文章编号:1671 - 6833(2012)05 - 0030 - 04

# 基于改进模糊 K-M 法的岩溶塌陷危险性预测

冯 永[1], 陈鹏举[2]

(1. 河南工业大学 土木建筑学院,河南 郑州 450001; 2. 浙江省工程物探勘察院, 浙江 杭州 310005)

**摘 要**: 岩溶塌陷受诸多因素影响,而这些影响因素在进行安全性评价中通常用定性的数据形式给出,尽管定量预测的数学理论已相当成熟,但是处理这种大量定性数据问题却有很大的局限性. 因此,研究处理定性数据的预测方法对于开展岩溶塌陷的危险性具有重要意义. 模糊 K - M 算法采用各类中各属性值的频率作为聚类的中心,通过各属性的权重来修正目标间的距离,采用非随机的方法选取初始聚类的中心,有效地保证了复杂定性数据计算结果的精确性和可靠性. 在分析武汉市岩溶塌陷影响因素的基础上,利用此方法,对武汉市岩溶塌陷危险性进行研究,结果表明该种算法具有计算过程方便及收敛速度快等优点,值得在岩溶塌陷危险性预测中推广应用.

**关键词**: 岩溶塌陷; 改进模糊 K-M 法; 危险性预测

**中图分类号**: TU472.6 **文献标志码**: A doi:10.3969/j.issn.1671 - 6833.2012.05.007

## 0 引言

岩溶塌陷的影响环境非常复杂,影响因素种类繁多且相互制约[1],影响因素如岩性、地下水波动、人类工程活动等通常用定性的数据形式给出,近年来,许多学者对岩溶塌陷危险性评价方法进行了研究,采用了灰色统计、神经网络等新方法,尽管定量预测的数学理论已相当成熟,但是处理这种大量定性数据问题却有很大的局限性.

聚类分析是数据处理中一种重要的方法,对于定量的数据,模糊 k-means 算法是一种很有效的算法,然而,对定性数据的问题处理,模糊 k-means 算法就不再适合了. Huang 通过定义一个组合的非相似度,将 k-means 算法推广 k-modes 算法(简称为 K-M 法)和 k-prototypes 算法,用于处理定性数据(qualitative data)和混合值(mixed numeric and categorical values)数据,取得了良好的效果[2-4]. 但遇到特殊的属性类时,会出现属性值丢失现象. 作者引入一种改进的模糊 K-M 法,将会有效地克服属性值丢失现象. 作者以武汉市岩溶地面塌陷危险性预测为例,在分析岩溶塌陷影响因素的基础上,利用此方法对武汉市岩溶塌陷危险性进行了预测,探讨利用该方法预测岩溶塌陷的可行性.

## 1 改进模糊 K-M 法

### 1.1 模糊 K-M 法

模糊 K-Modes 算法是模糊 K-Means 聚类的推广算法,具体模型为

$$\min_{w,v} \sum_{j=1}^{c} \sum_{i=1}^{n} w_{ij}^{p} d(x_i, v_j), \tag{1}$$

其中 $\sum_{j=1}^{c} w_{ij} = 1$, 和 $w_{ij} \geqslant 0$, 对 $\forall i,j$, $n$ 为样本 $\{x_1, x_2, \cdots x_n\}$ 的个数,$c$ 为聚类个数,$\{v_1, \cdots, v_c\}$ 表示聚类中心集.

根据拉格朗日定理对上述模型求最优,可得到以下命题:

**命题 1** 对于固定中心集 $v_j, j = 1, \cdots c$,最优权重为 $w_{ij}$ 为

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} (d(x_i, v_j)/d(x_i, v_k))^{1/(p-1)}}. \tag{2}$$

### 1.2 改进模糊 K-M 法

一般中心只选取多个属性值中的一个,就忽

**收稿日期**:2012 - 03 - 28;**修订日期**:2012 - 05 - 17

**基金项目**:国家"十一五"科技计划资助项目(2011BAD03B00);河南省教育厅自然科学研究项目(2011B560003);郑州市科技发展计划项目(2001NYXM225);河南工业大学高层次人才基金(2009BS031)

**作者简介**:冯永(1984 - ),汉族,江苏徐州人,河南工业大学副教授,博士,主要从事岩土工程及工程地质的工作,E-mail:.

略了出现次数较少的属性值，对聚类的精度造成一定的影响. 所以采用每类中各属性的各属性值的频率来表示这一类的中心[5]，具体做法如下.

不妨设所有属性均为分类属性，其中第 $j$ 个属性 $A_j$ 有 $n_j$ 个属性值，即 $A_j$ 是 $\{A_{j1},\cdots,A_{jn_1}\}$ 的集合，$j=1,\cdots,m$，则第 $p$ 个聚类中心 $v_p$ 的第 $j$ 个分量可表示为：

$$v_{pj}=v_{pj(1)}A_{j1}+v_{pj(2)}A_{j2}+\cdots+v_{pj(n_j)}A_{jn_j},\quad(3)$$

其中，$v_{pj(k)}\geqslant 0, k=1,\cdots,n_j$，且 $\sum_{k=1}^{n_j}v_{pj(k)}=1$.

由于 $x_{ij}$ 必为某个 $A_{jq}$，不妨设为 $A_{jt}$，则样本 $x_i$ 与中心 $v_l$ 的第 $j$ 个属性间的距离可表示为

$$\|x_{ij}-v_{lj}\|^2=|1-v_{lj(t)}|^2+\sum_{q\neq t}|v_{lj(q)}|^2.\quad(4)$$

所以可令样本 $x_i$ 与中心 $v_l$ 的距离 $d(x_i,v_l)$ 为：

$$\begin{aligned}d(x_i,v_l)&=d_n(x_i^{(n)},v_l^{(n)})+\beta d_c(x_i^{(c)},v_l^{(c)})\\&=d_n(x_i^{(n)},v_l^{(n)})+\beta\sum_{t\in A^{(c)}}\|x_{it}-v_{lt}\|^2.\end{aligned}\quad(5)$$

式中：$d_n(x_i^{(n)},v_l^{(n)})$ 和 $d_c(x_i^{(c)},v_l^{(c)})$ 分别表示数值属性和分类属性间的距离；$d_n(x_i^{(n)},v_j^{(n)})$ 可为欧氏距离；$A^{(c)}$ 是分类属性的集合，把以上的(4)式带入(5)式就得到了新的距离定义.

将以上属性间距离定义应用到模糊 k-modes 算法定义的距离当中就得到了改进的聚类方法. 一般来说用类中样本的属性频率反映聚类中心比简单的取出出现最多的属性值做聚类中心要精确的多. 不妨设 $L$ 为最大循环次数.

具体算法如下：

(1) 从样本集中选取 $k$ 个聚类中心 $V^0=\{v_1^0,\cdots,v_k^0\}\in\mathscr{R}^{k\times m}$，可根据前面介绍的自动选取方法.

(2) 根据命题 1 计算各样本属于每一类的隶属度 $W^{(0)}$，根据该隶属度确定分类，由上公式(3)重新选取初始中心 $V^{(0)}=\{v_1^{(0)},\cdots,v_k^{(0)}\}\in\mathbf{R}^{k\times p}$，置 $t=1, l=1$.

(3) 根据命题 1 确定隶属度矩阵 $\boldsymbol{W}^{(t)}$，如果目标函数 $|f(\boldsymbol{W}^{(t)},V^{(t-1)})-f(\boldsymbol{W}^{(t-1)},V^{(t-1)})|<\varepsilon$，其中 $\varepsilon$ 为足够小的数，则算法终止，否则，进入第 4 步.

(4) 根据隶属度矩阵 $\boldsymbol{W}^{(t)}$ 确定分类，由上公式(3)式计算聚类中心 $V^{(t)}$，如果目标函数 $|f(\boldsymbol{W}^{(t)},V^{(t)})-f(\boldsymbol{W}^{(t)},V^{(t-1)})|<\varepsilon$，则算法终止；否则，令 $t=t+1, l=l+1$，进入第 5 步.

(5) 如果 $l>L$，则终止循环，否则，重复第 3 步.

### 1.3　初始聚类中心的选取方法

笔者提出可以根据以下方法求得初始聚类中心，设需选取 $k$ 个聚类中心，则

(1) 从原始数据中选取距离最远的两个样本，并将其定为两个初始的聚类中心 $v_1^{(0)}, v_2^{(0)}$. 令 $t=2$.

(2) 如果 $k>t$，则在原始数据中剔出之前选出的样本后，找出一样本 $x_i$ 使得

$$\min(d(x_i,v_1^{(0)}),\cdots,d(x_i,v_t^{(0)}))=\max_j(\min(d(x_j,v_1^{(0)}),\cdots,d(x_j,v_t^{(0)}))).$$

则令 $v_{t+1}^{(0)}=x_i, t=t+1$. 如果 $t=k$，则停止.

(3) 重复第 2 步，直到找到 $k$ 个初始聚类中心 $v_1^{(0)},v_2^{(0)},\cdots,v_k^{(0)}$.

## 2　基于改进模糊 K-M 法的岩溶塌陷危险性预测

### 2.1　武汉市岩溶塌陷地质环境简介

武汉市属于岩溶地面塌陷严重、多发区，自 1930 年以来，武汉地区已发生过 10 多次不同规模的塌陷，而且近年来发生频率有提高的趋势，影响到了城市空间利用，因此对该地区岩溶塌陷危险性进行准确预测具有十分重要的现实意义. 武汉市地质环境简述如下.

(1) 地层岩性：研究地层主要由第四系松散地层、石炭系 — 三叠系碳酸盐岩、第三系黏土岩和粉砂岩组成.

(2) 地质构造：区内影响岩溶发育的主要褶皱和断裂有关山向斜，青菱断裂等.

(3) 水文地质条件：含水层类型主要有孔隙承压水和裂隙岩溶水，两者在局部地区水力联系密切，地下水动态特征受长江影响显著，存在地下水开采井.

(4) 岩溶地质特征：区内岩溶类型分为埋藏型和覆盖型，岩溶在构造活跃处比较发育.

### 2.2　预测指标选取

在岩溶塌陷区域危险性指标方面，由于研究区域的不同，各自取得的指标也不尽相同[1,6-7]，对于武汉市的地质条件来讲，岩性，覆盖型岩溶分布及岩溶发育是岩溶发育的基础条件，而区域的构造情况(如关山向斜，青菱断裂)对于岩溶发育及产生也具有重要的影响，覆盖层是岩溶地面塌陷发生的物质基础，而水文地质条件是塌陷的重要因素，结合以上有关岩溶地面塌陷影响因素的

分析,考虑武汉市的基本情况以及目前的资料,借鉴相关文献[1,6-8],通过征询专家意见,最终选定如下的危险性评价指标体系(见表 1).

**表 1 岩溶塌陷危险性评估指标**

**Tab.1 The evaluation system for karst collapse in Wuhan**

| 一级指标 | 一级指标 |
|---|---|
| 岩溶基础条件 | 岩性 |
| | 岩溶分布类型 |
| | 岩溶发育程度 |
| | 与构造关系 |
| 覆盖层特征 | 覆盖层结构 |
| | 覆盖层厚度 |
| 水文地质条件 | 孔隙水和岩溶水联系情况 |
| | 枯水位 |
| | 洪水位 |
| | 地下水波动情况 |
| | 距离长江远近 |
| | 是否受水井影响 |
| | 是否在水域影响范围内 |

## 2.3 模型的建立

将研究区域按分成 1 633 个单元(150 m × 150 m),见下图 1. 以上每一指标变量均为定性数据且类别个数也不一样,甚至有的指标变量如距离长江远有 6 个级别,由于各区域单元格中会有各级别相叠加的情况,所以根据叠加情况,再次把各单元格中各属性对岩溶地面塌陷危险性影响程度进行划分,且每一属性均由定性数据表示,经重新划分后,指标变量如距离长江远就有 10 个级别,分别用0 ~ 9的整数表示. 共可得1 633个样本数据且均为定性属性.

根据前面介绍的改进的模糊 k-modes 算法的计算过程,将数据样本分为4类,令参数 $p = 1.2$,利用 matlab 编制以上计算步骤,结果如下:

① 初始聚类中心 $\boldsymbol{V}^{(0)}$ 为:

$v_1^{(0)} = \{1\ 0\ 1\ 4\ 0\ 5\ 2\ 0\ 6\ 7\ 9\ 0\ 0\}$;

$v_2^{(0)} = \{0\ 0\ 0\ 0\ 4\ 3\ 0\ 6\ 1\ 0\ 0\ 3\ 0\}$;

$v_3^{(0)} = \{4\ 2\ 3\ 1\ 0\ 3\ 3\ 4\ 2\ 3\ 5\ 3\ 1\}$;

$v_4^{(0)} = \{2\ 2\ 2\ 2\ 3\ 3\ 0\ 4\ 3\ 3\ 7\ 0\ 3\}$.

② 最终的聚类中心. 由于最终聚类中心由各类的各属性值的频率表示,且每一属性值的个数不一样,所以下面仅列举了最终聚类中心的前两个指标(岩性、岩溶分布类型)的值:

$v_1^{(0)} = \{0.613\ 0.38\ 0.0069\ 0\ 0,1\ 0\ 0\}$;

$v_2^{(0)} = \{0.0429\ 0.93\ 0.0226\ 0.0045\ 0,0.9752\ 0.0181\ 0.0068\}$;

$v_3^{(0)} = \{0\ 0.0206\ 0.4845\ 0.1237\ 0.3711,0\ 0.2577\ 0.7423\}$;

$v_4^{(0)} = \{0.0324\ 0.0081\ 0.1541\ 0.0946\ 0.7108,0.0351\ 0.0919\ 0.8730\}$.

## 2.4 预测结果

依据聚类结果,将1 633个单元的归属类别导入 MapGIS 软件,根据各单元类别,输出预测图,武汉市覆盖型岩溶地面塌陷危险性预测图,如图 1 所示.

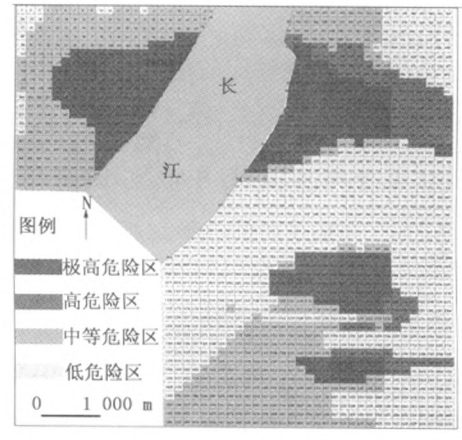


**图 1 武汉市岩溶塌陷危险性预测成果图**

**Fig.1 Karst collapse forecast in Wuhan city based on fuzzy k-modes algorithm**

通过以上分析,将研究区岩溶地面塌陷危险性分为极高危险区、高危险区、中等危险区及低危险区四级,与实际计算结果对比表明既往塌陷点均发生在极高危险区,而且研究成果和以往关于研究区的危险性分区成果大致相同,说明利用该改进模糊 K-M 法可以得到比较可靠的岩溶地面塌陷危险性预测结果.

# 3 结论

通过以上分析,作者提出的改进模糊 K-M 法岩溶地面塌陷预测中得到了很好的应用,可知这种改进的模糊聚类算法的突出优点是:

(1) 不需要人为量化指标数值,计算过程相对客观,提高了预测结果的可信性;

(2) 基于各属性值的频率来定义各数据间的距离,可以有效地避免属性值丢失的现象,从而可以综合考虑各因素的影响,使出现次数较少的属性值也对聚类结果有所贡献;

(3) 聚类中心的选取依据自动搜索的方法,该方法完全是基于聚类的基本思想,可以有效的保证选取的初始聚类中心,可代表不同的类别;

（4）在选取初始聚类中心时，其距离依据自定义距离计算，该距离不仅计算方便，而且可以更好地反映多值属性之间的差异，从而保证选取的聚类中心的有效性. 可以看出该算法不仅可以提高聚类精度，也提高了收敛速度.

笔者在分析武汉市岩溶塌陷影响因素的基础上，利用改进模糊 K-M 法对武汉市岩溶塌陷危险性进行了预测，研究结果表明该方法在岩溶地面塌陷危险性预测方面具有诸多优点，值得推广应用.

## 参考文献：


[1] 胡成，陈植华，丁国平，等. GIS 技术在岩溶塌陷预测中的应用[J]. 桂林工学院学报，2000，20(2)：117－119.

[2] HUANG Z. Extensions to the k-means algorithm for clustering large data set with categorical values[J]. Data Mining Knowledge Discovery，1998，2 (3)：283－304.

[3] BENATI S. Categorical data fuzzy clustering: An analysis of local search heuristics [J]. Computers &Operations Research，2006，7(4)：157－163.

[4] MICHAEL K N，JOYCE C W. Clustering categorical data sets using tabu search techniques[J]. Pattern Recognition，2002，35：2783－2790.

[5] 王宇，杨莉. 基于凝聚函数的混合属性数据聚类算法[J]. 大连理工大学学报，2006，46(3)：446－448.

[6] 雷明堂，蒋小珍，李瑜，等. 城市岩溶塌陷地质灾害风险评估－以贵州六盘水市为例[J]. 中国地质灾害与防治学报，2000，11(4)：23－27.

[7] 陈学军，陈植华，陈先华，等. 桂林市西城区岩溶塌陷模糊层次综合预测[J]. 桂林工学院学报，2000，20(2)：112－116.

[8] 张丽霞，熊大军，王集宁，等. 莱芜市岩溶塌陷原因分析与评价[J]. 山东地质，2002，18(3)：32－35.


## The Application of the Improved Fuzzy k-modes Algorithm to Forecast the Karst Collapse Hazard


Feng Yong[1]，Chen Peng-ju[2]

(1. Civil engineering school of Henan University of Technology，Zhengzhou 450001，China；2. Geology Exploration Institution of Zhejiang Province，Hangzhou 310005，China)



**Abstract**：The influence factors of karst collapse are very complex，and they are usually be expressed by categorical data. Although the mathematical theory of quantitative forecasting is quite mature，it has many limitations in processing large scale qualitative data. Therefore，it is necessary to introduce and apply processing qualitative data to forecast the ground collapse risk in karst region. For the improved fuzzy k-modes algorithm，the frequency of each attribute value of each attribute is taken as the cluster's center. The distance of objects can be modified by the weight of each attribute. A non-random method is applied to choose the clustering centers，which can ensure the clustering results of complex qualitative data accuracy and reliability. Therefore，in this paper，based on analysis on the influence factors in Wuhan karst collapses，the improved fuzzy k-modes algorithm was used to forecast the hazard potentiality of karst collapse in Wuhan city and good results have been obtained. It was found that the application of this model could get good effictiveness and this model should be adopted widely.



**Key words**：karst collapse；improved fuzzy k-modes algorithm；hazard forecast