赵晓萌，卫星君，王娜，等. 降雨型滑坡灾害的特征聚合决策树预测模型[J]. 灾害学，2020，35(1)：27－31. [ZHAO Xiaomeng，WEI Xingjun WANG Na，et al. Feature aggregation decision tree prediction model for rainfall landslide disaster[J]. Journal of Catastrophology，2020，35(1)：27－31. doi：10.3969/j.issn.1000－811X.2020.01.006.]

降雨型滑坡灾害的特征聚合决策树预测模型*

赵晓萌[1]，卫星君[2]，王 娜[1]，雷向杰[1]

(1. 陕西省气候中心，陕西 西安 710014；2. 陕西能源职业技术学院，陕西 咸阳 712000)

摘 要：为了有效预警降雨型滑坡灾害，以秦巴山区为研究区域，采集并处理大量不同时段降雨数据，构成降雨特征属性。利用 Fisher 最优分割，对降雨特征属性值进行分段统计，提出特征聚合转换表。使用信息增益和预测反馈筛选影响滑坡灾害的有效降雨特征，为预测模型提供有效数据集。利用特征聚合转换表和有效降雨特征，改进决策树，构建特征聚合决策树预测模型，进而提高预测效率和预测准确率。分析决策树的深度和叶子节点个数，给出决策树的反馈执行度，表明使用特征聚合转换表的决策树更优。比较特征聚合决策树、决策树、朴素贝叶斯和逻辑回归预测模型，结果表明，特征聚合决策树预测模型对降雨型滑坡灾害有更高的预测准确率，且平均预测准确率较高。

关键词：降雨；滑坡灾害；信息增益；最优分割；决策树

中图分类号：X43；X915.5；P642 **文献标志码**：A **文章编号**：1000－811X(2020)01－0027－05

doi：10.3969/j.issn.1000－811X.2020.01.006

滑坡灾害现已成为我国首要的地质灾害，尤其以降雨滑坡最为严重，影响着生产经济的发展且造成了巨大生命和经济损失[1-2]。以秦巴山区为例，2010－2012 年期间发生的滑坡灾害中，90%以上由降雨所致[3]。由于不同区域降雨对滑坡的影响不同，且监测降雨数据量大，给滑坡预测带来困难。因此，研究降雨型滑坡成为预防区域滑坡灾害的重点和热点。

费晓燕[4]等对滑坡灾害的前期雨量进行统计分析，采用逻辑回归构建灾害预警模型，但模型整体准确率仅为 78.36%。王芳[4]等利用不同算法对滑坡灾害进行预测，但没有综合考虑不同日降雨对滑坡灾害的影响。曹洪洋[5]等把降雨型滑坡数据和粗糙集理论结合预测滑坡是否发生，但缺少连续小时降雨对滑坡灾害的影响。李芳等[6]使用信息量法构建预警模型，提高预警水平，但没有分析不同降雨因子对滑坡的影响程度。Caraccio-lo[7]等依据滑坡和降雨数据推导临界降雨阈值，进行灾害预测，但未能全面分析不同时段连续降雨对滑坡的影响。Rossi[8]等利用统计方法对滑坡预警，但依赖于降雨经验阈值。降雨数据随时间不断累积而增加，且当前大多数的滑坡灾害预测过度依赖降雨经验阈值，使得提高预测效率和预测精度成为难点。因此，机器学习[9]作为数据分析和预测的有效手段，显的越发重要。如决策树、逻辑回归、朴素贝叶斯、支持向量机等模型对于不同实际场景应用广泛。

本文针对以上问题提出了特征聚合决策树。在综合考虑降雨对滑坡的影响下，筛选影响滑坡灾害的有效降雨因子，提高预测效率；提出特征聚合转换表，进而改进决策树，提高预测准确率；给出决策树的反馈执行度，表明该模型有效性。

1 数据采集与处理

地质灾害灾情数据来源于陕西省地质环境监测总站对 2010－2012 年秦巴山区的监测，主要包括灾害发生时间、地点、规模和经纬度。滑坡灾害点附近降雨数据，由不同的区域气象站(以下称区域站)对降雨数据进行逐时采集。距离灾害点最近的区域站采集的降雨数据，更能准确反映降雨对滑坡灾害的影响程度。

1.1 降雨特征属性

滑坡当日降雨和前期降雨均有可能诱发滑坡灾害[10-11]。选取当日最大连续 1 h(RH^1)、3 h(RH^3)、6 h(RH^6)、12 h(RH^{12})、当日(RH^{24})雨量和，滑坡

* 收稿日期：2019－05－29 修回日期：2019－07－12

基金项目：国家重点基础研究发展计划(973)(2013CB430202)；中国气象局气候变化专项项目(CCSF201845)；陕西省气象局秦岭和黄土高原生态环境气象重点实验室青年基金课题(2019Y－7)

第一作者简介：赵晓萌(1985－)，女，陕西咸阳人，工程师，主要研究方向为气候监测与灾害估. E-mail：xmzhao2011@163.com

通讯作者：卫星君(1983－)，男，甘肃平凉人，副教授，主要从事机器学习方面的研究. E-mail：shanxi_edu@126.com

发生前 1 d(RD^1) 至前 10 d(RD^{10}) 日综合雨量和，共计 15 个作为影响滑坡灾害的特征属性[12-13]。

滑坡灾害点最近区域站表示如下:

$$I = \min_{p \in P}\left(\sqrt{(C_x - S_x)^2 + (C_y - S_y)^2}\right)。 \quad (1)$$

式中: C_x 和 C_y 为滑坡灾害点经度和纬度; S_x 和 S_y 为区域站经度和纬度; P 为滑坡灾害点附近全部区域站。

计算灾害点附近全部区域站距离，获取最近区域站 I，进而计算日小时降雨量 RH_I 和前期日综合雨量 RD_I。

日小时降雨表示如下:

$$RH_I = \begin{cases} \max\left(\sum_{i}^{i+h-1-24} RH^h\right), i + h - 1 > 24; \\ \max\left(\sum_{i}^{i+h-1} RH^h\right), i + h - 1 \leqslant 24。 \end{cases} \quad (2)$$

式中: i 为发生降雨的时间; h 为滑坡发生前期小时; RH^h 为连续 h 小时雨量。

日综合雨量表示如下:

$$RD_I = RD^t + \sum_{d=1}^{10} \theta^d RD^d 。 \quad (3)$$

式中: RD^t 滑坡发生当日雨量; θ^d 降雨衰减系数; RD^d 为前 d 天综合雨量。

1.2 特征属性的信息增益

特征属性过多，分析特征和训练模型所需的时间就长，同时容易引起“维度灾难”，模型也越复杂; 特征属性过少模型太简单，影响预测准确率。因此，使用信息增益[14] (IG) 对降雨特征属性进行排序，得到 15 个特征属性对滑坡灾害的影响程度。

信息增益表示一个特征能够为分类系统带来多少信息，带来的信息越多，说明该特征越重要，信息增益也就越大。信息增益表示如下:

$$IG(S, Attr) = H(S) - C(S^v)$$
$$= -\sum_{i=1}^{n} p_k \log_2 p_k - \sum_{v=1}^{V} c^v / c H(S^v)。 \quad (4)$$

式中: S 为特征属性集; $H(S)$ 为特征属性的信息熵; $C(S^v)$ 为特征属性不同取值时，赋予权重 c^v/c 的信息熵; v 是特征属性 $Attr$ 可能取值。

2 滑坡灾害特征属性筛选

直接使用 15 个降雨滑坡灾害特征，模型预测效率和准确率不高。本文利用 Fisher 最优分割[15]，对每一个降雨特征属性值进行分段统计，提出特征聚合转换表，再利用预测反馈筛选影响滑坡灾害的有效特征属性，进而提高预测效率和准确率。

2.1 Fisher 最优分割

最优分割将特征属性值近似的样本划分到同一段内。因此，用{ V_1^{attr}, V_2^{attr}, ……, V_m^{attr}} 表示 m 个样本的 $attr$ 特征属性值。特征属性段内直径，即属性 $attr$ 的第 i 个特征值到 j 个特征值距离均值的距离，表示为:

$$J(i, j) = \sum_{t=i}^{j} (V_t^{attr} - \bar{V}^{attr})^2 。 \quad (5)$$

式中，$\bar{V}^{attr}$ 为 i 到 j 特征均值。表示为:

$$\bar{V}^{attr} = \frac{1}{j - i + 1} \sum_{t=i}^{j} V_t^{attr}。$$

n 个不同特征属性值分成 k 段，定义分段损失函数为:

$$W(b(n, k)) = \sum_{t=1}^{k} D(i_t, i_{t+1} - 1)。 \quad (6)$$

损失函数达到极小，k 段内距离最小，则损失函数最优解递推公式表示为:

$$W(b(n, 2)) = \min_{2 \leqslant j \leqslant n} \{D(1, j-1) + D(j, n)\}; \quad (7)$$

$$W(b(n, k)) = \min_{\leqslant j \leqslant n} \{W(b(j-1, k-1)) + D(j, n)\}。 \quad (8)$$

即先求解最优 2 分割，依次计算直到最优 k 分割。k 的取值依赖于段内距离 D，且随着 k 变化，D 趋于稳定。

2.2 聚合转换表

定义 1 特征聚合转换表，{ $attr_1$, $attr_2$, …… $attr_n$} 为样本的 n 个特征属性，$attr_i$ 特征属性的 m 个特征值的最优 k 个分割中，段内特征值表示相同，即: $V^k = \{v_1^k, v_2^k, \cdots v_q^k\}$, $v_i^k = v_j^k$, $v_i^s \neq v_j^q$, 则 n 个特征属性构成的表称为特征聚合转换表，记作 ATB。

由定义可知，特征属性最优 k 分割，同一段内特征属性有相同的“等级”表示，即特征属性值近似的样本，段内特征值表示相同。

特征聚合转换表的意义在于:

(1) 解决特征属性值的多样性，提高预测效率。

(2) 聚合特征值，提高特征属性相似样本的预测准确率。

特征聚会转换表的计算具体步骤如下:

Step1 对 n 个特征属性 $attr$ 的 m 个值排序。

Step2 依据公式(5) 和公式(8) 计算第 i 个特征属性的 k 段最优分割。

Step3 属性 $attr$ 中的特征值{ V_1^{attr}, V_2^{attr}, …, $V_{i\cdots}^{attr}$, V_m^{attr}} 段内表示相同。

Step4 重复 Step2 和 Step3，直到 $i = n$。

2.3 特征属性筛选

计算降雨特征属性信息增益并排序，对排序的特征属性构建聚合转换表。逐次插入降雨特征属性，构建决策树并对滑坡灾害进行预测。选择预测准确大于阈值 δ，且特征属性最少的为影响降雨滑坡灾害有效特征属性。

3 特征聚合决策树预测模型实现

决策树[16] 是样本进行分类的树形结构，各个结点用信息增益选择特征，递归地构建一颗决策树。

本文改进后的决策树算法:

(1) 使用信息增益和预测结果反馈确定影响滑坡灾害的有效降雨特征。

(2) 使用特征聚合转换表，解决了特征属性值的多样性；减少了无意义的树分枝，合并了分类能力强的属性树分枝，有效避免了过度拟合的问题。

(3) 其他算法模型与决策树模型相同。

3.1　算法流程

算法流程如图 1 所示。

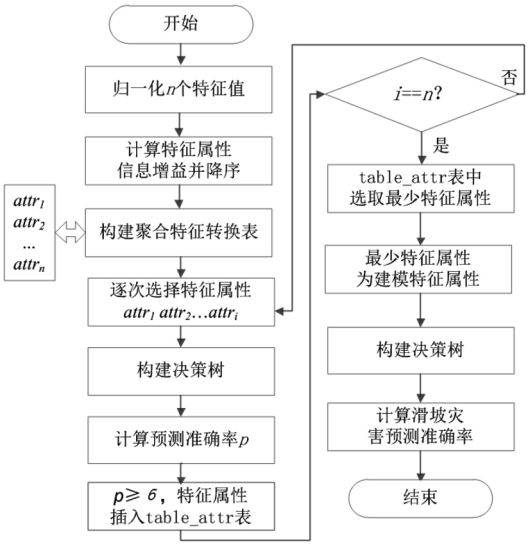

图 1　改进决策树算法流程图

算法具体步骤如下:

Step1 使用式(1)、式(2)式计算灾害样本的 15 个降雨特征属性，并归一化。

Step2 对样本的 15 个降雨特征属性，使用公式(4)计算信息增益大小，并降序。

Step3 利用上文 2.2 计算特征属性聚合转换表。

Step4 构建决策树，并逐次加入特征属性，计算预测准确率 p。

Step5 $p \geqslant \theta$，把特征属性插入 table_ attr 表中。

Step6 重复 Sep4 和 Step5，直到 $i = n$。

Step7 table_ attr 表中特征个数最少的为影响滑坡灾害有效降雨特征。

Step8 使用筛选的特征属性，构建决策树预测模型，对滑坡灾害进行预测。

3.2　反馈执行度

针对改进的决策树，本文提出决策树的反馈执行度，反映决策树的复杂程度。表示为:

$$FD_{eff} = -\frac{\sum_{i=1}^{I}(|T^i| + depth^i)}{I}, (if \quad p^i \geqslant \sigma) \quad 。 \qquad (9)$$

式中: I 为构建决策数的次数; $|T^i|$ 决策树的叶子节点个数; $depth^i$ 为决策树的深度; σ 为预测准确率阈值; p^i 表示构建第 i 次决策树时，大于等于 σ 的准确率。p^i 的表示如下:

$$P^i = \frac{TP^i}{TP^i + FP^i} 。 \qquad (10)$$

式中: TP 正类判断为正类; FP 反类判断为正类。从表达式(9)可以看出，当 $p^i \geqslant \theta$ 时，反馈执行度越大，构建次数为常数，则决策树的叶子节点个数和树的深度就越小，即低复杂度和高预测准确率。因此，可以使用 FD_{eff} 来描述决策树模型的预测效果。

4　实验设计与分析

4.1　实验设计

实验数据为 2010 – 2012 年秦巴山区 845 个地质灾害数据，计算日小时降雨和前期日综合雨量，作为降雨滑坡灾害特征属性，构成实验样本数据集。

4.1.1　筛选降雨滑坡灾害特征

降雨特征属性的信息增益大小比较如图 2。

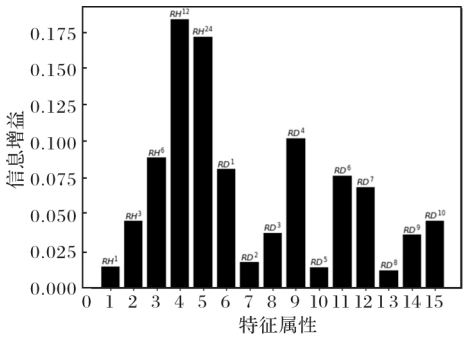

图 2　特征属性信息增益比较

表 1 为信息增益降序后，随机选取 75% 数据为训练集，25% 数据为测试集，进行 100 次实验，逐次加入特征属性的预测准确率。可以看出平均准确率 $p \geqslant 0.82$，特征属性最少个数为 5。因此，筛选 RH^6，RH^{12}，RH^{24}，RD^1，RD^4 为影响降雨滑坡灾害的有效特征属性。计算这 5 个特征属性的聚合转换表，构建降雨型滑坡灾害特征聚合决策树模型。

4.1.2　确定预测反馈准确率阈值

图 3 为以筛选特征属性，逐次构建深度为 1 ~ 200 的决策树的预测准确率。可以看出，当树的深度达到一定值后，准确率维持在 80% 以上，而深度的增加对准确率影响不大。

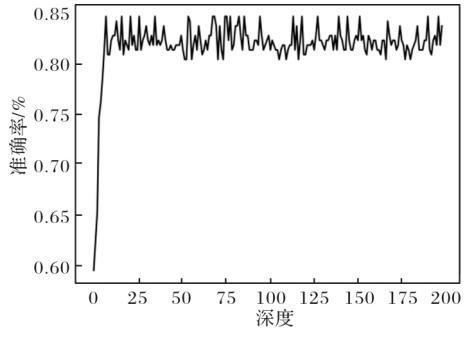

图 3　深度为 1 ~ 200 预测准确率

表 1 筛选特征属性的预测准确率

筛选特征属性															准确率/%
RH^{12}															57.6
RH^{12}	RH^{24}														73.7
RH^{12}	RH^{24}	RD^{4}													76.9
RH^{12}	RH^{24}	RD^{4}	RH^{6}												80.6
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}											84.3
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}										82.1
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}									79.1
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}	RD^{10}								82.1
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}	RD^{10}	RH^{3}							83.1
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}	RD^{10}	RH^{3}	RD^{3}						75.0
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}	RD^{10}	RH^{3}	RD^{3}	RD^{9}					73.1
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}	RD^{10}	RH^{3}	RD^{3}	RD^{9}	RD^{2}				73.6
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}	RD^{10}	RH^{3}	RD^{3}	RD^{9}	RD^{2}	RH^{1}			73.6
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}	RD^{10}	RH^{3}	RD^{3}	RD^{9}	RD^{2}	RH^{1}	RD^{5}		73.6
RH^{12}	RH^{24}	RD^{4}	RH^{6}	RD^{1}	RD^{6}	RD^{7}	RD^{10}	RH^{3}	RD^{3}	RD^{9}	RD^{2}	RH^{1}	RD^{5}	RD^{8}	77.4

图 4 为深度 1 到 20 的决策树的预测准确率，可以看到，当深度≥7，预测准确率超过 82%，而深度增加，准确率增幅范围在 0 ~ 0.05 之间，因此实验中选取决策树深度 $depth \geqslant 7$。

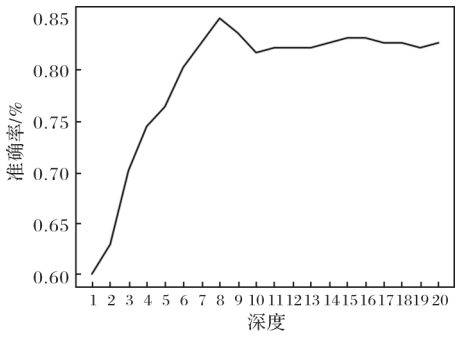

图 4 深度为 1 ~ 20 预测准确率

由 4.1.1 和 4.1.2 确定实验参数：筛选特征属性个数 $m = 5$（RH^{6}，RH^{12}，RH^{24}，RD^{1}，RD^{4}）；降雨衰减系数 $\theta_d = 0.8$[10,17]；预测结果阈值 $\sigma = 0.82$；决策树深度 $depth = 8$ 和叶子节点 $Leaf\_nodes = 7$。

4.2 实验结果比较

表 2 为 100 次实验，使用特征聚合表（ATB）和未使用情况下（NATB），平均叶子节点个数，平均树的深度，平均准确率和决策树的反馈执行度。可以看出使用决策转换表后，有更高的反馈执行度和预测准确率。

表 2 ATB 和 NATB 决策树比较

决策树	*Leaf_nodes*	*depth*	*ave_acc*/%	*FDeff*
ATB	6.6	8.2	82.1	−14.8
NATB	6.8	8.6	78.6	−15.4

图 5 为随机抽取滑坡灾害样本 100，200，300，400，500，600，700，800，比较特征聚合决策树、决策树、朴素贝叶斯和逻辑回归准确率。从图 5 可知，预测模型的准确率，随着滑坡灾害样本的增加而提高。滑坡灾害样本超过 250 个后，特征聚合决策树的预测准确率高于其他三个模型，说明特征聚合决策树模型能更好的预警降雨型滑坡灾害的发生。

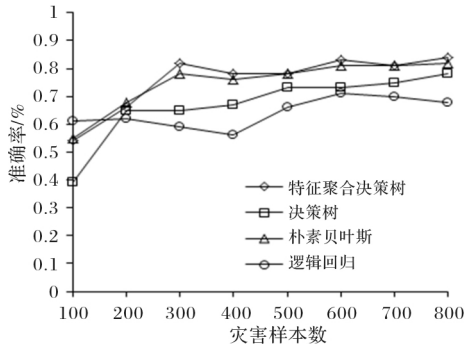

图 5 模型准确率比较

5 结论与讨论

本文收集处理大量降雨数据，计算最近区域站不同时段降雨量，利用最优分割、信息增益和预测反馈筛选影响滑坡灾害的有效降雨因子，建立特征聚合决策树预测模型，对降雨型滑坡灾害进行了预测，结果表明：

（1）利用最优分割，提出特征聚合转换表，解决了特征值多样性的问题，提高了预测效率和准确率。

（2）给出决策树的反馈执行度，计算使用和未使用特征聚合转换表的决策树反馈执行度，表明特征聚合决策树预测模型更优。

（3）相比决策树、朴素贝叶斯和逻辑回归有更

高的预测准确率。

下一步研究的重点是，考虑影响滑坡灾害的其他诱因，如土体、岩体、地下水和人工活动等，进一步分析影响降雨型滑坡灾害的关键因素，为此类滑坡灾害预防预警提供参考。

参考文献:

[1] 文海家，张岩岩，付红梅，等. 降雨型滑坡失稳机理及稳定性评价方法研究进展[J]. 中国公路学报，2018，31(2)：15－29.

[2] 王卫东，瞿霞，刘攀，等. 基于最优权重联合模型的滑坡位移预测研究[J]. 灾害学，2018，33(1)：59－64.

[3] 费晓燕，柳锦宝，屈伯强，等. 四川省降雨诱发滑坡灾害的气象预警模型[J]. 水土保持通报，2017，37(5)：315－321.

[4] 王芳，殷坤龙，桂蕾，等. 不同日降雨工况下万州区滑坡灾害危险性分析[J]. 地质科技情报，2018，37(1)：190－195.

[5] 曹洪洋，任晓莹. 基于粗糙集理论的区域降雨型滑坡预测预报[J]. 水文地质工程地质，2017，44(2)：117－123.

[6] 李芳，梅红波，王伟森，等. 降雨诱发的地质灾害气象风险预警模型：以云南省红河州监测示范区为例[J]. 地球科学，2017，42(9)：1637－1646.

[7] Caracciolo D，Arnone E，Conti F L，et al. Exploiting historical rainfall and landslide data in a spatial database for the derivation of critical rainfall thresholds [J]. Environmental Earth Sciences，2017，76(5)：222.

[8] Rossi M，Luciani S，Valigi D，et al. Statistical approaches for the definition of landslide rainfall thresholds and their uncertainty using rain gauge and satellite data [J]. Geomorphology，2017，285：16－27.

[9] 杨剑锋，乔佩蕊，李永梅，等. 机器学习分类问题及算法研究综述[J]. 统计与决策，2019(6)：36－40.

[10] 魏来. 降雨诱发滑坡预测模型研究[D]. 重庆：重庆交通大学，2013.

[11] 赵奎锋. 诱发陕西秦巴山区地质灾害的强降水形成机制及预报预警研究[D]. 兰州：兰州大学，2012.

[12] 卫星君，赵晓萌，马长玲，等. 降雨型滑坡灾害的约简和逻辑回归预测模型[J]. 中国安全科学学报，2018(8)：1－6.

[13] 赵晓萌，蔡新玲，雷向杰，等. 基于 Logistic 回归的陕南秦巴山区降雨型滑坡预测方法[J]. 冰川冻土，2019，41(1)：175－182.

[14] 毛伊敏，彭喆，陈志刚，等. 基于不确定决策树分类算法在滑坡危险性预测的应用[J]. 计算机应用研究，2014，31(12)：3646－3650.

[15] 朱燕燕，武鹏林. 基于 FAHP－Fisher 的最优分割法在汛期分期中的应用[J]. 水电能源科学，2016，34(6)：57－59.

[16] 赵建民，黄珊，王梅，等. 改进的 C4.5 算法的研究与应用[J]. 计算机与数字工程，2019，47(2)：261－265.

[17] 唐红梅，魏来，唐云辉，等. 重庆地区降雨型滑坡相关性分析及预报模型[J]. 中国地质灾害与防治学报，2013，24(4)：16－22.

Feature Aggregation Decision Tree Prediction Model for Rainfall Landslide Disaster

ZHAO Xiaomeng[1]，WEI Xingjun[2]，WANG Na[1] and LEI Xiangjie[1]

(1. *Shaanxi Provincial Climate Center*，*Xi'an* 710014，*China*；2. *School of Electrical and Information Engineering*，*Shaanxi Energy Institute*，*Xianyang* 712000，*China*)

Abstract: In order to forewarn the rain-type landslide disaster effectively，the Qinba Mountain area is used as the research area. In the research，a large number of rainfall data are collected and processed to form rainfall feature attributes. Fisher optimal segmentation is used to propose a feature aggregation conversion table by segmentation statistics of rainfall feature attribute values. Effective rainfall characteristics of landslide hazards are screened by information gain and predictive feedback to provide valid data sets for predictive model. Feature aggregation conversion table and effective rainfall characteristics are used to improve the decision tree. Meanwhile，the prediction model of the feature aggregation decision tree is constructed to achieve high prediction efficiency and accuracy. Having analyzed the depth of the decision tree and the node number of the leaves，the feedback execution of the decision tree shows that the decision tree of the feature aggregation conversion table is better. Feature aggregation decision trees，decision trees，naive Bayesian prediction models and logistic regression are compared. The results show that the prediction model the feature aggregation decision tree has higher prediction accuracy for rainfall landslide disasters，and the average prediction accuracy is higher.

Key words: rainfall；landslide hazard；information gain；optimal segmentation；decision tree