基于 C4.5 算法的洪涝灾害应急响应级别分析①

徐 国[1,2]，乌 云[1]，王儒敬[1,2]，张 洁[1,2]

[1](中国科学院 合肥智能机械研究所，合肥 230031)
[2](中国科学技术大学 自动化系，合肥 230039)

摘 要：自然灾害突发时，有效准确地启动应急响应级别是科学合理地实施应急救援的前提条件。目前，我国的应急响应分级标准是依据民政部制定的三个应急响应指标而确立的。它规定了用于划分响应级别的各指标的取值范围，但未给出如何根据实地灾情进行弹性调整的量化尺度，即酌减比率。针对上述问题本文提出利用历史案例数据库，在 C4.5 算法的基础上训练出应急响应级别分类器，得出具有实际灾情信息的分级准则。为传统的应急响应级别提供一个合理的酌减比率。通过对洪涝灾害进行实验表明该方法能够提供非常有参考价值的酌减比率。

关键词：C4.5 算法；洪涝灾害；应急响应；酌减比率

Flood Disaster Emergency Response Level Analysis Based on C4.5 Algorithm

XU Guo[1,2], WU Yun[1], WANG Ru-Jing[1,2], ZHANG Jie[1,2]

[1](Institute of Intelligent Machine, Chinese Academy of Sciences, Hefei 230031, China)
[2](Department of Automation, University of Science and Technology of China, Hefei 230026, China)

Abstract: Launch the emergency response level effectively is a prerequisite for implementing emergency rescue scientifically and reasonablely when natural disasters occur. At present, emergency response grading standard is based on three emergency response indicators, which launches by Ministry of Civil Affairs. It provides the range of each response level's indicators, however, this scheme lacks quantization scale for elastic adjustment, according to specific disaster. For the above problem, emergency response level classifier was trained in this paper base on C4.5 algorithm and historical cases. It also obtain a reasonable reduction ratio for the traditional emergency response level. Through experimental results of flood disaster show that this method provides valuable reference to reduction ratio.

Keywords: C4.5 algorithm; flood disaster; emergency response; reduction ratio

中国是自然灾害频繁发生的国家之一，由此造成的损失也极大。经民政部核定，2010 年全国各类自然灾害共造成 4.3 亿人次受灾，因灾死亡失踪 7844 人，紧急转移安置 1858.4 万人次；农作物受灾面积 3742.6 万公顷，其中绝收面积 486.3 万公顷；倒塌房屋 273.3 万间，损坏房屋 670.1 万间；直接经济损失高达 5339.9 亿元。如何建立一套科学、有效的应急管理机制，增强应对突发自然灾害的能力，把损失减小到最大限度是我国现代化建设的一项战略性任务。

自然灾害的发生具有一定的突发性和随机性，且有时是难以避免的。因此，灾后应急救援管理就显得至关重要。其中，应急响应级别的确立是合理有效地实施应急救援预案的前提引导条件。传统的应急响应级别是以描述性评价方法为基础，根据民政部制定的因灾死亡人口等三个应急响应指标硬性地进行判断的。这样容易导致现启动的应急救援预案级别与期望级别之间出现差距。也就是说面对自然灾害不能够进行全面的，合理的规避和应对。

针对上述问题，国家减灾委提出了应急响应级别的酌减比率这一要求，但缺乏量化的依据。为此，本

① 基金项目:国家科技支撑计划(2008BAK49B05)
收稿时间:2011-07-29;收到修改稿时间:2011-09-05

文提出了一套改进的应急响应模型。该模型一方面保留了传统的应急响应分级标准。另一方面利用历史案例数据库，通过C4.5算法构造了应急响应级别决策树，得出了基于实际灾害信息的应急响应分级准则。即提供了一个弹性的，具有参考价值的、量化的酌减比率。最后，通过专家决策系统，综合分析得出最终的应急响应级别。

本文首先对 C4.5 算法的相关概念流程进行了介绍，其次阐述了改进的应急响应级别启动模型，最后以洪涝灾害为例验证了改进模型的可行性，表明该方法能够为应急响应级别的准确判断提供有价值的参考依据。

1　C4.5算法基本原理

1.1 基本定义

C4.5 是 Quinlan 在 1993 年提出的[1]，目前已经成为一种构造决策树分类规则的经典算法。该算法采用信息增益率(Gain_Ratio)来选择属性，在树构造过程中进行剪枝。它不仅可以处理离散值属性，还能处理连续值属性[2]。

在 C4.5 算法中，设 A={A1,A2,…Am}为特征属性，C={C1,C2,…Ck}为类别，D 为训练样本，Cj 为 C 的一个取值， ai 为 A 的一个取值，H(C) 是类别信息熵，H(C|A) 是按 A 将 D 分割后，其类别条件的信息熵。则信息增益率的计算公式如下[3,5]：

$$Gain\_Ratio = \frac{\text{信息增益 } I(C|A)}{\text{属性 A 的信息熵 } H(A)} = \frac{H(C)-H(C|A)}{-\sum_j P(a_j)\log_2(P(a_j))}$$

其中信息熵为：

$$H(C) = -\sum_j P(C_j)\log_2(P(C_j))$$

$$H(C|A) = -\sum_j \sum_i P(a_i)P(C_j|a_i)\log_2(P(C_j|a_i))$$

信息熵所对应的概率计算可以用训练样本中相应的统计值予以近似。假设，选择一个属性A_i，且A_i有不相交的属性值集合$\{a_1,a_2,\cdots,a_n\}$。那么A_i可将D划分为多个子集$D_1,D_2,\cdots D_n$。且D_i $(i=1,2,\cdots,n)$中对应的属性A_i所有的属性值均取a_i。设$|D|$为D的样本数，则$|D_{a(i)}|$为D中$a=a_i$的样本数；设$|C_j|$为整个训练集中C_j类的样本数，则$C_{ja(i)}$为D中$a=a_i$的样本中，属于C_j类的样本数。根据以上假设，则概率计算分别为：$P(C)=|C|/|D|$；$P(a_i)=|D_{a(i)}|/|D|$；$P(C_j|a_i)=|C_{ja(i)}|/|D_{a(i)}|$。

1.2 C4.5 算法流程

C4.5 算法的主要流程如下所述[6]：

输入：训练样本 D, 候选属性集合为 attri_ lists

输出：一棵决策树

具体步骤：

Step1：对 D 各项属性数据进行预处理

Step2：创建根结点，并确定 attr_ lists 叶结点属性

Step3：计算候选属性 attr_lists 中的每个属性，选取 Gain_Ratio 最大且同时获取的信息增益 Gain 属性又不低于所有属性平均值的属性作为测试属性

Step 4：将当前选中的属性赋值给当前结点，将该属性的属性值作为该属性的分叉结点，并且将这些分叉结点插入队列中

Step5：从候选属性 attr_ lists 中将当前使用属性删除

Step6：从队列中取出一个节点，递归进行 Step3 到 Step5，直到候选属性 attr_ lists 为空

Step 7：为每个叶子节点分配类别属性，对相同的类别属性进行合并，将其进行约减

2　应急响应级别启动模型的设计

2.1 传统的应急响应级别启动模型

传统的自然灾害应急响应级别是在描述性评价方法的基础上建立的。它包括{因灾死亡人口(R1)、紧急转移安置人口(R2)、倒塌房屋间数(R3)}等三个属性组成的规则集，本文分别表示为{RuleSet_1, RuleSet_2, RuleSet_3}。规则的具体内容如表 1 所示。其中，按照突发自然灾害的严重性和紧急程度，应急响应级别分为：I 级响应，II 级响应，III 级响应，IV 级响应，V 级响应[7]。特别地，这里的 V 级响应表示不启动响应。通过三组规则集可以分别求出对应的三项应急响应结果，以其最小值作为最终的应急响应级别。

从表 1 可知传统的应急响应存在如下问题。假设因灾死亡 29 人；紧急转移安置 87750 人；倒塌房屋 9983 间。那么，按照上述规则判断则不启动应急响应。虽然它们非常接近四级响应的临界值。通过分析实际的历史案例发现，上述假设应启动四级响应。对于这样的边缘问题如何根据灾害发生的实际特点定量的给出一个酌减比率是非常必要的。

表 1 传统应急响应级别规则集

响应 / 规则	I 级响应	II 级响应	III 级响应	IV 级响应	V 级响应
*RuleSet*_1	IF $R_1 \geqslant 200$ THEN I	IF $100 \leqslant R_1 < 200$ THEN II	IF $50 \leqslant R_1 < 100$ THEN III	IF $30 \leqslant R_1 < 50$ THEN IV	IF $R_1 < 30$ THEN V
*RuleSet*_2	IF $R_2 \geqslant 100$ 万 THEN I	IF 80 万$\leqslant R_2 <$100 万 THEN II	IF 30 万$\leqslant R_2 <$80 万 THEN III	IF 10 万$\leqslant R_2 <$30 万 THEN IV	IF $R_2 <$10 万 THEN V
*RuleSet*_3	IF $R_3 \geqslant$20 万 THEN I	IF 15 万$\leqslant R_3 <$20 万 THEN II	IF 10 万$\leqslant R_3 <$15 万 THEN III	IF 1 万$\leqslant R_3 <$10 万 THEN IV	IF $R_3 <$1 万 THEN V

2.2 改进的应急响应级别启动模型

针对传统的应急响应级别存在的边缘问题，本文提出了一种改进的应急响应级别启动模型。如图 2 所示。改进的模型一方面保留了传统的方法，得到一个规则响应级别。另一方面，利用历史案例数据库在基于C4.5算法训练出能够满足实际要求的应急响应级别分类器，如图 3 所示。通过它得到当前灾害的分类响应级别。最后，由减灾领域专家通过远程会商平台，对两种应急响应级别进行综合分析判断，最终得出应急响应级别。

图 3 所示的训练过程和测试过程都是利用存在于数据库中的灾害案例数据。数据预处理主要是指节选灾害数据属性值完整的案例；特征提取是根据那些对灾害损失影响比较大的致灾因子，并通过特征属性集预测目标属性应急响应级别。

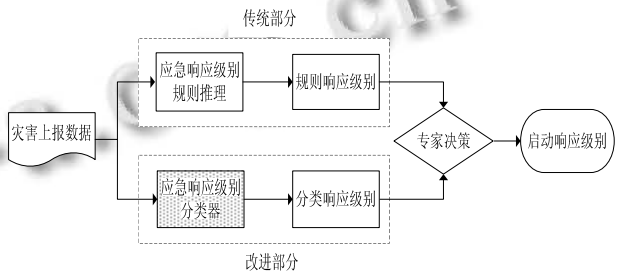

图 2 改进的应急响应启动模型

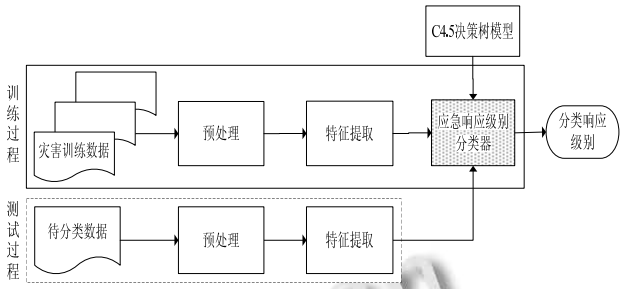

图 3 应急响应级别分类器的生成

表 2 洪涝灾害案例数据

No.	特征属性							目标属性
	A1	A2	A3	A4	A5	A6	A7	objA
1	13651000	822200	1240000	540000	33800	144000	600000	2
2	12520000	286000	673000	182000	76000	175000	535500	4
3	4502000	3700	264000	38000	4900	39000	94000	5
4	27053800	1151300	2318993	644700	272600	572000	1473400	1
5	2459000	7801	250500	55700	2679	12000	73000	5
6	6302000	457000	427000	100000	63000	151000	467000	3
…	…	…	…	…	…	…	…	…

3 实验结果与分析

根据民政部自然灾害情况统计报表，文中选取了 8 个特征属性：受灾人口(人)A1，紧急转移安置人口(人)A2，农作物受灾面积(公顷)A3，农作物绝收面积(公顷)A4,倒塌房屋间数(间)A5，损坏房屋间(间)A6，直接经济损失(万元)A7，灾害响应级别 objA 作为目标属性。

本文从历史案例数据库中抽取 100 条洪涝案例数据，如表 2 所示。其中随机选取 70%作为训练样本，

剩余的 30%作为测试样本。经过 8 次反复重新分组，进行实验得出分类器判断应急响应级别的分类正确率，如图 4 所示。

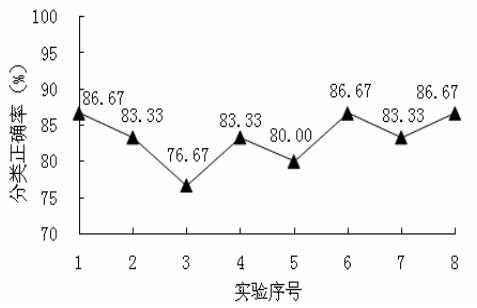

图 4　应急响应级别分类器的分类正确率

由图 4 可见，训练出来的应急响应级别分类器的分类正确率分布区间为[76.67, 86.67]，平均正确率达到 83.33%，说明分类结果具有较高的可信度。

以第六次实验为例进行分析可知，在 30 条测试数据中，有 26 条分类是正确的。其中，V 级响应即不启动响应的预测分类数目与实际分类数目完全符合；IV 级响应的预测分类数目比实际分类数目多一条；III 级响应的预测分类数目比实际分类数目少一条；II 级响应的预测分类数目比实际分类数目多两条；I 级响应的预测分类数目为零，而实际分类数目是两条。

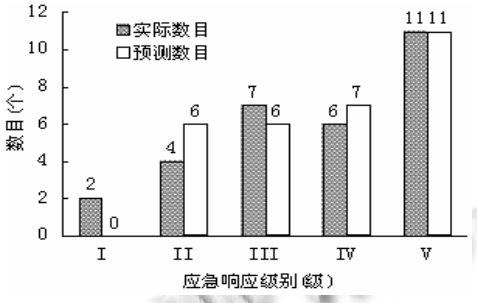

图 5　各级应急响应级别总数的对比

出现这种情况的原因和实验数据不足有关，特别是缺乏 I 级响应训练样本，导致无法正确预测 I 级响应。数据偏少会导致分类器得不到足够的训练。如果训练样本足够大，例如几千条数据甚至更多，那么分类器就会得到充分的训练，分类正确率也会相应提高。另外，本次实验通过信息增益率对生成的决策树进行修剪得到修剪后的应急响应级别决策树如图 6 所示。

结合传统的应急响应级别规则集可得如下结论：①影响洪涝灾害的主要元素为：倒塌房屋间数、紧急转移安置人口和受灾人口三个特征属性。这与国家减灾委的三个指标基本吻合；②对于“倒塌房屋间数”而言，四级响应与不响应级别之间的临界值由 1 万间降为 9300 间，酌减比率为 0.7%；③对于“紧急转移安置人口”而言，三级响应与四级响应级别之间的临界值由 30 万人降为 28.6 万人，酌减比率为 4.67%；

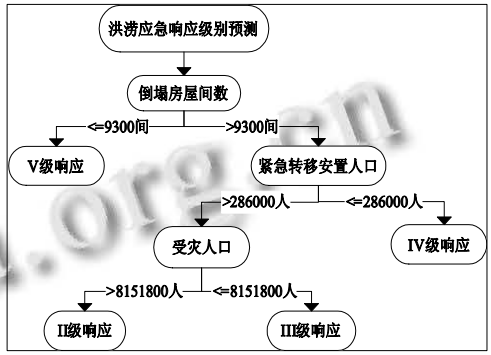

图 6　修剪后的应急响应级别决策树

可见，改进的应急响应级别启动模型可以为专家进行科学决策提供一个弹性的、有价值的参考依据。

4　结语

本文将 C4.5 算法应用于洪涝灾害的应急响应级别预测中。通过对洪涝灾害案例的八项特征属性进行有效的数据挖掘，构造出改进的应急响应级别启动模型。实验结果表明，该模型有效地解决了关于应急响应级别的酌减比率问题，使之可以成为减灾领域专家决策应急响应级别的参考。本文只进行了 100 条案例数据的实验，后续工作考虑增大历史案例数据量，得到分类正确率更高的分类器，从而提高该模型的性能。

参考文献

1 Quinlan JR.C4.5:Programs for Machine Learning.NewYork: Morgan Kaufman,1993.
2 朱明.数据挖掘.合肥:中国科学技术大学出版社,2002.
3 陈文才,黄金才.数据仓库与数据挖掘.北京:人民邮电出版社,2004.
4 Han JW, Kamber M. 数据挖掘概念与技术.北京:机械工业出版社,2004.
5 Tan PN, Steinbach M, Kumar V.数据挖掘导论.北京:人民邮电出版社,2006.
6 李楠,段隆振,陈萌.决策树 C4. 5 算法在数据挖掘中的分析及其应用.计算机与现代化,2008,12–1602–04.
7 李志宪.事故应急救助预案范例精选.北京:煤炭工业出版社,2007.