计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(6):195−201 [doi: 10.15888/j.cnki.csa.006386]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

不确定 GM-CFSFDP 聚类算法在滑坡危险性预测中的应用①

胡　健[1], 覃　慧[2], 毛伊敏[2]

[1](江西理工大学 应用科学学院, 赣州 341000)
[2](江西理工大学 信息工程学院, 赣州 341000)
通讯作者: 覃　慧, E-mail: 1406054966@qq.com

摘　要: 针对滑坡危险性预测中降雨等不确定诱发因素难以有效处理, CFSFDP 算法需要人工尝试设置密度阈值以及对大规模数据集无法进行准确聚类等问题, 为了提高滑坡危险性预测准确度, 提出一种基于网格与类合并的不确定 CFSFDP(简称不确定 GM-CFSFDP) 聚类算法. 该算法首先引入不确定数据处理方法, 设计了 *E-ML* 距离公式, 有效刻画降雨不确定因素; 其次通过网格划分的思想把大规模数据集划分到多个网格空间中, 实现大规模数据有效编码; 计算网格平均密度, 建立网格密度阈值分布模型, 动态获得网格密度阈值; 最后利用层次聚类思想对关联性较高的类进行合并, 构建不确定 GM-CFSFDP 算法模型, 在延安宝塔区进行滑坡实例验证. 实验结果表明不确定 GM-CFSFDP 聚类算法获得较高的预测精度, 从而验证了该算法在滑坡危险性预测中的可行性和先进性.

关键词: 不确定数据; 滑坡; CFSFDP 聚类算法; 危险性预测

引用格式: 胡健,覃慧,毛伊敏.不确定 GM-CFSFDP 聚类算法在滑坡危险性预测中的应用.计算机系统应用,2018,27(6):195–201. http://www.c-s-a.org.cn/1003-3254/6386.html

Application of Uncertain GM-CFSFDP Clustering Algorithm in Landslide Hazard Prediction

HU Jian[1], QIN Hui[2], MAO Yi-Min[2]

[1](Collage of Applied Science, Jiangxi University of Science and Technology, Ganzhou 341000, China)
[2](Faculty of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: Since the rainfall and other uncertainties are difficult to effectively deal with in landside hazard prediction, as well as the density threshold in CFSFDP algorithm is required to be set manually and its low accuracy for large-scale data clustering, in order to improve the prediction accuracy, this study proposed an uncertain CFSFDP algorithm based on Grid and Merging clusters (uncertain GM-CFSFDP). Firstly, the *E-ML* distance formula based on uncertain data processing method is designed to effectively describe the uncertain factors of rainfall. Secondly, the idea of meshing is used to effectively encode the large-scale data by dividing it into multiple grid spaces. The average density of the mesh is calculated to establish the grid density threshold distribution model and obtain the grid density threshold dynamically. Finally, the hierarchical clustering idea is used to merge the higher association class and the uncertain GM-CFSFDP algorithm model is established. The experiments conducted in the Baota district of Yan'an show that the uncertain GM-CFSFDP clustering algorithm achieves a higher prediction accuracy and proves the feasibility and advancement of the algorithm in landslide hazard prediction.

Key words: uncertain data; landslide; CFSFDP clustering algorithm; hazard prediction

① 基金项目: 国家重点自然基金(41530640); 国家自然科学基金(41562019, 41362015); 江西省自然科学基金(20161BAB203093); 江西省教育厅科技项目 (GJJ151531); 江西省社科规划项目 (13YD020)
收稿时间: 2017-10-02; 修改时间: 2017-10-24; 采用时间: 2017-11-06; csa 在线出版时间: 2018-05-28

引言

滑坡灾害严重危害人类的生命财产安全, 并对环境、资源构成严重威胁[1], 给人们的生活带来了巨大影响. 滑坡的发生伴随着多种因素, 其中降雨是一个重要的诱发因素之一[2]. 由于降雨具有不确定性和随机性, 无法对其进行有效刻画, 因此给滑坡预测的准确性带来了一定的挑战.

聚类技术能够根据数据对象之间的较高相似度、聚簇之间的较高分离度实现数据对象的有效划分, 因而被广泛应用在滑坡灾害预测的研究中. 张俊等[3]使用滑坡面积比与分级面积比曲线对指标因子分级, 选取 7 个致灾因子作为滑坡易发性的评价指标, 采用 K-means 聚类算法对三峡库万州区滑坡易发性评价体系进行分级, 实验表明滑坡灾害易发性评价体系预测精度较高. 文建华等[4]提出同伦模糊 C-均值聚类算法, 以三峡库岸为研究区对边坡的稳定性进行分类, 研究表明同伦模糊 C-均值聚类算法是一种较好的边坡稳定性分级聚类分析方法. 孙树林等[5]以南京地区滑坡作为研究对象, 提取影响因素并计算其熵值, 利用 K-PSO 方法生成南京地区滑坡敏感图, 并行研究对比表明 K-PSO 聚类准确度高, 验证了其在滑坡敏感性分析的可行性. 吴亚子等[6]采用灰色聚类法, 并选取 11 个评价因子, 建立了阿里地区地质灾害危险性的评价模型, 结果表明利用灰色聚类方法对阿里地区公路沿线的危险性评价精度较高, 说明该方法具有一定可行性. 传统聚类技术在滑坡预测应用上取得了一定成果, 但是还不能满足人们的需求, 主要是存在以下两个问题: 1) 传统聚类算法很难实现对不确定数据降雨量的有效处理; 2) 传统聚类方法需要预先设定聚簇数目 k 值, 而在实际应用中 k 值难以准确给定, 致使对大规模数据集聚类结果影响较大. 针对传统聚类算法预先设定 k 值问题, Miin-Shen 等[7]构建一个基于学习的模糊聚类框架, 可自动找到最佳簇的数量, 实验结果证明该算法具有先进性; 赵文冲等[8]通过对 k 值的自动获取, 提高实验聚类结果, 但难以处理不确定数据. 以上两个问题致使传统聚类算法在滑坡危险性预测中的聚类结果不是很理想, 因此需要一种能够有效处理不确定数据和能够提升聚类效果的方法, 从而提高滑坡危险性预测精度.

快速搜索和发现密度峰值聚类算法 (CFSFDP)[9]可自动获得类的个数, 能够有效避免聚类数目 k 的预先设定, 算法复杂度相对较低, 可对任意形状的数据集进行聚类且实现简单聚类速度快. 但是 CFSFDP 算法无法有效处理不确定数据, 并且需要人工尝试设置密度阈值以及对大规模数据处理效果不佳, 因此文中在传统 CFSFDP 算法基础上, 提出不确定 GM-CFSFDP 聚类算法. 该算法首先建立不确定数据模型, 设计 *E-ML* 距离公式, 使其能够描述不确定属性之间的相似度, 有效刻画不确定因素降雨; 通过网格划分的思想按照维度将数据集进行网格化, 使之能够有效处理大规模数据; 借鉴平均密度思想建立网格密度阈值模型, 动态确定网格密度阈值, 避免 CFSFDP 需要人工尝试确定密度阈值; 利用层次聚类思想合并关联性较高的类, 解决大规模滑坡数据集密度分布不均匀的问题, 构建不确定 GM-CFSFDP 聚类算法滑坡预测模型, 以延安市宝塔区为例进行预测. 实例结果证明不确定 GM-CFSFDP 算法比 CFSFDP 算法在滑坡危险性预测中聚类效果更佳, 具有可行性.

1 不确定 GM-CFSFDP 算法

1.1 不确定数据的模型

假设不确定性数值属性A_{ij}, 其取值在一定范围内, 即$A_{ij}\in[a_{ij}{}^L, a_{ij}{}^R]$, $a_{ij}{}^L < a_{ij}{}^R$, 其中$a_{ij}{}^L$和$a_{ij}{}^R$分别称为A_{ij}的左界值和右界值. 若$A_{ij}\cdot g(x)$为A_{ij}的概率密度函数, 则有:

$$\begin{cases}\int_{-\infty}^{a_{ij}{}^L} A_{ij}\cdot g(x)dx = 0\\ \int_{a_{ij}{}^L}^{a_{ij}{}^R} A_{ij}\cdot g(x)dx = 1\\ \int_{a_{ij}{}^R}^{+\infty} A_{ij}\cdot g(x)dx = 0\end{cases}$$

1.2 不确定数据的处理

传统 CFSFDP 聚类算法能够处理离散型和连续型数据, 但难以对不确定数据进行有效处理. 文中结合不确定数据模型, 采用积分形式[10,11]考虑范围内点与点之间的差值, 再利用不确定数据的中点和长度, 替换左右界值对距离公式进行重新定义. 最后考虑含有离散型、连续型和不确定型的混合型属性数据, 对传统 Euclidean 距离进行拓展, 得到一种新的描述相似度的距离 (*E-ML* 距离) 公式.

定理 1. 设两个P维数据对象a和b均含有不确定属性, 则a和b的 *E-ML* 距离$d_{E-ML}(a,b)$为:

$$d_{E-ML}(a,b)=\sqrt{\sum_{i=1}^{p}\{[M(a_i)-M(b_i)]^2+\frac{1}{12}[L(a_i)-L(b_i)]^2\}}$$
$$\text{其中, } P\geq 1 \tag{1}$$

其中, $M(a)=\frac{a^L+a^R}{2}$和$L(a)=a^R-a^L$分别为不确定数据$a=[a^L,a^R]$的中点和长度. 离散型数据和连续型数据经过归一化处理之后均可看作是特殊的不确定数据, 此时$M(a)=a$, $L(a)=0$, 则 *E-ML* 距离可处理P维数据中包含离散属性、连续属性和不确定属性数据间的距离.

证明: 设不确定数据的区间为$a=[a^L,a^R]$, $b=[b^L,b^R]$, 给出如下定义[7]:

$$\begin{aligned}D^2(a,b)&=\int_{-1/2}^{1/2}\int_{-1/2}^{1/2}\times\{[(\frac{a^L+a^R}{2})+x(a^R-a^L)]\\&\quad-[(\frac{b^L+b^R}{2})+y(b^R-b^L]\}^2dxdy\\&=(\frac{a^L+a^R}{2}-\frac{b^L+b^R}{2})^2+\frac{1}{3}[(\frac{a^R-a^L}{2})^2\\&\quad+(\frac{b^R-b^L}{2})^2]\end{aligned} \tag{2}$$

指定$D(a,b)=\sqrt{D^2(a,b)}$为不确定数据a和b的距离. 但当$a=b$的时侯, 却存在$D(a,b)\neq 0$, 由定义易知, 对于任意的不确定数a和b恒有$D(a,b)>0$. 对公式 (2) 进行修正如下:

$$\begin{aligned}{d_{ML}}^2(a,b)&=\int_{-1/2}^{1/2}\{[M(a)+xL(a)]-[M(b)+xL(b)]\}^2dx\\&=[M(a)-M(b)]^2+\frac{1}{12}[L(a)-L(b)]^2\end{aligned} \tag{3}$$

记$d_{ML}(a,b)=\sqrt{{d_{ML}}^2(a,b)}$是$a$和$b$之间的距离, 其中$M(a)$和$L(a)$分别$a$的中点和长度, 可容易验证条件$d_{ML}(a,b)=0\Leftrightarrow a=b$.

当a和b为两个任意不确定P维数据时, 其 Euclidean 距离为$E(a,b)=\sqrt{\sum_{i=1}^{P}(a_i-b_i)^2}$, 结合公式 (3) 和传统 Euclidean 距离公式, 得到 *E-ML* 距离公式:

$$d_{E-ML}(a,b)=\sqrt{\sum_{i=1}^{p}\{[M(a_i)-M(b_i)]^2+\frac{1}{12}[L(a_i)-L(b_i)]^2\}}$$
$$\text{其中, } P\geq 1 \tag{4}$$

容易证明公式 (1) 满足距离定义的条件: 非负性、对称性和三角不等性, 说明 *E-ML* 距离具有合理性.

1.3 GM-CFSFDP 聚类算法

CFSFDP 聚类算法[9]可聚类任意形状数据集, 并且能够自动获取类的个数, 算法复杂度低, 然而仍存在不足: a) 算法聚类质量依赖于给定的密度阈值d_c; b) 大规模数据集存在规模大和密度分布不均匀, 算法虽然可以对数据点按密度值大小进行排序, 但聚类效果不够理想. 针对 CFSFDP 聚类算法需人工设置密度阈值、无法对大规模数据集进行准确聚类问题, 设计一种基于数据空间网格化的 CFSFDP 聚类算法 (GM-CFSFDP), 首先对数据进行数据空间网格化, 划分为不同的网格单元, 实现大规模数据的有效编码; 其次对密度阈值d_c进行动态选择, 引入平均密度思想, 将网格单元划分为稠密、中度、稀疏三种状态, 根据网格密度实现动态选择d_c; 最后借鉴层次聚类思想, 选取具有相关度较高的类进行合并, 获得聚类结果, GM-CFSFDP 聚类算法设计概念如下.

(1) 数据空间网格化

假设存在数据集$D=\{D_1,D_2,\cdots,D_d\}$, 采用自顶向下的网格划分方法[12]来对数据集进行划分, 将其归一化处理, 遍历数据集, 获得每个维度的长度l_i, 将数据空间按照维度$l_{i=m}$进行划分, 获得两部分数据空间, 再次对两个数据空间进行分割, 直至数据子空间满足点数目小于或等于阈值以及最短长度小于 2 倍密度阈值d_c, 得到空间集合U.

$$L=\{l_i/l_i=g(d_i)\} \tag{5}$$

$$m=g_{\max}(L) \tag{6}$$

$$U=\{d_1,d_2,\cdots,d_n\} \tag{7}$$

其中, L为长度l_i的集合, d为数据维度, $i\in d$, 函数g为求出d_i的长度, m为最长维度, 函数$g_{\max}$为L中最大值的编号.

(2) 网格密度阈值

采用平均密度公式计算所有网格平均密度阈值[13,14], 获取所有网格单元密度的最大值和最小值, 定义网格密度阈值d_c, 使网格单元分为稠密 ($f_i\geq f_{\text{Minpts}}$)、中度 ($f_{\text{Low}}\leq f_i<f_{\text{Minpts}}$)、稀疏 ($f_i<f_{\text{Low}}$)3 种, 若$d_c<f_{\text{Low}}$说明多数稠密的网格单元成为独立簇, 此时阈值设置过低需要增加调整, 若$d_c>f_{\text{Minpts}}$说明部分簇作为中度或稀疏单元格进行处理, 阈值设置过高需要降低调整, 依此保证d_c的取值范围和准确性. 根据网格密度选取网格所属密度阈值.

平均密度公式:

$$f_{ave}=\frac{\sum_{i=1}^{n} f_i}{n} \tag{8}$$

网格密度阈值公式:

$$f_{Minpts}=(f_{ave}+f_{max})/2 \tag{9}$$

$$f_{Low}=(f_{ave}+f_{min})/2 \tag{10}$$

其中, n为所有网格单元数目, f_i为第i个网格单元密度值, f_{max}为最大的网格单元密度, f_{min}为最小的网格单元密度.

目前在确定阈值的研究中, 学者们做了很多贡献, 其中近邻距离曲线[15]变化情况来确定密度阈值的方法, 解决了人工设置阈值的不足, 计算方法简述为先求出数据集的第 1 至第$2\%\times|S|$(其中 S 为数据集) 近邻距离曲线, 再找到曲线斜率变化明显的曲线, 记为第 r 条曲线, d_c取$i\sim j$数据点的所有第 r 条近邻距离的均值. 李宗林等[16]采用非参数核密度估计理论分析数据的分布特征来自动确定阈值. 两种方法都避免了人工尝试确定密度阈值的不确定性, 对于数据集规模较小时, 能得到明显的效果, 但在多数实际问题中数据集规模大, 上述方法确定密度阈值过程更复杂, 采用文中提到的阈值计算方法, 复杂度更小, 占用内存更少, 运行速度更快.

(3) 类合并

CFSFDP 算法无法准确对数据密度分布不均匀的数据集进行聚类[17], 原因是当数据集密度分布不均匀时, 算法可能会将一个类划分成两个或多个类, 此时需要进行子类合并. 借鉴层次聚类算法思想[18,19], 通过对比密度阈值d_c, 选择相关性较高的类进行合并, 从而实现准确聚类. 假设任意两个类 A、B, 其对应的网格密度阈值表示为d_{cA}、d_{cB}, 类 A、B 的边界区域点集S_A,S_B, 边界区域中的点数为$|S_A|$,$|S_B|$, p_i和q_j分别为S_A,S_B中的数据点, $dS_{p_iq_j}$为数据点p_i和q_j之间的距离, 公式如下:

$$\forall p_i\in S_A,\forall q_j\in S_B \tag{11}$$

$$d(A,B)=\min\{d_{cA},d_{cB}\} \tag{12}$$

若 A、B 满足类间相似度条件, 如公式 (13) 所示, 则将类 A、B 进行合并.

$$\frac{\sum_i\sum_j dS_{p_iq_j}}{|S_A|\times|S_B|}\le d(A,B) \tag{13}$$

1.4 不确定 GM-CFSFDP 聚类算法设计

不确定 GM-CFSFDP 聚类算法聚类过程如下:

Step 1. 数据进行归一化处理, 获得有效数据集;

Step 2. 根据数据空间网格化方法对有效数据集进行网格划分, 获得对应的数据空间集合;

Step 3. 使用平均密度思想和不确定数据处理方式对数据空间集合的各数据点进行局部密度和距离计算, 对网格单元密度进行划分, 进而动态确定密度阈值d_c;

Step 4. 使用 CFSFDP 算法对网格数据对象进行聚类, 确定聚类中心和初始聚类个数;

Step 5. 利用密度阈值d_c, 确定类的核心区域与边界区域, 指定边界区域中最高点密度值作为去除噪声点的阈值;

Step 6. 计算类之间的距离, 采用类合并方法, 判断类之间能否合并, 若满足合并条件则进行合并, 否则返回 Step 5;

Step 7. 退出合并操作, 输出数据集聚类结果.

2 实例研究及结果分析

2.1 实例研究

2.1.1 数据来源

实验数据来源于西安地质调查中心数据库, 采用 ARCGIS 将延安市宝塔区进行栅格化处理, 每个栅格单元尺寸设计为 5 m×5 m, 得到 5 672 922 个栅格单元, 每个栅格单元看成一个点, 借鉴刘卫明[20]的属性提取方法, 获得坡型、坡向、坡高、坡度数据信息, 以及岩土体结构数据、植被覆盖数据、降雨量值.

依据宝塔区的地质环境条件及地质灾害发生机理和原始数据集中各属性对聚类结果的影响程度选取坡型、坡向、坡高、坡度、岩土体、植被、降雨作为评价因子, 滑坡危险性等级作为决策因子. 其中坡型、植被、岩土体为离散属性, 先将其数值化再进行归一化处理; 坡度、坡高、坡向为连续属性可直接进行归一化方式处理; 降雨为不确定属性, 只能确定其大致取值范围, 无法直接用传统方法进行刻画, 因此采用文中提出的不确定数据处理方式进行处理.

2.1.2 不确定 GM-CFSFDP 聚类算法滑坡预测模型的构建

由延安市宝塔区经过栅格化处理的的 5 672 922 个栅格单元, 每个栅格单元被看成一个点, 这些点形成的数据集规模大, 因此首先采用不确定 GM-CFSFDP 聚类算法中的数据空间网格化步骤, 通过网格划分的思想把大规模滑坡数据划分到相应的数据空

间中, 最后得到数据空间网格单元 283 375 个; 初始化设置聚簇数目为空, 计算各个网格单元的平均密度, 依据密度阈值求解方法动态获得网格密度阈值d_c, 使用文中不确定数据距离公式 (*E-ML* 距离) 计算数据对象之间的距离; 然后使用 CFSFDP 聚类算法对各个滑坡数据空间网格单元进行聚类, 聚类时各个网格单元根据其合适的d_c进行聚类, 确定初始聚类中心位置和聚类个数, 初始得到聚簇数目为 558 个; 对其余非聚类中心的数据点进行归簇, 并利用密度阈值d_c确定簇边界区域, 计算两个相邻簇之间的相似度, 对所有相邻簇的相似度进行排序, 合并相似度较高的两个簇, 直到所有簇簇之间的相似度不满足合并条件为止, 最终得到 483 个簇, 依据簇内具有较高的相似度和簇间具有较高的分离度特征, 预测滑坡危险性等级.

2.1.3 滑坡危险性等级划分

滑坡危险性等级是滑坡危险性预测的决策因子, 因此如何正确划分滑坡危险性等级影响着滑坡危险性预测的精度. 聚类算法会把具有相似特征的栅格单元聚在一个子集中, 则子集内具有较高的相似度, 文中根据“具有相似特征的滑坡同时具有相似的滑坡发生趋势[21]”这一特性, 利用已知含有降雨信息的 293 个滑坡观测点的危险性等级, 采用直接搜索法和专家评分法[22]定各个区域的危险性等级. 首先利用直接搜索法, 对评价单元进行逐一搜索, 评价单元若只含有一个确定的危险性等级单元, 则该聚类子集的危险性等级为该单元的危险性等级, 若评价单元含有的各危险性等级单元不等, 则按照少数服从多数原则评定, 若未含有确定危险性等级单元和含有相同数目的不同危险性等级单元的聚类子集危险性等级则由专家根据经验进行评定, 结合区域调查结果判定滑坡危险性等级从而划分出其余单元的危险性等级.

2.2 实验结果分析与比较

2.2.1 实验环境

为了验证 GM-CFSFDP 聚类算法的有效性以及不确定数据处理方式能否提高滑坡危险性预测精度, 实验选择 Windows 7 旗舰版操作系统, 计算机硬件配置为 Inter i5 处理器、主频 3.3 GHz、8 G 内存, 实验数据通过 ARCGIS10.2 获取, 算法通过 JAVA 语言实现.

2.2.2 评价标准

基于误差矩阵的*Kappa*系数精度评价方法能够反映预测值和真实值的一致性[23], 其范围为[–1, 1], 其值越大, 表示预测值和观测值的一致性越大, 是一种滑坡危险性预测评价较好的方法, *Kappa*系数定义为:

$$Kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{14}$$

$$Pr(a) = \frac{\sum_{i=1}^{n} p_{ii}}{N} \tag{15}$$

$$Pr(e) = \frac{\sum_{i=1}^{n} P_{i+} \times P_{+i}}{N^2} \tag{16}$$

其中, $Pr(a)$表示观测和预测一致的数量与所有观测点的比例, $Pr(e)$表示同等级观测总和、预测总和占所有观测点的比例求和, p_{ii}为第 i 类型被正确分类的数目, P_{i+}为第 i 类型所在列的数目之和, P_{+i}为第 i 类型所在行的数目之和.

2.2.3 算法性能分析

为了验证 GM-CFSFDP 聚类算法的有效性, 分别按照 5%、10%、15%、20% 的比例对数据空间网格化后的 283 375 的网格单元进行采样, 对比 CFSFDP 聚类算法和 GM-CFSFDP 聚类算法的运行时间, 进行多次实验求取实验运行时间均值作为最后的聚类算法运行时间. 两种聚类算法的时间性能分析如表 1 所示.

表 1 滑坡数据集聚类算法运行时间对比

采样比例 (%)	CFSFDP(min)	GM-CFSFDP(min)
5	10.28	9.02
10	28.45	26.81
15	62.15	56.82
20	165.78	119.85

从表 1 可得, 数据采样比例为 5% 时, CFSFDP 算法的运行时间为 10.28 min, GM-CFSFDP 算法的运行时间为 9.02 min, 采样比例为 10% 时, 二者的运行时间分别为 28.45 min 和 26.81 min, 采样比例较小时, 二者算法运行时间相差不大, 这是因为对于小规模数据的处理, CFSFDP 和 GM-CFSFDP 都能快速的实现聚类效果. 当采样比例增大到 15% 时, GM-CFSFDP 算法的运行时间要比 CFSFDP 少 6 min 左右, GM-CFSFDP 算法在采样比例为 20% 时运行时间明显低于传统 CFSFDP 聚类算法, 这是因为在处理大规模数据时, GM-CFSFDP 算法设计了数据空间网格化思想, 能够快速的实现数据的编码, 密度阈值的动态选择实现了聚

类中心选择和聚类个数, 避免了需人工设置密度阈值和设置聚簇个数带来的问题, 类合并解决了数据集密度分布不均匀的问题, 提高了聚类效果. 通过整体采样实验发现, GM-CFSFDP 聚类算法的性能要高于 CFSFDP 聚类算法, 当数据规模越大, 效果越明显, 因此, GM-CFSFDP 聚类算法对于大规模数据而言聚类速度更快, 效果更佳, 可以作为一种处理滑坡大规模数据的方法.

2.2.4 滑坡预测精度分析与比较

为了验证不确定数据处理方式是否可以提高滑坡危险性预测精度, 比较传统 CFSFDP 聚类算法和不确定 GM-CFSFDP 聚类算法在滑坡实验中的预测精度. 传统聚类算法滑坡危险性预测中降雨通常以离散值进行处理, 采用定量法[24]将降雨分为六类: 小雨, 中雨, 大雨, 暴雨, 大暴雨, 特大暴雨, 使用传统 Euclidean 公式计算两个数据对象之间的距离, 构建传统的 CFSFDP 聚类算法滑坡危险性预测模型. 野外勘测获得延安宝塔区有 428 个滑坡灾害观测点, 其中有 293 个观测点含降雨量信息, 所有灾害观测点被栅格化为 1367 个单元, 其中 1036 个单元含降雨信息, 剩余 331 个为不含降雨信息的单元. 不确定 GM-CFSFDP 聚类算法利用不确定属性对降雨进行刻画, 使用 *E-ML* 距离公式计算数据对象之间的距离, 构建不确定 GM-CFSFDP 聚类算法滑坡危险性预测模型. 分别采用两类算法在宝塔区进行滑坡危险性预测, 依据滑坡危险性等级划分标准获得其等级划分, 并计算两种算法的预测精度 Pr(*a*) 和 *Kappa* 系数, 如表 2 所示.

表 2 两种算法滑坡危险性预测等级划分及预测精度比较表

聚类算法	预测 观测	低危	中危	高危	*Pr*(*a*)(%)	*Kappa*
CFSFDP 聚类算法	低危	385	36	12	88.88	0.8250
	中危	38	573	27		
	高危	15	24	257		
不确定 GM-CFSFDP 聚类算法	低危	393	26	14	93.27	0.8939
	中危	21	608	9		
	高危	12	10	274		

在满足相同的聚类条件时, 不确定 GM-CFSFDP 聚类算法的预测精度为 93.27%, 比传统 CFSFDP 聚类算法高出约 4 个百分点, *Kappa* 系数值是 0.8939, 传统 CFSFDP 聚类算法的 *Kappa* 为 0.8250, 说明不确定 GM-CFSFDP 聚类算法具有较好的滑坡危险性预测准确性. 结果分析不确定 GM-CFSFDP 的预测精度和 *Kappa* 系数值比传统 CFSFDP 聚类算法的较好, 原因是设计了数据空间网格划分理念, 实现对大规模数据的有效编码, 定义不确定数据距离公式, 有效的刻画了不确定属性降雨, 网格密度阈值的有效计算方法避免了人为设置阈值带来的误差, 利用层次聚类合并思想解决了由于大规模数据集密度分布不均匀导致的聚类效果不佳问题, 提高了滑坡危险性预测的精确度.

3 结束语

针对滑坡危险性预测中的诱发因素降雨刻画难、CFSFDP 算法对大规模数据集聚类不准确以及人为设置密度阈值等问题, 文中提出了不确定 GM-CFSFDP 聚类算法, 结合延安市宝塔区进行实例验证. 该算法设计新型 *E-ML* 距离公式, 实现不确定数据的有效刻画; 通过网格划分的思想对滑坡数据集进行数据空间网格划分, 实现了大规模数据有效编码, 利用平均密度思想构建密度阈值选择模型, 动态确定密度阈值, 对滑坡数据对象进行初始聚类, 最后合并关联性较高的类, 解决算法需人工设置密度阈值及处理大规模数据聚类效果不佳的问题. 实验结果表明不确定 GM-CFSFDP 聚类算法滑坡危险性预测具有较高的精度, 证明了该算法的可行性, 也为进一步的相关研究打下了基础.

参考文献

1 Huang FM, Huang JS, Jiang SH, *et al*. Landslide displacement prediction based on multivariate chaotic model and extreme learning machine. Engineering Geology, 2017, 218: 173–186. [doi: 10.1016/j.enggeo.2017.01.016]

2 Salciarini D, Fanelli G, Tamagnini C. A probabilistic model for rainfall —induced shallow landslide prediction at the regional scale. Landslides, 2017, 14(5): 1731–1746. [doi: 10.1007/s10346-017-0812-0]

3 张俊, 殷坤龙, 王佳佳, 等. 三峡库区万州区滑坡灾害易发性评价研究. 岩石力学与工程学报, 2016, 35(2): 284–296.

4 文建华, 周翠英, 黄林冲, 等. 边坡稳定性分类评价的同伦模糊 C-均值聚类算法. 岩土力学, 2012, 33(5): 1457–1461.

5 孙树林, 余文平, 刘小芳, 等. 基于信息熵与 KPSO 聚类法滑坡敏感性分析. 环境保护科学, 2014, 40(6): 88–96.

6 吴亚子, 杨敏. 灰色聚类法在阿里地区地质灾害危险性评价中的应用. 水资源与水工程学报, 2010, 21(6): 155–158.

7 Yang MS, Nataliani Y. Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters. Pattern Recognition, 2017, 71: 45–59. [doi: 10.1016/j.patcog.

2017.05.017]
8 赵文冲, 蔡江辉, 张继福. 改进 k 值自动获取 VDBSCAN 聚类算法. 计算机系统应用, 2016, 25(9): 131–136. [doi: 10.15888/j.cnki.csa.005325]
9 Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014, 344(6191): 1492–1496. [doi: 10.1126/science.1242072]
10 Tran L, Duckstein L. Comparison of fuzzy numbers using a fuzzy distance measure. Fuzzy Sets and Systems, 2002, 130(3): 331–341. [doi: 10.1016/S0165-0114(01)00195-6]
11 刘华文. 基于距离测度的模糊数排序. 山东大学学报 (理学版), 2004, 39(2): 30–36.
12 王飞, 王国胤, 李智星, 等. 一种基于网格的密度峰值聚类算法. 小型微型计算机系统, 2017, 38(5): 1034–1038.
13 邢长征, 王晓旭. 基于扩展网格和密度的数据流聚类算法. 计算机工程, 2014, 40(12): 188–194. [doi: 10.3778/j.issn.1002-8331.1207-0101]
14 米源, 杨燕, 李天瑞. 基于密度网格的数据流聚类算法. 计算机科学, 2011, 38(12): 178–181. [doi: 10.3969/j.issn.1002-137X.2011.12.040]
15 蒋礼青, 张明新, 郑金龙, 等. 快速搜索与发现密度峰值聚类算法的优化研究. 计算机应用研究, 2016, 33(11): 3251–3254.
16 李宗林, 罗可. DBSCAN 算法中参数的自适应确定. 计算机工程与应用, 2016, 52(3): 70–73.
17 孙昊, 张明新, 戴娇, 等. 基于网格的快速搜寻密度峰值的聚类算法优化研究. 计算机工程与科学, 2017, 39(5): 964–970.
18 乔端瑞. 基于 K-means 算法及层次聚类算法的研究与应用[硕士学位论文]. 长春: 吉林大学, 2016.
19 吕琳, 尉永清, 任敏, 等. 基于蚁群优化算法的凝聚型层次聚类. 计算机应用研究, 2017, 34(1): 114–117.
20 刘卫明, 高晓东, 毛伊敏, 等. 不确定遗传神经网络在滑坡危险性预测中的研究与应用. 计算机工程, 2017, 43(2): 308–316.
21 Yeon YK, Han JG, Ryu KH. Landslide susceptibility mapping in Injae, Korea, using a decision tree. Engineering Geology, 2010, 116(3-4): 274–283. [doi: 10.1016/j.enggeo.2010.09.009]
22 王磊, 张春山, 杨为民, 等. 基于 GIS 的甘肃省甘谷县地质灾害危险性评价. 地质力学学报, 2011, 17(4): 388–401.
23 邱海军. 区域滑坡崩塌地质灾害特征分析及其易发性和危险性评价研究[博士学位论文]. 西安: 西北大学, 2012.
24 辛鹏, 吴树仁, 石菊松, 等. 基于降雨响应的黄土丘陵区滑坡危险性预测研究——以宝鸡市麟游县为例. 地球学报, 2012, 33(3): 349–359.