用于医学诊断试验性能的评价[10]。

ROC 曲线是以预测结果的每一个值作为可能的判断阈值，由此计算得到相应的灵敏度和特异度，以假阳性率即(1 - 特异度)为横坐标，以真阳性率即灵敏度为纵坐标绘制而成。ROC 曲线下的面积即为 AUC 值(area under curve)。AUC 是很好的衡量模型预测准确度的指标，其取值范围为 [0.5，1]，值越大表示模型判断力越强。理想情况是模型预测分布区与滑坡实际分布区完全吻合，此时 AUC 值为 1。文中假阳性率即未发生滑坡单元被正确预测的比例，真阳性率即滑坡单元被正确预测的比例。

将两个模型最后的预测值和相应的诊断值导入到 SPSS15.0 中进行 ROC 分析，得到两个模型的 ROC 曲线和 AUC 值(图 4)。其中逻辑回归模型和人工神经网络模型的 AUC 值分别为 0.806 和 0.799，说明这两个模型的预测结果可以相互验证。但逻辑回归模型在较高危险和高危险区域的预测结果比 ANN 模型要准确，分析主要有以下三方面原因：(1)初始权重是随机分配的；(2)变量大部分是根据经验选择，而且数据不够完备，有些数据只能以区间来分类；(3)在学习训练的过程中，执行时间过长，造成数据的过度拟合。这是需要进一步研究和改进的。

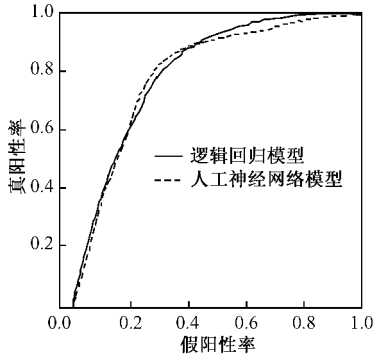

图 4　逻辑回归和人工神经网络模型的 ROC 曲线

Fig. 4　ROC curve of the logistic regression model and the artificial neural network model

6　结论及建议

(1)以三峡坝区到巴东段为研究区，选择坡度、坡向、软弱夹层、水系影响范围和土地利用状况等 9 项评价指标，分别采用逻辑回归和人工神经网络模型，在 ArcGIS 平台上进行滑坡灾害危险性预测，结果表明滑坡灾害危险性预测区划图和实际的滑坡灾害分布图基本一致。

(2)运用 ROC 曲线分析法对这两种模型进行评价，结果表明这两个模型的预测结果可以相互验证，逻辑回归模型在较高危险和高危险区域的预测结果比 ANN 模型要准确。但这并不能说明逻辑回归模型在其它地质环境下的预测效果也比 ANN 模型好，还有待在其他实例中进行验证。

(3)在对滑坡灾害进行空间预测的过程中,空间数据的管理和输入数据的修改很不方便。因此,为了提高预测模型的运行效率,及时修改相关输入数据,加强空间数据的分析功能,将 GIS 和空间预测模型集成对 GIS 进行二次开发很有必要。

致谢:文中滑坡各项调查及地质数据均来源于三峡库区地质灾害防治工作指挥部,在此表示衷心的感谢!

参考文献:

[1] 李媛,孟晖,董颖,胡树娥. 中国地质灾害类型及其特征:基于全国县市地质灾害调查成果分析[J]. 中国地质灾害与防治学报,2004,15(2):29 – 34.

[2] 刘传正,李铁锋,温铭生,等. 三峡库区地质灾害空间评价预警研究[J]. 水文地质工程地质,2004,31(4):9 – 19.

[3] 毕华兴,中北理,阿部和时. GIS 支持下的滑坡空间预测与危险等级划分[J]. 自然灾害学报,2004,13(3):50 – 57.

[4] 张桂荣,殷坤龙. 区域滑坡空间预测方法研究及结果分析[J]. 岩石力学与工程学报,2005,24(23):4297 – 4302.

[5] 吴益平,殷坤龙,陈丽霞. 滑坡空间预测数学模型的对比及其应用[J]. 地质科技情报,2007,26(6):95 – 100.

[6] 范磊,张运陶,程正军. 基于 MATLAB 的改进 BP 神经网络及其应用[J]. 西华师范大学学报:自然科学版,2005,26(1):70 – 74.

[7] 陈桦,程云艳. BP 神经网络算法的改进及在 Matlab 中的实现[J]. 陕西科技大学学报,2004,22(2):45 – 47.

[8] 葛哲学,孙志强. 神经网络理论与 MATLAB R2007 实现[M]. 北京:电子工业出版社,2007.

[9] 夏金梧,李长安,王旭. 基于神经网络理论的三峡水库诱发地震预测研究[J]. 水文地质工程地质,2007,34(5): 17 – 20.

[10] 王运生,谢丙炎,万方浩,等. ROC 曲线分析在评价入侵 物种分布模型中的应用[J]. 生物多样性,2007,15(4):365 – 372.

Application of logistic regression and artificial neural networks in spatial assessment of landslide hazards

LIU Yi-liang, YIN Kun-long, LIU Bin

(*Engineering Faculty*, *China University of Geosciences*, *Wuhan* 430074, *China*)

Abstract: The main purpose of this study is to analyze and compare the results of spatial assessment of landslide hazards by means of logistic regression analysis, back – propagation artificial neural network (ANN) through the use of geographic information system (GIS) techniques. A region from the Three Gorges Dam to the Badong Town, which suffered from deformation for many years, was selected as the application site of this study. Firstly, nine evaluation factors were selected, including topographic slope, topographic aspect, soft interlayer, water system influence area, land use, and so on. Then, this paper highlighted the discrepancies between a logistic regression model and an ANN model, made a landslide susceptibility map, and evaluated the performance of these two models by ROC curve analysis. Finally, based on the above analyses, satisfactory agreement was obtained between the susceptibility map and the existing data on landslide location. The area under curve (AUC) values of the logistic regression and the ANN model are 0.806 and 0.799. Accuracies of these two models can be evaluated relatively similarly.

Key words: Three Gorges Dam; logistic regression; artificial neural network; GIS; ROC curve

责任编辑:汪美华