Journal of Computers Vol. 32 No. 4, 2021, pp. 67-79
doi:10.53106/199115992021083204006

Uncertain GM-CFSFDP Clustering Algorithm for Landslide Hazard Prediction

Yimin Mao[1], Binbin Guo[1], Ruey-Shun Chen[2],
Yeh-Cheng Chen[3], Tao Tao[1], Deborah Simon Mwakapesa[1*]

[1] Institute of information engineering, jiangxi university of science and technology, GanZhou 341000, China
mymlyc@163.com, guobinbin@mail.jxust.edu.cn,
taotao@mail.jxust.edu.cn, and debrahmwaky@gmail.com

[2] Department of computer science, National Chiao Tung University, Hsinchu 30050, Taiwan
chen1868@gmail.com

[3] Department of Computer Science, University of California, Davis, CA, 95616, USA
ycch@ucdavis.edu

Received 10 December 2020; Revised 7 January 2021; Accepted 25 February 2021

Abstract. Due to difficulties in obtaining and effectively processing rainfall in landslide hazard prediction, as well as the existing limitation in dealing with large-scale data sets in clustering by Fast Search and Find of Density Peaks (CFSFDP) algorithm, a novel CFSFDP algorithm based on grid and merging clusters (GM-CFSFDP) has been proposed to assess landslide susceptibility model. Firstly, this method adopted a new two-phase clustering algorithm, which is suitable for large-scale data sets. Secondly, the uncertain data model is presented to effectively quantify triggering factors (precipitation). At the same time, a novel Euclidean distance formula based on midpoint and length of uncertain data ($E-ML$ distance formula) is designed, which makes the new method to manage the uncertain data. Finally, the prediction model of landslide hazards was constructed and verified in Baota district of Yan'an city. The experimental results show that the uncertain GM-CFSFDP clustering algorithm can effectively improve the accuracy of landslide hazard prediction.

Keywords: GM-CFSFDP clustering algorithm, hazard prediction, landslide, uncertain data

1 Introduction

Landslides are considered as movement of masses, mud flows, and failure in slopes. They have been recorded as among the worst natural hazard that have led to severe damages to properties and loss of lives in the Loess Plateau of Shaanxi Province [1]. Thus, the study of landslide prediction has an important significance. Landslides can occur as a result of topography, geological, physical factors and human activities, which are categorized as intrinsic and extrinsic causes. In most cases, landslides are usually triggered by unexpected torrential rainfall [2]. To quantitatively inspect the nature of landslides and forecast their occurrence is usually complicated due to perplexity and evaluation uncertainty of the causing factors [3].

Clustering methods divide data sets into groups (clusters) to maximize the intra-cluster similarity and minimize the inter-cluster similarity, thus play a significant role in the data mining field [4]. Thus, many scholars have conducted many studies on spatial prediction of landslide hazards using clustering technology. Based on the landslide hazard data in Badong County of Hubei Province, Gui Lei et al. [5] first chosen intrinsic and extrinsic factors closely related to landslide hazard as index in the zoning evaluation, then combined with Entropy method and Analytic Hierarchy Process (AHP) method to give a comprehensive assessment of the index weight. On the basis of the above results, all the evaluation units

* Corresponding Author

in this study area were classified and were identified automatically to different hazard levels by using a novel clustering algorithm. From the experiments, the results reflects the actual landslides distribution as they show closeness of the prediction result to the actual local situation. In particular, choosing four variables as the predominant factors of the susceptibility mapping, Hu Kaiheng et al. [6] employed the clustering analysis and Maximum Likelihood Classification (MLC) methods to map the susceptibility of post-quake geo-hazard in the Wenchuan earthquake area. According to the characteristic of landslide deformation, Deng Yong et al. [7] partitioned landslide hazard levels in the study area based on the fuzzy clustering algorithm. In the process of division, the threshold value K was precisely confirmed to evaluate the clustering result, which contributed to a zoned map of landslide susceptibility (landslide zoning classification). Pece V. Gorsevski1 et al. [+1] used fuzzy k-means method to classify continuous terrain, and used Bayesian probabilistic modeling method to study the case of the Clearwater National Forest (CNF) in central Idaho. Zhang Jun et al. [8] took Wanzhou district of the Three Gorges Reservoir region as the focus of this study. The slope stability was graded by adopting the K-means algorithm. The experimental results indicate that this method has better performance in the partitioning result. However, there are still two limitations based on the existing clustering algorithm for predicting landslide hazards. One is that precipitation is hard to quantify [9]. The other is that the threshold value K is difficult to accurately confirm in the existing K-means algorithm [10].

Aiming at the limitations of existing hierarchical schemes, such as the K-means algorithm, which requires manual determination of some parameters, Miin-Shen et al. [+2] constructed a learning-based fuzzy clustering framework that can automatically find the optimal number of clusters. Experimental results prove that the algorithm is advanced; ZHAO Wen-Chong et al. improved the experimental clustering effect by automatically acquiring the k value, but it is difficult to deal with uncertain data. The above two problems make the clustering effect of traditional clustering algorithm in landslide risk prediction not very ideal.

The CFSFDP algorithm can automatically obtain the number of clusters, which can effectively avoid the pre-setting of the number of clusters k, the algorithm complexity is relatively low, and it can cluster data sets of any shape [12-13]. However, there are three limitations in CFSFDP algorithm for predicting landslides hazards:1) CFSFDP algorithm needs to determine the cluster centers by setting a threshold of density artificial attempt, it is hard to be automated to run the algorithm; 2) CFSFDP algorithm is not suitable for managing large data sets, and it does not work well on large data sets. 3) It is difficult to obtain and effectively process rainfall in landslide hazard prediction. Therefore, an algorithm based on CFSFDP is needed, which can automatically determine the grid density threshold, can handle big data, and can effectively process uncertain data.

In this paper, an algorithm named GM-CFSFDP is presented to solve the above limitations. The GMCFSFDP algorithm first establishes a grid density threshold model, dynamically determines the density threshold, avoiding manual determination of the grid density threshold; then grids the data set so that it can handle large-scale data; finally establishes an uncertain data model, characterizing uncertain rainfall factors. According to the above theory, we finally constructed the prediction model based on the uncertain GM-CFSFDP algorithm. The experiment was carried out in the Baota district of Yan'an City and the results indicated better prediction accuracy from the proposed prediction model.

This paper is organized as follows: In section 2, a model of uncertain data and a processing of uncertain data are proposed, and the GM-CFSFDP algorithm based on them is described; In section 3, experiments were conducted using uncertain GM-CFSFDP algorithm based on data from Baota District of Yan'an City; In section 4, the experimental results are discussed and analyzed; Finally, in section 5, we summarize the experimental results and draw some perspectives for future developments.

2 Uncertain GM-CFSFDP Clustering Algorithm

2.1 Uncertain Data Model

Given a variable A_{ij}, which is an uncertain numerical attribute in a certain interval: $A_{ij} \in [a_{ij}{}^R, a_{ij}{}^L]$, $a_{ij}{}^L < a_{ij}{}^R$; in addition, $a_{ij}{}^L$ and $a_{ij}{}^R$ are called the left and right bounds of A_{ij}, respectively. If $A_{ij} \cdot g(x)$

is the probability density function of $A_{ij,}$ then $g(x)$ satisfies the conditions $\int_{-\infty}^{a_{ij}{}^L} g(x)dx = 0$, $\int_{a_{ij}{}^L}^{a_{ij}{}^R} g(x)dx = 1$, $\int_{a_{ij}{}^R}^{+\infty} g(x)dx = 0$ [14].

2.2 Processing of Uncertain Data

Landslide occurrence is nonlinear and a complicated process which is associated with various factors. Among those causative factors, rainfall value appears to lie within a certain interval, thus it is termed as uncertain factor. Processing such uncertain values using the existing clustering algorithms is difficulty. As a new clustering algorithm, CFSFDP clustering algorithm can operate both continuous and discrete data, but fails to effectively manage uncertain data such as rainfall. Therefore, to explore better the uncertain attributes features and then improve the prediction accuracy of the landslide model, we propose a new uncertain numerical data model called distance by combining with the Euclidean distance and the midpoint and length of uncertain data.

Theorem 1: Suppose a and b are two p-dimensional objects with uncertain attribute, their $E-ML$ distance is defined as:

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

where $M(a) = (a^L + a^R)/2$ and $L(a) = a^R - a^L$ are the midpoint and length of the uncertain data $a = [a^L, a^R]$, respectively. The discrete and continuous data that are normalized can be regarded as special uncertain data, such as $M(a) = a$, $L(a) = 0$. Therefore, a new distance definition ($E-ML$ distance) can be applied to p-dimensional data, including discrete attributes, continuous attributes and uncertain attributes.

Proof: Suppose the interval of uncertain data is $a = [a^L, a^R]$ and $b = [b^L, b^R]$, the definitions are as follows:

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

Suppose the distance between uncertain data a and b is $D(a,b) = \sqrt{D^2(a,b)}$ [15]. On the condition of $a = b$, so $D(a,b) \neq 0$. In view of definition, given arbitrarily two uncertain data a and b, therefore, $D(a,b) > 0$. The equation (2) is modified as follows:

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

Suppose the distance between a and b is $d_{ML}(a,b) = \sqrt{d_{ML}{}^2(a,b)}$, where $M(a)$ and $L(a)$ are the midpoint and length of a, respectively. Subsequently, $d_{ML}(a,b)$ satisfies the conditions $d_{ML}(a,b) = 0 \Leftrightarrow a = b$.

When a and b are two arbitrary uncertain p-dimensional data, the Euclidean distance between them is $E(a,b) = \sqrt{\sum_{i=1}^{p}(a_i - b_i)^2}$. According to the equation (3) and the Euclidean distance, the $E-ML$ distance formula is defined as:

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

2.3 GM-CFSFDP Clustering Algorithm

As a density-based clustering algorithm, CFSFDP is designed to cluster data sets of arbitrary shape, as well as automatically obtain the number of clusters. So CFSFDP clustering algorithm is a low-complexity algorithm. However, its shortcomings remain as follow: a) the clustering quality of the algorithm is sensitive to the density threshold d_c ; b) the result of this algorithm is heavily affected by the large-scale data sets, which are uneven density distribution. Considering the above problems, in this paper we propose a novel clustering algorithm called GM-CFSFDP algorithm to solve these problems. In this method, firstly the large-scale data sets are divided into several grid cells with different size, which realizes efficient coding for large-scale data sets. Secondly, we adopt the average density to divide the grid cell into dense, moderate and sparse states, and then the density threshold d_c can be dynamically got. Finally, the cluster of higher relevance is merged by using the hierarchical clustering to obtain the clustering results. Some definitions of the GM-CFSFDP clustering algorithm are as follows.

Data Space Meshing. Suppose the data set is $D=\{D_1, D_2, \cdots, D_d\}$, which is divided by the method of top-down grid partition [16], then the length of each dimension (l_i) is got after finishing the processes of normalization and traversal for the data sets. Subsequently, we select the dimension $l_{i=m}$ to divide the data space into two parts. Finally, we make the data space continually subdivide until the data subspace satisfies the situation that the number of data subspace is not more than the density threshold d_c , and the shortest length is twice less than the density threshold d_c . From above we get the space set U .

$$L=\{l_i \mid l_i=g(d_i)\}, i \in d, \tag{5}$$

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

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

where L and d are the set of length l_i and the data dimension, respectively. Then, the function $g(d_i)$ denotes the length of d_i , m is the maximum value of dimension, and the function $g_{\max}$ is the maximum value of L .

Grid Density Threshold. The maximum and minimum density threshold of all grid cells is got by using the average density formula [17], and then we determine the density threshold d_c to divide the grid cell into dense ($f_i \geq f_{Minpts}$), moderate ($f_{low} \leq f_i < f_{Minpts}$) and sparse ($f_i < f_{low}$) states. If $d_c < f_{low}$, it shows that most of the dense grid cells are regarded as independent cluster, so d_c should be increased. If $d_c > f_{Minpts}$, it indicates that the partial cluster is divided into moderate and sparse grid cell, thus d_c should be reduced. According to the above analysis, it can accurately obtain the certain range of d_c . The average density formula is defined as:

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

The grid density threshold formulas are as follows:

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

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

where n and f_i are the number of the grid cell and the density threshold of the i-th grid cell, respectively. Furthermore, $f_{\max}$ and $f_{\min}$ is the maximum and minimum density threshold, severally.

Thus far, many classical and effective methods have been proposed in this field of determining the density threshold. The method of determining the density threshold by the change of neighbor distance curve [18] can solve the problem that the density threshold is manually determined. The steps are as follows: (1) calculating the neighbor distance curve of data sets from 1 to $2\% \times |S|$ ($|S|$ is the scale of the data sets); (2) finding the curve whose slope changes obviously, then this curve is viewed as the r-th curve; (3) d_c is the mean value of all the r-th neighbor distance between the data point from i to j . In

addition, Li ZongLin et al. [19] used nonparametric kernel density estimation theory to analyze the distribution characteristics of the data to automatically determine the density threshold. Therefore, the above two methods avoid the uncertainty that determines the density threshold manually. However, they are complicated for dealing with large-scale data sets. In view of this disadvantage, in this paper we propose the method for calculating the threshold, which has been shown to be effective in processing large-scale data sets.

Merging Clusters. When the data is uneven density distribution, the cluster can be divided into multi-clusters which need to be merged. So, CFSFDP clustering algorithm cannot accurately cluster the data sets of uneven density distribution [20]. In order to solve the above problem, we contrast the density threshold dc by using the concept of hierarchical clustering [21] to merge the clusters with higher relevance. Suppose d_{cA} and d_{cB} are the density threshold of the random cluster A and B, respectively. S_A and S_B represent the point set of boundary region, p_i and q_j are the data in set S_A and S_B, severally. According to the above data, the formula is:

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

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

If A and B accord with the condition that the inter-cluster is similar, the cluster A and B will be merged as shown in the equation 13.

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

where dS_{piqj} is the distance between p_i and q_j . $|S_A|$ and $|S_B|$ are the number of point in the boundary region.

2.4 Design of the Uncertain GM-CFSFDP Clustering Algorithm

The processing of uncertain GM-CFSFDP clustering algorithm is as follows:

Step 1. Normalize the data, then the valid data set could be obtained.

Step 2. The valid data set is divided into the grid spaces by using the data space meshing, then the corresponding data space set is got.

Step 3. In order to automatically determine the density threshold d_c , we combine with the average density and method of the uncertain data processing to calculate the local density and distance of the data space set, and the grid cell is divided.

Step 4. The CFSFDP algorithm is used to cluster the grid data objects, and then determines the initial clustering center and the number of initial clusters.

Step 5. On the basis of known density threshold d_c , determine the core region and the boundary region of the cluster. Subsequently, density value of the highest point in the boundary region is designated as the threshold of noise points of removal.

Step 6. Calculate the distance between clusters according to the novel $E-ML$ distance formula. Then we adopt the merging clusters, if the clusters cannot be merged, turn to step e, otherwise merge the clusters.

Step 7. Exit merging operation, and output clustering results of data sets.

3 Landslide Hazard Prediction Using GM-CFSFDP

3.1 Data Preparation

The Baota district of Yan'an city was divided into 5,672,922 grids with of 25 m×25 m size, in the ArcGIS software. These were then imported into the digital elevation map (DEM) using their mean values, with a scale of 1:5000. From the DEM, the slope angle, slope type, slope height and slope aspect thematic maps were obtained [22]. Thereafter, from the thematic maps, topography and geomorphology as well other significant information were extracted. To obtain geotechnical data, a scale of 1:10,000 was

used. Furthermore, using the Near-infrared band B3 and the visible red band B2 of Spot5 in ENVI, the remote sensing images, the Normalized Difference Vegetation Index (NDVI) was computed. Precipitation is obtained based on the data in the past seven days and coming seven days from meteorological rainfall graph [23].

3.2 Data Processing

During this process, the original data was thoroughly inspected, whereby, some experiment-independent recorded data were erased; recorded data with missing values as well as repeated records needed to undergo processing before use, which also led to elimination of repetitions. In accordance to the geographic environmental conditions, research experience related to the Baota district geological disasters and previous landslide hazards reports, seven attributes including slope angle, slope type, slope height, slope aspect, type of rock and soil, rainfall, and vegetation index, were selected to evaluate landslide hazard in the area [24]. These attributes are categorized into 3: continuous, discrete (these two can directly be used in modelling) and uncertain attributes (which needs to be processed, in this paper, $E-ML$ distance formula was applied) as shown in Table 1. After all these procedures, a total of 4,856,723 records containing the seven landslide attributes.

Table 1. Seven attributes of landslide data sets

Property item	Property type	Discrete attribute value
Slope angle/°	Continuous	
Slope height/m	Continuous	
Slope aspect	Continuous	
Vegetation index	Discrete	Low, lower, high, higher
Slope type	Discrete	Straight, stepped, convex, concave
Rock and soil	Discrete	Loess + nearly horizontal paleo-soil, Loess + inclined paleo-soil, Loess + paleo-soil layers +bedrock, Loess + paleo-soil layers + the Neogene clay
Rainfall/mm	Uncertain	

3.3 Model Construction

Experimental Data. In study area, 428 landslides were investigated, in which 293 points with precipitation information were as associated with landslides. All of these landslides have been rasterized, we finally obtained 1367 valid data, which were used to evaluate prediction model of landslide hazard and to verify the accuracy of the prediction model. Part of the records are shown in Table 2.

Table 2. Experimental data

Sequence	Slope height /m	Slope angle/°	Slope aspect /°	Slope type	Vegetation index	Rock and soil	Rainfall/mm
325	150	60	250	Straight	High	loess+inclined, paleo-soil	80-100
326	140	35	110	Stepped	Higher	loess +inclined, paleo-soil	56-99
327	180	27	305	Straight	Higher	loess + paleo-soil layers, + the Neogene clay	87-103
328	160	40	45	Stepped	High	loess +inclined paleo-soil	88-112
329	180	30	125	Convex	Higher	loess + paleo-soil layers, + the Neogene clay	12-43
…	…	…	…	…	…	…	…

Mapping of Landslide Hazard. In the clustering algorithm, the spatial unit with similar characteristics (topography, geology, etc.) is aggregated into a sub-cluster, which means that the evaluation unit in the

same sub-cluster is always similar. According to the above characteristics, the prediction model based on the uncertain GM-CFSFDP clustering algorithm is constructed. Subsequently, 4856723 records are finally clustered into 483 clustering subsets, as shown in Fig. 1, and the grid cells in the same sub-cluster are always similar. Based on the theory of "one is similar to the characteristic of landslide development, which has the similar tendency of landslide occurrence" [25] and combining with the direct search method and expert evaluation method [26], The process of determining the hazard level of sub-clusters is as follows.

The example of each grid data in the cluster subset

ID	Angle	Elevation	Curvature	Aspect	Rock Soil	Plant Cover	Precipitation
1	25	39	-0.039	S	1	0.531	22-247
2	19	32	-0.042	SE	2	0.534	24-235
...	...	...	...	...	...	...	...
98	23	34	-0.0034	SE	2	0.581	31-198
...	...	...	...	...	...	...	...
156	30	38	-0.045	S	1	0.648	32-178
...	...	...	...	...	...	...	...

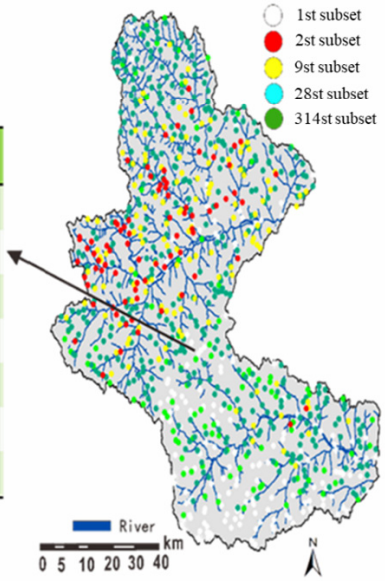

Fig. 1. 483 clustering subsets based on uncertain GM-CFSFDP algorithm

On the basis of direct search method, the evaluation units in the sub-clusters are searched one by one, and the hazard level of the sub-clusters is determined according to the hazard levels of the evaluation units in the sub-clusters. The process of determining the hazard level of sub-clusters is shown in Fig. 2. First, search for a sub-cluster and calculate the number of evaluation units that determine the hazard level in the sub-cluster and the number of evaluation units for each hazard level. Then, assume that the sub-cluster contains only one evaluation unit that determines the hazard level, the hazard level of sub-clusters is equal to that of the evaluation unit within sub-clusters, assume that the hazard levels of two or more evaluation units are determined, and the number of evaluation unit of different hazards varies, the hazard level of sub-clusters is confirmed based on the principle of majority rule [27], assuming that there is no evaluation unit for determining the hazard level, or the hazard level of two or more evaluation units is determined and the amounts of the evaluation unit of different hazards is same., combining with previous experience and relevant knowledge of regional geological environment, the experts make full use of geological investigation results to divide the landslide hazards of the maintaining evaluation unit in the study area. Finally, if there are sub-clusters that have not been searched, repeat the above operation, else, the process of determining the hazard level of the sub-cluster is over.

Through the sub-cluster hazard level determination process and the known hazard level of 293 landslides recorded precipitation information, these hazard levels of 483 clustering subsets are obtained. In these 483 sub-clusters the proportions of the five levels of very high, high, medium, low and very low hazard are 13.86%, 12.13%, 40.25%, 11.87% and 21.89% respectively. The very high and high hazard areas are mainly scattered in the Yanhe River Valley area where the vegetation coverage of the Yanhe River Basin is low; the medium hazard areas are located in the loess landform area and the southern area of the Fenchuan River outside the Yanhe River Basin in the northern part of the Baota District; the low and very low hazard areas are mainly scattered in the southern area of Baota District. The results of the zoning are consistent with the development status of geological hazards in Baota District, and compared with the planning map of the geological hazard zone provided by the Xi'an Geological Disaster Management Center, the results of the zoning are also consistent with it.

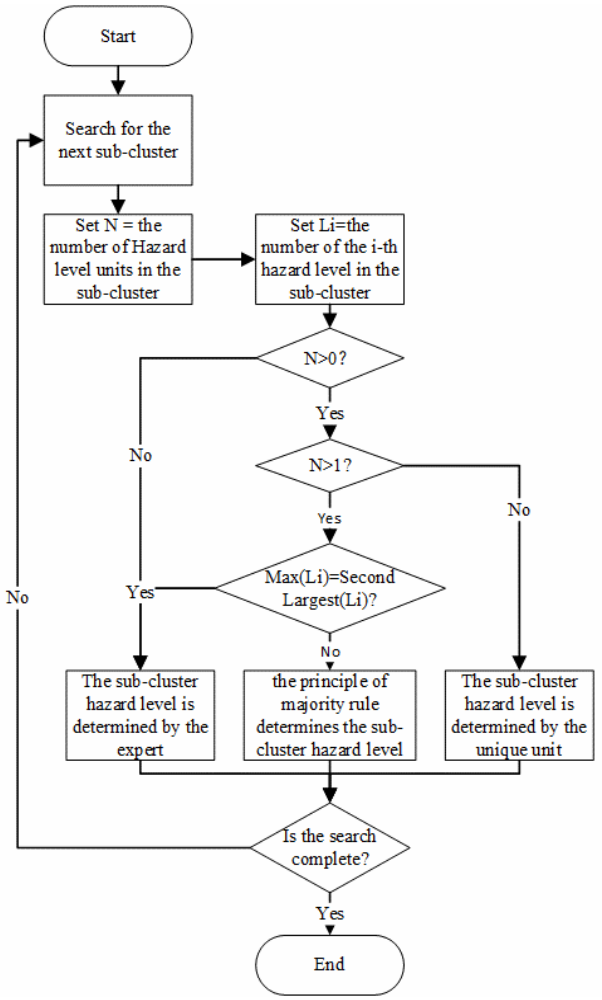

Fig. 2. Determination of sub-cluster hazard level

4 Result and Discussion

4.1 Experimental Environment

An experimental computer was configured with the Windows 7 operating system, an Intel i5 dual-core processor, 3.3-GHz frequency and 8.0 GB of memory. The algorithm was programmed to construct landslide hazard mapping and run on high-performance and mainframe computer, its maximum computing speeds can reach four billion times per second. Experimental data were obtained by using ARCGIS 10.2.

4.2 Evaluation Standard

To evaluate the landslide prediction model, error matrix method is commonly used. In this paper, we establish the error matrix through the statistics of landslide actual survey data and landslide prediction results in the study area. Besides, *Kappa* coefficient (k) [28] is used to estimate the consistency between the predicted value and the actual value, its value range is $[0,1]$. k value close to 0 is an indication that there is no agreement between the landslide model and actual data, while k value of 1 indicates complete agreement. The calculation formula of *Kappa* coefficient is shown as follow:

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

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

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

where $Pr(a)$ denotes the prediction accuracy of evaluation methods. P_{i+} and P_{+i} indicate the total records numbers of i row and i column, respectively. Besides, N is the size of samples. n shows the number of categorical types, and we set its value equal to 5 in this experiment.

4.3 The Analysis and Comparison

The effectiveness analysis of distance. In order to verify the effectiveness of $E-ML$ distance in measuring the uncertainty of rainfall, we selected the Euclidean distance, Hausdorff distance and $E-ML$ distance to measure the rainfall to analysis clustering results, respectively. Fig. 3 shows the clustering results of the algorithm with three kinds of distances under different scale data sets. It can be seen from the Fig. 3 that the sil (Silhouette) value of the $E-ML$ distance reaches 0.8, so it is superior to the other two distances. Euclidean distance completely ignores the uncertainty of data in measuring rainfall, therefore, the clustering results of algorithm with Euclidean distance are not accurate. Although the uncertainty of rainfall is considered by Hausdorff distance, which still loses some internal important information. It is better than Euclidean distance, but it is lower than $E-ML$ distance. The $E-ML$ distance makes full use of the internal information of the uncertain data, so it obtains a better clustering effect.

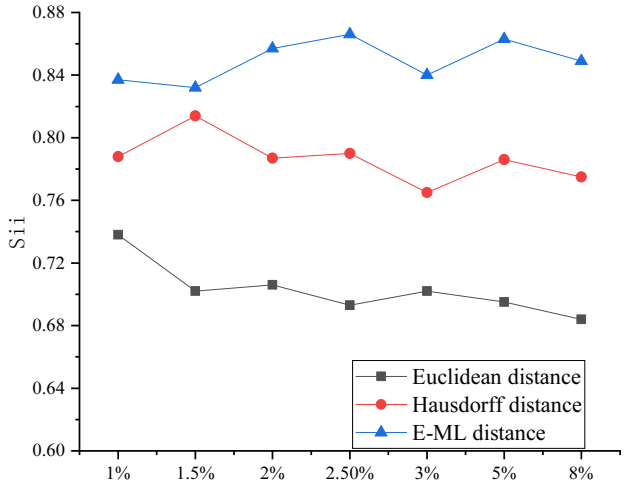

Fig. 3. Comparison of sil value of different distance

Comparison of Time Performance. To verify the performance of GM-CFSFDP clustering algorithm, we randomly choose 1%, 2% 5%, 8% and 10% data volume from 4856723 records as experimental data. All evaluation units were used to confirm the susceptibility of landslides. According to the clustering characteristic, they are clustered into 483 clusters. A certain amount of data was randomly selected as samples to test time performance of CFSFDP, GM-CFSFDP, SYNC and FAKCS algorithms. The average running time of these four algorithms was obtained based on 10 repeated experiments by using the same processing approach. The experimental results were obtained as shown in Fig. 4.

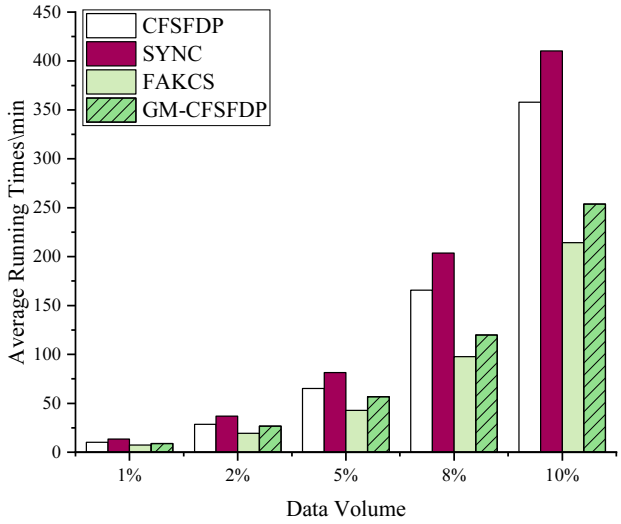

Fig. 4. Comparison of average running time based on two methods

As we can see, the average running time of these four algorithms is similar at the 1% data size. Suppose the data volume is 2%, the average running time of the GM-CFSFDP algorithm is slightly lower than the FAKCS algorithm, and the average running time of the CFSFDP algorithm is the highest. As the number of samples increases, the difference in average running time based on the four algorithms gradually increases. In general, the time performance of the GM-CFSFDP algorithm and the FAKCS algorithm is better under five different sampling ratios. This is mainly because the SYNC algorithm needs to perform iterative calculations on each component, which has high time complexity; The CFSFDP algorithm needs to calculate the boundary area of each cluster, and the time complexity is only lower than that of the SYNC algorithm; The FAKCS algorithm greatly reduces the calculation amount of the algorithm by compressing the original data set, and has better time performance when processing large data sets, but the compressed data set cannot fully represent the original data set, so in the next paragraph, the accuracy of the FAKCS algorithm is lower than the GM-CFSFDP algorithm. Considering the prediction accuracy and time performance, the effectiveness of the GM-CFSFDP algorithm is better than the other three algorithms.

Comparison of Accuracy. To evaluate the performance of the uncertain GM-CFSFDP algorithm, the different proportions of data as the experimental data is selected to compare the Sil (Silhouette) value of the GM-CFSFDP algorithm with the CFSFDP algorithm, the SYNC algorithm, as well as the FAKCS algorithm. The experimental data were taken for 1%, 2%, 5%, 8%, and 10% of all sample data, and each algorithm deals with continuous and discrete attributes data in the existing way, and the $E-ML$ distance formula is used to deal with uncertain data.

The comparison is shown in Fig. 5. As it can be seen, the GM-CFSFDP algorithm has better prediction accuracy than the other three algorithms. The clustering quality of CFSFDP algorithm is poorer whose average prediction is just about 75%. The reason is that the existence of some indistinct cluster of clusters, and the CFSFDP algorithm is difficult to discovery this type cluster. From the line of the SYNC algorithm, it is easy to find that the larger the data set, the worse the prediction accuracy, so the SYNC algorithm has limitations on the large datasets. And for the FAKCS algorithm, the quality is slightly better than CFSFDP algorithm and SYNC algorithm, and its average prediction accuracy is about 85%, which is still lower than the GM-CFSFDP algorithm with prediction accuracy of 90%. The reason is as follows: Firstly, the large-scale data sets are divided into several grid cells with different size, which realizes efficient coding for large-scale data sets. Secondly, the GM-CFSFDP algorithm simplifies the parameter selection and adaptively adjust the grid density threshold, which makes the algorithm convergence faster in the early phase. Finally, the cluster of higher relevance is merged by using the hierarchical clustering to obtain the clustering results, which speeds up the global optimization speed and the convergence speed. In conclusion, the GM-CFSFDP algorithm has the best prediction effect.

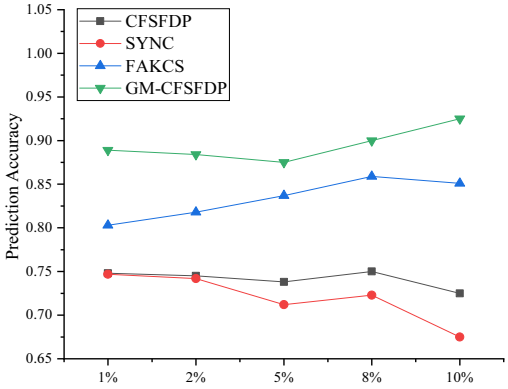

Fig. 5. Comparison of prediction accuracy of different algorithms

4.4 Comparing the Performance of the Prediction Models

To evaluate the performance of the uncertain data processing method in this paper, we used separately quantitative method and uncertain data processing method to availably quantify the rainfall. The following compared their clustering results.

The uncertain rainfall value, at the moment, are processed using a quantitative method [29] that classifies the value as: below 20 mm is light rain, 20.0~44.9 mm is moderate rain, 45.0~69.9 mm is heavy rain, a rainstorm is 70.0~99.9 mm, a heavy rainstorm is 100.0~250.0 mm and above 250.0 mm is super heavy rain. Using the quantitative method, the rainfall value was directly processed as a discrete variable. Subsequently, similarity degree of objects was calculated by the Euclidean distance formula. In contrast, the uncertain GM-CFSFDP clustering algorithm considered precipitation as an uncertain numerical value. The uncertain data processing method was used to quantify rainfall. Then a novel $E-ML$ distance formula was designed to calculate the similarity degree of objects. Therefore, the traditional CFSFDP, SYNC, FAKCS clustering algorithm and the Uncertain GM-CFSFDP algorithm landslide risk prediction model are established.

According to formulas 14, 15 and 16, using the CFSFDP algorithm, the prediction accuracy and Cohen K*appa* of the four algorithms are calculated, as shown in Table 3. The prediction accuracy of the uncertain GM-CFSFDP algorithm and Cohen *Kappa* is the highest, with a prediction accuracy of 93.27%, and Cohen Kappa is 0.8939. Comparing the four models, the uncertain GM-CFSFDP model outperforms the CFSFDP, SYNC and FAKCS models. Experiments show that the predicted results are basically consistent with the actual situation. Due to the proposed uncertain data model, the prediction accuracy and Cohen coefficient have been generally improved. In addition, this uncertain data processing method used to quantify rainfall, effectively unravel the limitations of the present methods. Besides, the Euclidean distance for calculating similarity degree was extended so that the new method could process the data with uncertain attributes. Thus, the total prediction accuracy was improved to a certain extent.

Table 3. Comparison of the accuracy and Cohen kappa coefficient

Clustering model	prediction observation	Low-hazard	Middle-hazard	High-hazard	Pr(a) (%)	*Kappa*
CFSFDP	Low-hazard	385	36	12	88.88	0.8250
	Middle-hazard	38	573	27		
	High-hazard	15	24	257		
Uncertain GM-CFSFDP	Low-hazard	393	26	14	93.27	0.8939
	Middle-hazard	21	608	9		
	High-hazard	12	10	274		
SYNC	Low-hazard	371	28	34	85.74	0.8047
	Middle-hazard	49	542	47		
	High-hazard	16	34	246		
FAKCS	Low-hazard	384	15	34	89.35	0.8434
	Middle-hazard	25	587	26		
	High-hazard	18	13	265		

5 Conclusion

It is generally difficult to obtain and effectively deal with rainfall in landslide hazard prediction. Moreover, CFSFDP algorithm has poor performance in processing large-scale data sets. Subsequently, in order to deal with these problems, the uncertain GM-CFSFDP algorithm has been proposed. In this method, the $E-ML$ distance formula is designed by combining with Euclidean distance formula as well as the midpoint and length of uncertain data, which solves the problem that the rainfall value cannot be effectively depicted and proposed. Finally, the experimental results illustrate that the overall accuracy of the prediction model is higher than those of the existing prediction model and provides a scientific method for landslide prediction.

The performance of GM-CFSFDP algorithm needs more experiments and larger-scale data to verify. The stability of the algorithm is not very good, and the accuracy of the algorithm can be further improved. The next work will try to combine with the Ensemble Learning to improve the stability of the algorithm.

Acknowledgements

This study was supposed by the National Natural Science Foundation of China (41562019, 41530640), the National Key Research Development Projects of China (2018YFC1504705), and Jiangxi Provincial Technology of Education Department (GJJ161566, GJJ181504).

References

[1] Y. Mao, M. Zhang, P. Sun, G. Wang, Landslide susceptibility assessment using uncertain decision tree model in loess areas, Environmental Earth Sciences 76(22)(2017) 752.

[2] D. Salciarini, G. Fanelli, C. Tamagnini, A probabilistic model for rainfall—induced shallow landslide prediction at the regional scale, Landslides 14(5)(2017) 1731-1746.

[3] J.S. Pan, C.Y. Lee, A. Sghaier, M. Zeghid, J. Xie, Novel systolization of subquadratic space complexity multipliers based on toeplitz matrix–vector product approach, IEEE Transactions on Very Large Scale Integration (VLSI) Systems 27(7)(2019) 1614-1622.

[4] T.Y. Wu, C.M. Chen, K.H. Wang, C. Meng, E.K. Wang, A provably secure certificateless public key encryption with keyword search, Journal of the Chinese Institute of Engineers 42(1)(2019) 20-28.

[5] L. Gui, K.L. Yin, J.J. Wang, Landslide hazard zonation based on cluster analysis, Hydrogeology & Engineering Geology 40(1)(2013) 100-105.

[6] K.H. Hu, P. Cui, Y.S. Han, Y. You, Susceptibility mapping of landslides and debris flows in 2008 Wenchuan earthquake by using cluster analysis and maximum likelihood classification methods, Science of Soil and Water Conservation 10(1)(2012) 12-18.

[7] Y. Deng, Z.L. Zhang, N.S. Xie, G.C. Lv, Research on fuzzy clustering and its application, Science of Surveying and Mapping 35(4)(2010) 163-165.

[8] J. Zhang, K. Yin, J. Wang, L. Liu, F. Huang, Evaluation of landslide susceptibility for Wanzhou district of Three Gorges Reservoir, Chin. J. Rock Mech. Eng 35(2016) 284-296.

[9] C.M. Chen, K.H. Wang, K.H. Yeh, B. Xiang, T.Y. Wu, Attacks and solutions on a three-party password-based authenticated key exchange protocol for wireless communications, Journal of Ambient Intelligence and Humanized Computing 10(8)(2019) 3133-3142.

[10] C.M. Chen, B. Xiang, Y. Liu, K.H. Wang, A secure authentication protocol for internet of vehicles, Ieee Access 7(2019) 12047-12057.

[11] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344(6191)(2014) 1492-1496.

[12] J.S. Pan, P. Hu, S.C. Chu, Novel parallel heterogeneous meta-heuristic and its communication strategies for the prediction of wind power, Processes 7(11)(2019) 845.

[13] T. Ma, Q. Liu, J. Cao, Y. Tian, A. Al-Dhelaan, M. Al-Rodhaan, LGIEM: Global and local node influence based community detection, Future Generation Computer Systems 105(2020) 533-546.

[14] Y. Mao, Z. Peng, Z. Chen, D. Liu, Landslide hazard assessment based on uncertain decision tree classification method, Application Research of Computers 31(12)(2014) 3646-3650.

[15] L. Tran, L. Duckstein, Comparison of fuzzy numbers using a fuzzy distance measure, Fuzzy sets and Systems 130(3)(2002) 331-341.

[16] F. Wang, G.-Y. Wang, Z.-X. Li, S.-Y. Peng, Clustering by fast search and find of density peaks based on grid, Journal of Chinese Computer Systems 38(5)(2017) 1034-1038.

[17] C. Xing, X. Wang, Data Stream Clustering Algorithm Based on Extensible Grid and Density, Computer Engineering (12)(2014) 36.

[18] L. Jiang, M. Zhang, J. Zheng, Optimization of clustering by fast search and find of density peaks, Application Research of Computers 33(11)(2016) 3251-3254.

[19] Z. Li, K. Luo, Research on adaptive parameters determination in DBSCAN algorithm, Computer Engineering and Applications 52(3)(2016) 70-73+80.

[20] H. Sun, M.-X. Zhang, J. Dai, Z.-W. Shang, Optimization of grid based clustering by fast search and find of density peaks, Computer Engineering and Science 39(5)(2017) 964-970.

[21] D. Bubboloni, M. Gori, Symmetric majority rules, Mathematical Social Sciences 76(2015) 73-86.

[22] L. Lyu, Y. Wei, M. Ren, X. Pan, Agglomerative hierarchical clustering based on ant colony optimization algorithm, Application Research of Computers 34(1)(2017) 114-117.

[23] S. Shimizu, S. Shimada, K. Tsuboki, Assimilation Impact of Different GPS Analysis Methods on Precipitation Forecast: A Heavy Rainfall Case Study of Kani City, Gifu Prefecture on July 15, 2010, Journal of Disaster Research 12(5)(2017) 944-955.

[24] Q. Su, Research on Loess landslide hazard zonation based on DEM, [Doctoral dissertation] China University of Geosciences (Beijing), 2006.

[25] F. Guzzetti, A. Carrara, M. Cardinali, P. Reichenbach, Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy, Geomorphology 31(1-4)(1999) 181-216.

[26] F. Guzzetti, A. Carrara, M. Cardinali, P. Reichenbach, Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy, Geomorphology 31(1-4)(1999) 181-216.

[27] D. Bubboloni, M. Gori, Symmetric majority rules, Mathematical Social Sciences 76(2015) 73-86.

[28] W.-N. Xu, P.-X. Wang, P. Han, T.-L. Yan, S.Y. Zhang, Application of Kappa coefficient to accuracy assessments of drought forecasting model: a case study of Guanzhong Plain, Journal of Natural Disasters 20(6)(2011) 81-86.

[29] H.X. Gao, K.L. Yin, Discuss on the correlations between landslides and rainfall and threshold for landslide early-warning and prediction, Yantu Lixue(Rock and Soil Mechanics) 28(5)(2007) 1055-1060.