中華水土保持學報, 49(3): 187-198 (2018)
Journal of Chinese Soil and Water Conservation, 49 (3)*: 187-198* (2018)
DOI: 10.29417/JCSWC.201809_49(3).0006

A Study of Landslide Image Classification through Data Clustering using Bacterial Foraging Optimization

Shiuan Wan[1*] Shih-Hsun Chang[1] Tein-Yin Chou[2] Chen Ming Shien[2]

Abstract Generation landslide susceptibility maps are to study the relations among the image data of band variables concerning the occurrence/nonoccurrence on investigated samples of landslide. A feasible solution on generating landslide susceptibility map through land cover classification of remote sensing data is an important topic for studies of image processing and classification. Image classification considering clustering technique is well-accepted when ground truth data is scarce. However, applying the clustering technique, the initial guess of cluster centers may lead to different results. As a result, the traditional K-means data clustering technique may fail to arrange the data to the appropriate target groups. Accordingly, this study employs bacterial foraging algorithm (BFA), which successfully resolves the image data of clustering problems in landslide. On the other hand, the constrained clustering is a useful clustering technique to improve the classification outcomes when few label data are available. Accordingly, the study focused on the classifier by using BFA optimized constrained clustering to study landslide area in which the evaluation of landslide occurrence by remote sensing image data is rationally studied. The results show constrained BFA clustering yields the better classification results (93.7%) than those of BFA clustering (81%) and K-means (77.6%).
Keywords: Landslide, image classification, bacterial foraging algorithm

Introduction

Landslides generally cause the loss of human lives and their properties. Comparing to diverse geologic, topographic and climatic conditions, landslides are frequent phenomena in Taiwan. Conventionally, monitoring of landslides for their locations and distributions are generally used in-situ or field geotechnical techniques through aerial photos by human-power or unmanned aerial devices. These attempts with profound techniques can lead to evaluate landslides locations become effectively due to their spatially remote distributions. To sum up, generating the landslide susceptibility map manually results in a lot of time spending and human power. As a matter fact, it may not be completed in time. Thus, a more effective solution to estimate landslide area is desired through considering the remote sensing data. Excessive researches study on landslide various evaluation/estimation through Geographic Information System (Wan et al. 2012) with different techniques. Generating a landslide susceptibility map is a very crucial solution for landslide hazard management and analysis for various useful steps and scope by analyzing the landslide with its conditioning factors. Most of the past studies used the statistical analysis. Alternatively, the data mining based models are widely accepted for creating the landslide susceptibility map. For instance, researchers have used fuzzy theory (Pradhan 2011; Akgun 2012), artificial neural networks (Fan et al. 2012; Pradhan and Lee 2010a, b, c; Yilmaz 2009b, 2010a, b), decision tree or support vector machine (Wan et al. 2009) to aid in landslide modeling. These studies obtain satisfactory results, however, most of them are through supervised classification, which requires labeled data in order to build the prediction system. Due to the difficulties of landslide field survey, it is therefore anticipated to create a landslide classifier with a limited number of landslide data points.

Among the well-known data classification techniques, the unsupervised clustering method requires no prior knowledge of the class label of sample data and is adequate for grouping similar land covers (Wan 2013). That is, clustering approaches are unsupervised process as well as a common technique utilized in many fields in Geosciences. Its application includes machine learning (Fan et al. 2012), pattern recognition (Bishop 2006), data mining (Jain 2010), image analysis (Chaang and Wan 2015) and bioinformatics (Higham et al. 2007). Clustering is a very useful method to understand Geosciences data to the physical behavior without in-situ investigation. Among the optimization methods for enhancing the cluster analysis, there are quite a few popular optimization methods. For example, genetic algorithm (GA) has been explored extensively (Kirschbaum et al. 2009) and Particle swarm optimization (PSO) is an evolutionary optimization technique. PSO with k-means is also applied to data clustering optimization (Cura 2012, Dasgupta et al. 2009). However, those aforementioned algorithms suffered from following shortcomings. First, they converge speed are very slow. On the other hand, it sometimes converges to local minimal points, which lead to a wrong solution (Trelea 2003). As part of our study, it is decided to use the Bacterial Foraging Algorithm (BFA) which is an optimization algorithm simulated by the foraging behavior of bacteria. In essence, considering the weaker foraging ability of bacteria that is more possible to be wipe out than those with stronger foraging ones. The computer program simulates the behavior of generating new bacteria in each of the calculation iterations. Thus, if the program is assigned to run many iterations, it can represent the weaker foraging ability bacteria are either vanished or changed to new nutrient states. This action can be seen as such as E.Coli which lives in the human intestine (Kim et

〔1〕Information Networking and System Administration, Ling Tung University, Taiwan
〔2〕GIS center Feng Chia University.
* Corresponding Author. E-mail: shiuan123@teamail.ltu.edu.tw

al. 2007). To sum up, BFA presents a novel search technique which can solve the practical optimization problem in different fields. Considering the ability to avoid the process of searching function falls into in local minimum for optimizing process, the BFA perform s better than GA, SA, and PSO (Das et al. 2009).

In general, the additional information (or ancillary information) are employed to the domain knowledge. Considering various types of cluster techniques in each research are basically searching an appropriate solution in aggregating data into unknown groups. For example, through field survey over the landslide region, we have the ground truth data of a few samples. Field survey often costs a lot of human effort and research funds. Consequently, acquiring fewer samples is preferable to financial consideration. It turns out that we opt to make good use of these labeled objects to improve clustering performance. Constrained clustering, one kind of semi-supervised classification, grew out of the need to find ways to accommodate this information when it is available (Basu 2008 et al.; Wagstaff et al. 2008). There are details for further information about constrained clustering (Wagsta et al. 2001; Bilenko et al. 2004; Davidson and Ravi 2007).

In this study, a parallel study using K-means clustering and BFA clustering is performed in classifying the remote sensing data. Furthermore, BFA clustering is enhanced by constrained clustering algorithm, which a few labeled examples are added in the clustering process. The outcomes are compared and discussed. The rest of the papers are arranged as follows. Section 2 presents the study area and study plan in question. Section 3 reviews the K-means, BFA and constrained BFA clustering methods. A simplified example using *UCI Iris data set is used to explain the evolutionary process in BFA. The clustering results re presented and discussed in Section 4, followed by the Section 5, the conclusion.

*Note :UCI is a standard database for computer science researcher to test their model is good or not. IRIS is a very popular for the best raw data)

Study Area and Study Plan

1. Study Area

To perform the clustering technique, the study area is selected of downstream on the Wushe reservoir, Nantou County, Taiwan (E:121.182, N:23.945). The slate and shale with dark gray sandstone is the basic geological condition which belongs to sub-tertiary metamorphic region. The study region has a very fragile material of soil as well as the slope is steep. (Wan et al. 2014). Figure 1 shows the remote sensing image of site location and the shape of a reservoir. This landslide area is 4.3 ha with an average slope of 30 degrees (Wan et al. 2015). The average elevation area about 1000 meters and the slide area has the landslide mass. The slide moves along a roughly planar surface as well as it has little rotation or backward tilting. Accordingly, it is recognized that it is the translation type landslide (Cruden and Varnes 1996). Since the slope is very steep (26.7 degrees with standard deviation of 4.3 degrees), the area is very fragile while earthquake or huge rainfall struck for inducing landslide. The study ignored some of the bare-land or rock areas along the river have which has the same image property such as landslide area. As aforementioned, the typhoon usually brings heavy precipitations, and this triggers landslides in this area. Two violent typhoons occurred in the landslide area during March 1, 2006 to Oct. 1, 2008. They produced a great amount of rainfall (Wan et al. 2014). The cumulative precipitation depth of the Sinlaku typhoon is 976mm, and that of Jangmi typhoon is 451mm (Wan et al. 2014). The worst event happens while the movement of a landslide may cut off Wan Da stream, and it will suddenly destroy the Wan Da hydraulic power plant. Many built protection constructions (such as rock anchor) and monitoring equipment (such as the monitoring wells of ground-water levels) are made by Taiwan Power Company (TPC) to prevent and supervise this landslide area (Wan et al. 2014).

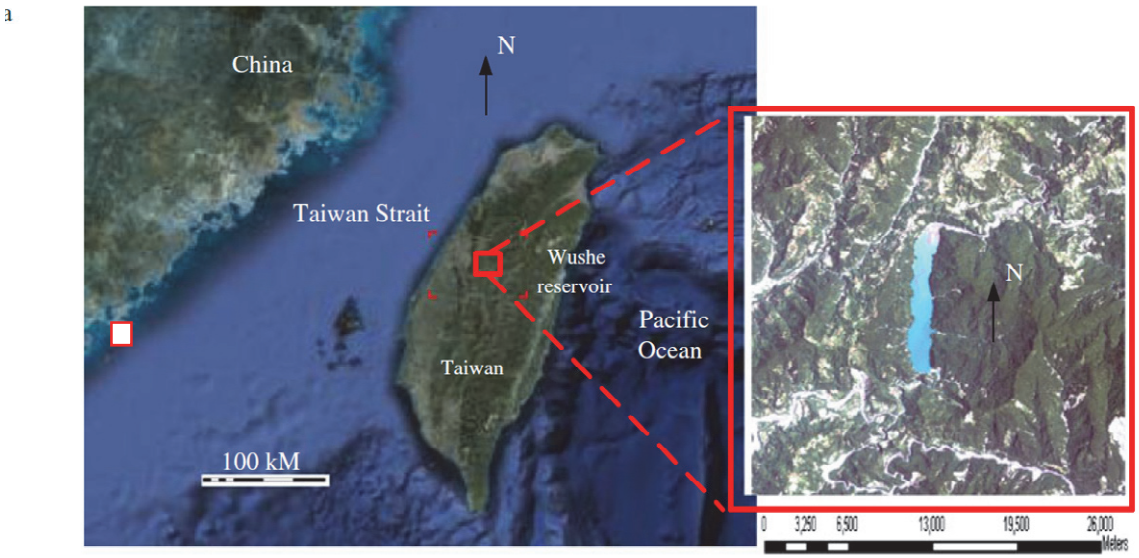

Fig.1 The location of the Wushe reservior in Taiwan

2. Data Preparation

As part of this study, the field survey was prepared to identify the landslide and non-landslide samples to be used in building the clustering model. These samples (36 non-land slide and 22 landslide data points) and their locations are depicted in Figure 2. Besides the fundamental spectral bands of data (R, G, B and NIR), a Digital Terrain Model (DEM) data for this area was also considered to retrieve the elevation attribute (Chang and Wan 2015). The Normalized Difference Vegetation Index is usually considered as the behavior of green vegetation of surface ground by using the red and near-infrared spectral reflectance measurements of:

$$NDVI = \frac{NIR - R}{NIR + R} \quad (1)$$

Hence, each sample pixel has six attributes (R, G, B, NIR, NDVI, DTM) are employed for the training model. To reduce the effect of quantity scale of variables on classification performance, all input attributes are normalized to the range [0,1] before training. Another region in the map was selected to serve as test area in order to examine the performance of classifier. Figure 3 shows the test area used in this study.

Fig.2 The field survey samples selected (blue circle: non-lanslide, purple cross: landslide)

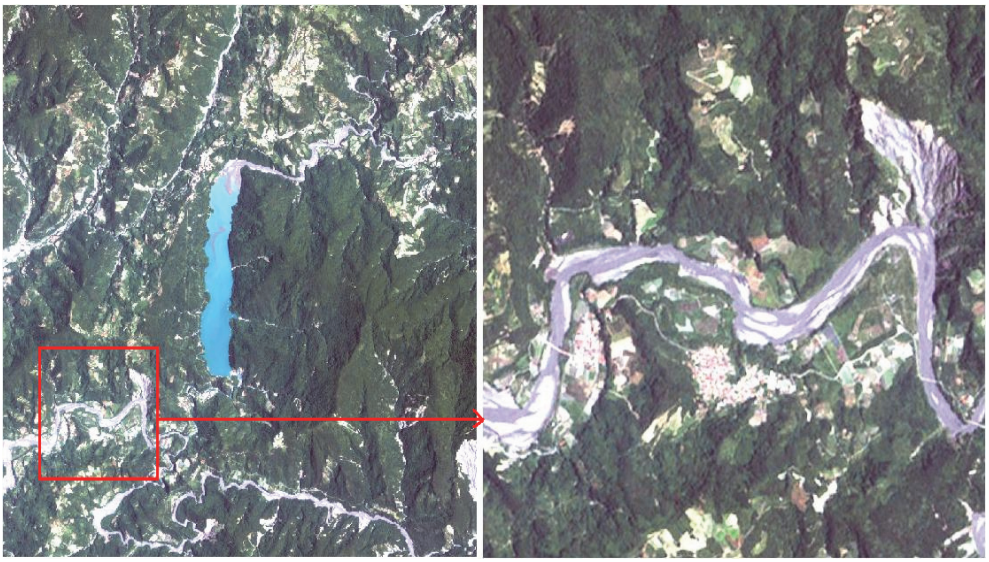

Fig.3 The test area selected

3. The Plan of Study

Landslide analysis of occurrences generally relies on a given inventory map. Unfortunately, there is no existing landslide. In this research, the study area is the Wan-Da reservoir, and the data collection time is after the occurrence of the Chi-Chi earthquake in 1999 (Wan et al. 2014). Nearby the study area, there is Wan-Da reservoir which is located at a 20-40 km away from the Chelung Pu fault. The fault may transmit tremendous energy to the central part of Taiwan such as Chi-Chi earthquake took place. Using digital elevation models (DEMs), geological maps and SPOT-image data, the researchers investigated maps and GIS spatial attribute data (morphology, geology, landslide, slope, and soil type). Those data were preprocessed by using a spatial information system to build a database of landslide events. Figure 4 presents the steps for this study.

The study breaks into two parts. In the first part, the K-means and BFA clustering algorithms are used to build the classifier with training samples as inputs. The built classifier is then used to classify the test area, thus generating a thematic map of a landslide. In the second part, an instance-level constrained clustering approach based on BFA is performed as a parallel study. The outcomes of the above three approaches are compared. The K-means clustering algorithm is directly from Matlab's built-in function, whereas BFA and constrained clustering are home built with Matlab.

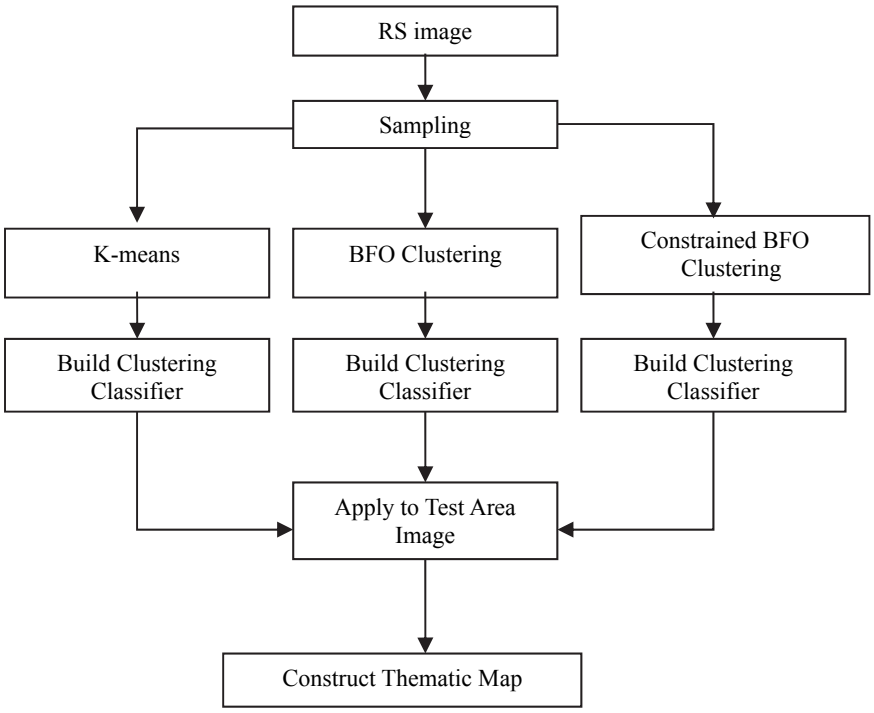

Fig.4 Research steps illustrated

Research Method

1. K-means algorithm

K-means (Macqueen 1967) employs an iterative clustering algorithm to approach the best groups of aggregated data. The concept of K-means is applying data items that are moved among set of group data until the optimal desired set is reached. Please see Figure 5(a). The cluster center is defined as the average value of attributes for all data points in each cluster. While implementing K-means, the first step is to assign the number of clusters and the initial value for each cluster center. Secondly, assign each data point to a cluster by the rule that the data point is closer to its grouped cluster center than to those of the other clusters. Please see Figure 5(b). Finally, calculate the new mean values of all data points in each cluster to obtain new cluster centers. The process will repeat until the convergence criteria are satisfied. However, the object function for clustering is generally non-convex and non-linear. Therefore the traditional approaches, especially a standard K-means algorithm, are sensitive to data initializations and easy to be trapped in a local optimal solution. The effectiveness of K-means is usually determined by the initial positions of the cluster centers, therefore it tends to have more trials on cluster center selection. On the other hand, the new approach proposed in this study, BFA clustering, is to obtain a better solution by searching the best cluster center and avoiding the searching process being trapped in a local minimum. Please see Figure 5(c) for illustration. Figure 5(c) is a good explanation for how to avoid solution being trapped in a local minimum. Applying improper cluster center will produce to aggregate the wrong scope of data to the cluster.

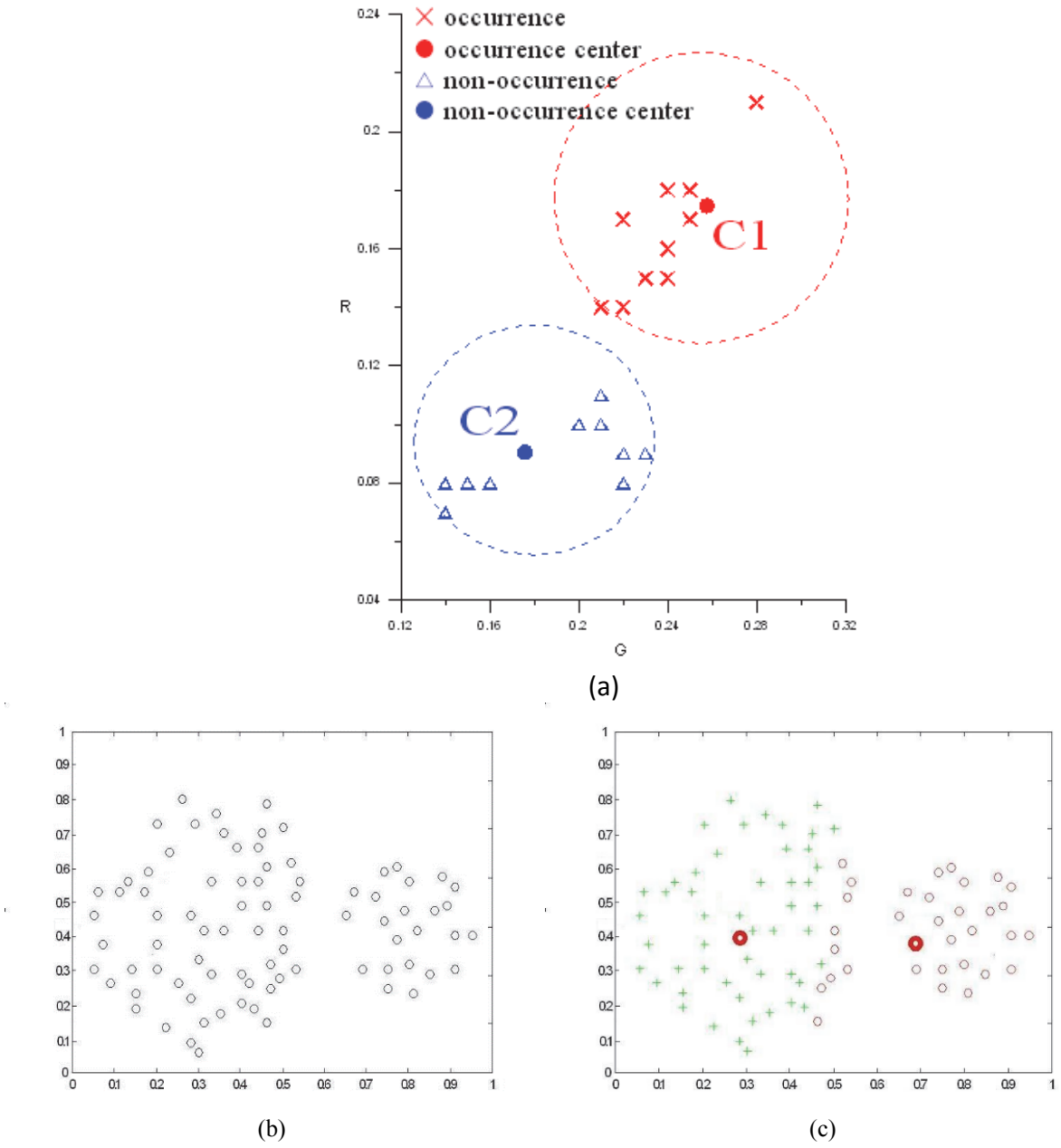

Fig.5 Concept of Clustering (a) cluster center (b) distribution of data (c)improper cluster center illustration

2. Bacterial foraging algorithm (BFA)

BFA is an algorithm simulated by the foraging behavior of E. Coli bacteria. The biological concept of the bacterial foraging strategies and their motile behavior and their decision-making solutions are studied and applied in different fields. More specifically, BFA is designed to solve non-gradient optimization problems and to handle those complex objective functions. Each bacterium carries a solution, guided by its foraging behavior to move in solution space by optimizing the object function defined in the later paragraph. It searches the hyperspace of data distribution to perform three main operations, namely chemotaxis, reproduction and elimination-dispersal. Chemotactic, reproduction, swarming, and elimination-dispersal are the four major steps of bacterium foraging process. The chemo-taxis process is simulated as the bacteria of tumbling and swimming. The bacteria spend the lifetime alternating between the above two modes of motion. In BFA, $\phi(i)$ is represented by a unit length in a random direction, which presents a movement with a given random direction after a tumble. In addition, the size of the steps is represented by the constant run-length unit, $C(i)$. For a population of bacteria, the location of the i^{th} bacterium at the j^{th} chemo-tactic step, k^{th} reproduction step and the l^{th} elimination-dispersal event is represented by *X(j,k,l)*. After a tumble, the location of the i^{th} bacterium is represented as

$$X_i(j+1,k,l) = X_i(j,k,l) + C(i) \times \phi(i) \tag{2}$$

where

$$\phi(i) = \frac{X_i(j,k,l) - X_{rand}(j,k,l)}{X_i(j,k,l) - X_{rand}(j,k,l)} \tag{3}$$

$$C(i) = rand() \times step \tag{4}$$

Where rand( ) is a symbol for computer program to generate random number.

The object function is constructed and described in the following procedure. Given a dataset of N objects $\{x_1, x_2, ..., x_N\}$ in $\Re^n$-dimensional space, and the goal is to partition the dataset into K clusters. One popular approach to measure the effectiveness of clustering is through within-class compactness J_c, of which its mathematical formulation can be described as

$$J_c = \sum_{j=1}^{K}\sum_{i=1}^{N}\sum_{p=1}^{n} w_{ij} \parallel x_{ip} - c_{jp} \parallel^2 \tag{5}$$

where x_{ip} is a value of *p*th attribute of *i*th sample; *c* a cluster center matrix of size $K \times n$; c_{jp} an average of the pth attribute values of all samples in the cluster *j*; w a weight matrix of size $N \times K$; w_{ij} an associated weight of object x_i with cluster *j* which can be assigned as

$$w_{ij} = \begin{cases} 1, \text{if object i is contained in cluster j} \\ 0, \text{otherwise} \end{cases}$$

$i = 1,2,.....N; j = 1,2,....K$

and

$$c_{jp} = \frac{\sum_{i=1}^{N} w_{ij} x_{ip}}{\sum_{i=1}^{N} w_{ij}} \quad j = 1,2,.....,K; p = 1,2,....n$$

The smaller within-class compactness (J_c) is, the better the cluster result. The fitness function of each bacteria is defined as

$$\text{fitness function } f = \frac{1}{J_c} \tag{6}$$

A reproduction is carried out after taking a maximum number of chemotactic execution, *Nc*. The population will be change to half of the number due to the least healthy half dies. Every bacterium in the other healthiest one splits into two bacteria that take the same position.

$$S_r = \frac{S}{2} \tag{7}$$

S is the total number of the bacterium in next generation; S_r is the original numbers of the bacterium.

After N_{re} reproduction steps, an elimination-dispersal event takes place for N_{ed} number of executions. In this operation, each bacterium could be moved to explore other parts of the search space. The probability for each bacterium to experience the elimination-dispersal event is determined by a predefined fraction p_{ed}.

To help further understand the BFA clustering process, an example which includes part of UCI Iris dataset is used to explain the BFA clustering algorithm in details. The selected Iris dataset has 12 instances, shown in Table 1. The first 6 samples are selected as Iris-versicolor class and the rest of the 6 samples are selected as Iris-virginica class. Since it is a clustering process, it randomly assigns to two labels, 1 and 2, to these 12 instances, as shown in Table 2. The number of bacteria is set to 3, which indicates 3 potential solutions are present. The maximum number of chemotactic execution times, *Nc,* is set to 2, with a maximum number of chemotactic steps, *Ns* set to 2. The reproduction steps, $N_{re,}$ is set to 2. The number of executions for elimination-dispersal events, N_{ed}, is set to 3, and the probability, p_{ed}, is 0.3. Marching *step* is set to 0.1. The initial fitness of 3 bacteria is set to zero. Each bacterium carries two cluster center locations, which jointly depicts the bacterium current location, or the solution in question. Our goal is to guide the bacterium to move to a best foraging location where the fitness function is maximum. The initial cluster centers, C_1 and C_2, are obtained through the initial random label assignment to instances. For the first bacterium, its cluster centers evolutionary movement is

C_1= (6.350, 3.050, 5.300, 1.875), C_2=(6.525, 2.887, 5.125, 1.700)

Table 1 Dataset used to explain BFA clustering process, extracted from UCI Iris dataset (source from https://archive.ics.uci.edu/ml/datasets/Iris)

No.	sepal length	sepal width	petal length	petal width	class
1	7.0	3.2	4.7	1.4	Iris-versicolor
2	6.4	3.2	4.5	1.5	Iris-versicolor
3	6.9	3.1	4.9	1.5	Iris-versicolor
4	5.5	2.3	4.0	1.3	Iris-versicolor
5	6.5	2.8	4.6	1.5	Iris-versicolor
6	5.7	2.8	4.5	1.3	Iris-versicolor
7	6.3	3.3	6.0	2.5	Iris-virginica
8	5.8	2.7	5.1	1.9	Iris-virginica
9	7.1	3.0	5.9	2.1	Iris-virginica
10	6.3	2.9	5.6	1.8	Iris-virginica
11	6.5	3.0	5.8	2.2	Iris-virginica
12	7.6	3.0	6.6	2.1	Iris-virginica

Table 2 The initial class label assigned to data instances

No.	1	2	3	4	5	6	7	8	9	10	11	12
class	2	2	1	2	2	1	1	2	2	2	1	2

The random unit direction rand()* $\phi(i)$ from Eq.(4) is (0.4148, -0.1691, 0.8325, -0.3261) and *step* = 0.1. Therefore after the chemotactic movement, C_1 = (6.391, 3.033, 5.383, 1.842), C_2 =(6.566, 2.871, 5.208, 1.667), where the first attribute in C_1 is calculated according to Eq.(2) as 6.391 = 6.350 + 0.1 * 0.4148. All other attributes in C_1 and C_2 are obtained in the same manner. After the new cluster centers are obtained, the data points are reassigned possibly to new cluster by a shorter Eulerian distance between themselves and cluster centers. Then, after new cluster centers are relocated as possible cluster assignment of data points. The fitness of a bacterium for the current clustering stage is then obtained by Eq.(6). Another chemotactic movement will be executed, and the same calculation follows. In the first two rounds (iterations) of bacteria, evolutionary process is depicted as Table 3. A half of bacteria with better fitness will be preserved in the reproduction process. Thereafter the elimination-dispersal events take place and a part of bacteria are discarded. New bacteria will be created randomly in the solution space with the goal to avoid being trapped in local minimum. The whole iteration process continues until the termination criteria is met, as illustrated in Figure 6.

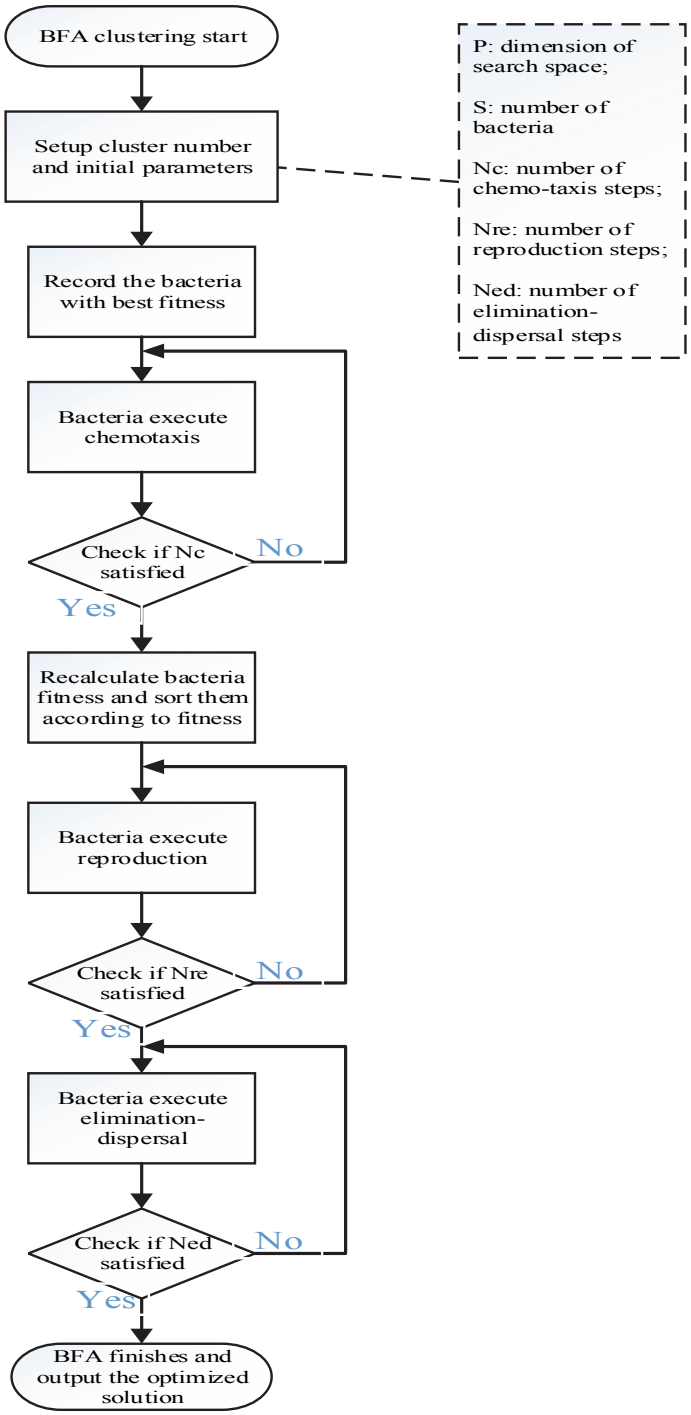

Fig.6 Computational steps for BFA clustering

Table 3 The cluster center evolutionary process

	Cluster Center of Bacterium #1	Cluster Center of Bacterium #2	Location change (rand* $\phi(i)$ *step)	Fitness
Initial cluster centers	(6.350, 3.050, 5.300, 1.875)	(6.525, 2.887, 5.125, 1.700)	(0.04148, -0.01691, 0.08325, -0.03261)	NA
After the first chemotactic movement	(6.391, 3.033, 5.383, 1.842)	(6.566, 2.871, 5.208, 1.667)	NA	NA
Calculate fitness and relocate the cluster centers	(6.600,2.983,5.833,2.100)	(6.333,2.900,4.533,1.417)	NA	0.12542
Before the second chemotactic movement	(6.600,2.983,5.833,2.100)	(6.333,2.900,4.533,1.417)	(-0.02332,-0.01958,0.08778, 0.03697)	NA
After the chemotactic movement	(6.577,2.964,5.921,2.137)	(6.310,2.880,4.621,1.454)	NA	NA
Calculate fitness and relocate the cluster centers	(6.760,3.040,5.980,2.140)	(6.257,2.871,4.614,1.486)	NA	0.12809

3. Constrained BFA clustering

Clustering is a crucial technique for data mining because it can recognize major patterns or trends without supervisory information such as data labels. It is usually defined as a process of moving a set of objects into clusters, each of which represents a meaningful group. However, there are cases when additional information or domain knowledge is available about the types of a cluster in the dataset. This supplemental information may comprise class labels for a subset of objects, a true similarity between pairs of objects, or user preferences about how objects should be grouped. Constrained clustering is a kind of clustering method when some of the aforementioned additional information is available, and it incorporates this information into its clustering process to enhance the overall performance. In this study, some samples are selected in prior because field survey work was performed on them. Therefore their labels are known and should be classified into the clusters of associated classes. We state the instance-level constrained BFA clustering algorithm as follows.

Assume a given data set X, number of clusters k, must-link constrains $C_{=} \subset X \times X$, cannot link constraints $C_{\neq} \subset X \times X$.

1. Let $\mu_1 ... \mu_k$ be the k initial cluster centers.
2. Search the attribute hyper space for a better cluster center set $\{ \mu_1 ... \mu_k \}$ through BFA, which optimizes the fitness function given.
3. For each instance $x_i \in X$, assign it to the closest center c such that VIOLATE-CONSTRAINTS ($x_i, c, C_{=}, C_{\neq}$) is false.
4. Update each cluster center μ_i by averaging all of the instances x_i that have been assigned to it.
5. Iterate between step 2 and 4 until convergence is reached.
6. Return $\{ \mu_1 ... \mu_k \}$

VIOLATE-CONSTRAINTS (instance x_i, cluster c, must-link constraints $C_{=}$, cannot link constraints $C_{\neq}$)

1. For each $c_{=}(i, j) \in C_{=}$: If $x_j \notin c$, return true.
2. For each $c_{\neq}(i, j) \in C_{\neq}$: If $x_j \in c$, return true.
3. Otherwise, return false.

4. Developing computer program for BFA clustering

A computer program is developed to carry out the computation of BFA clustering, as illustrated in Figure 6. The first step is to set up the cluster numbers and the initial conditions. It includes the dimension of search space, the number of bacteria, the number of chemotaxis steps, the length of swim, and the number of reproduction steps. The second step is to execute the chemotaxis process iteratively with the number of chemotaxis steps, N_c. Similarly, the third step is to execute the reproduction and the fourth step is elimination-dispersal iteratively with their associated number of steps, N_{re} and N_{ed} respectively. The program terminates when the maximum number of iteration is reached or the fitness does not change over a number of iterations.

Results and Discussions

1. Comparing constrain clustering of BFA vs. traditional BFA

In this study, three different parallel clustering approaches are designed and applied. Figure 4 illustrates our approaches. In this study, the clustering technique was employed to generate a susceptibility map of a selected test area based on 58 site survey samples. The remote sensing image consists of four attributes (R, G, B, NIR) and two ancillary attributes (NDVI, DTM), which are obtained by Eq.(1) and DTM map, respectively. The decision attribute is binary, landslide or non-landslide, acquired by site survey. A small part of the sample data is presented in Table 4.

Table 4 A part of data for constructing the BFA landslide image clustering model

ID	R	G	B	NIR	DTM(m)	class
1	57	55	70	79	1255	landslide
2	82	74	81	86	1326	landslide
3	36	40	58	52	1295	landslide
4	56	54	72	38	1581	landslide
5	25	35	51	101	1154	non-landslide
6	29	39	51	100	1084	non-landslide
7	25	34	51	92	1335	non-landslide
8	37	49	56	181	1520	non-landslide

The study employs the data samples (see Table 4) to construct the model through BFA clustering using the concept depicted in Figure 5. When the cluster centers are determined by the model, the rest of the imagery data points can be input into the model and the landslide susceptibility map can be generated accordingly. The distance of the test data between two cluster centers (non-landslide vs landslide) determines the pixel on the image is of non-landslide or landslide (red color). Three different clustering approaches are studied and compared. The first one is the traditional K-means clustering algorithm, and its predicted landslide susceptibility map is depicted in Figure 7. It is seen from the figure that it predicts most of the river as landslides, which is not a satisfying result. The second approach is BFA clustering, and its predicted landslide susceptibility map is shown in Figure 8. Comparing Figure 7 with Figure 8, the result of BFA clustering is somewhat better than that of K-means. There is less misclassification on predicting the riverside to landslides. This is attributed to the BFA heuristic optimization that generally finds a better solution in a limited time. However, a great amount of misclassification is still present in the river area. Lastly, the constrained BFA clustering is employed and its result is given in Figure 9. It is observed from the figure that there is much less misclassification. The riverside can be distinguished with landslide clearly. Only a small portion of the river bank is predicted as a landslide, and the overall susceptibility map is in more agreement with the actual site survey result. The error matrices of the predicting approaches, K-means, BFA and constrained BFA, are given in Table 5, 6 and 7, respectively. The K-means renders a 77.6% overall accuracy, but a poor user accuracy 63.0%, indicating its predicting ability is limited in classifying landslide land covers. On the other hand, BFA clustering improves a small amount on overall accuracy, but its user accuracy on predicting landslide is only 66.7%, marginally improved over the K-means. Constrained BFA clustering gives the best result, rendering an overall accuracy of

93.1% and a much improving user accuracy, 86.4%. This indicates that, at least in our case, the constraint algorithm imposed on BFA clustering produces a significant benefit in building the classifier; while employing BFA clustering alone still outperforms K-means marginally.

In this study, the threshold method by Liu (2010) was adopted. The digital elevation model (DEM) was used in the study area of satellite imagery at Wushi Reservoir to extract slope values for every 15 points in the collapsed and non-collapsed lands and recorded them as well. Thirty data points were applied to the slope gradient threshold was 24.9^0 which is found a dominant factor to distinguish the landslide and riverbed. Therefore, in this study, it depicts that there is a threshold slope, roughly 24.9^0, which separates landslide and non-landslide classes. Figure 10 is plotted based on the threshold value (24.9^0) to improve the thematic map on constrained BFO.

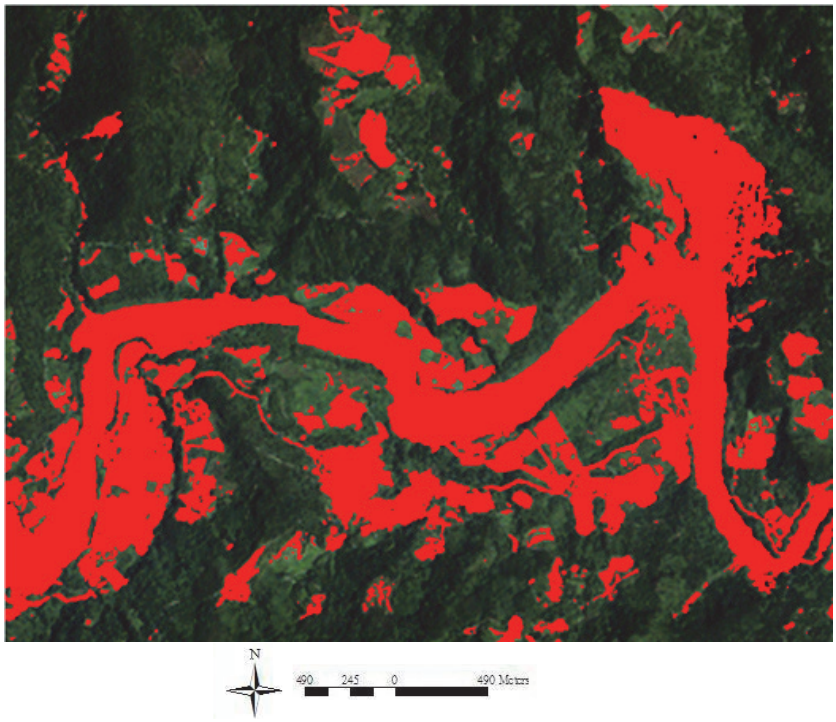

Fig.7 Thematic map generated by K-means classifier

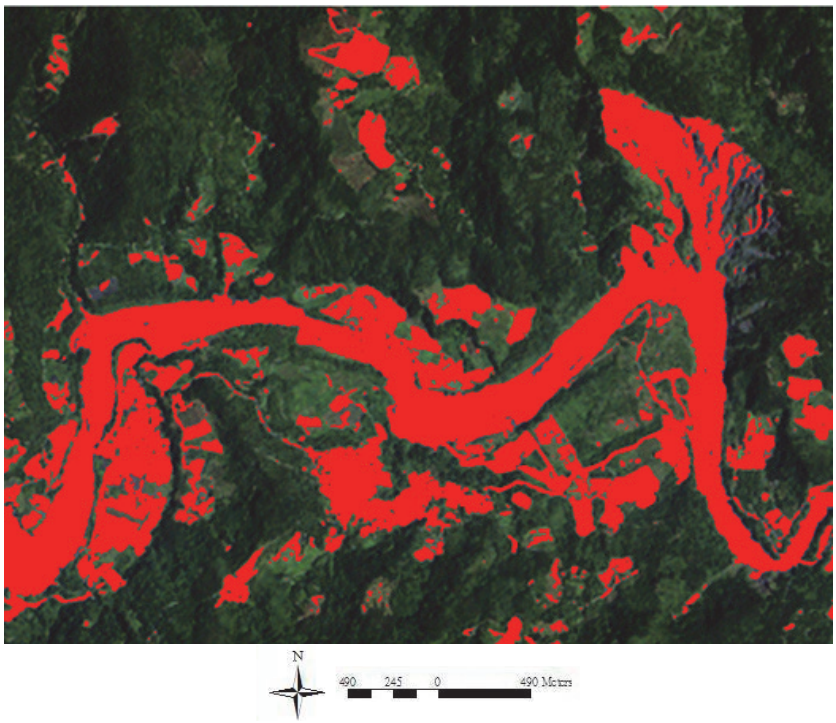

Fig.8 Thematic map generated by BFO classifier

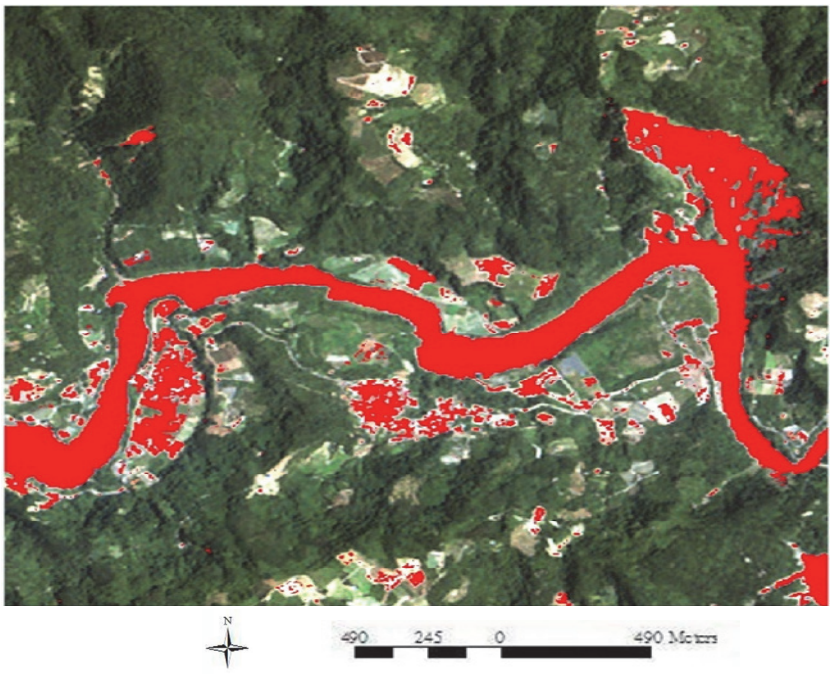

Fig.9 Thematic map generated by Constrained BFO classifier

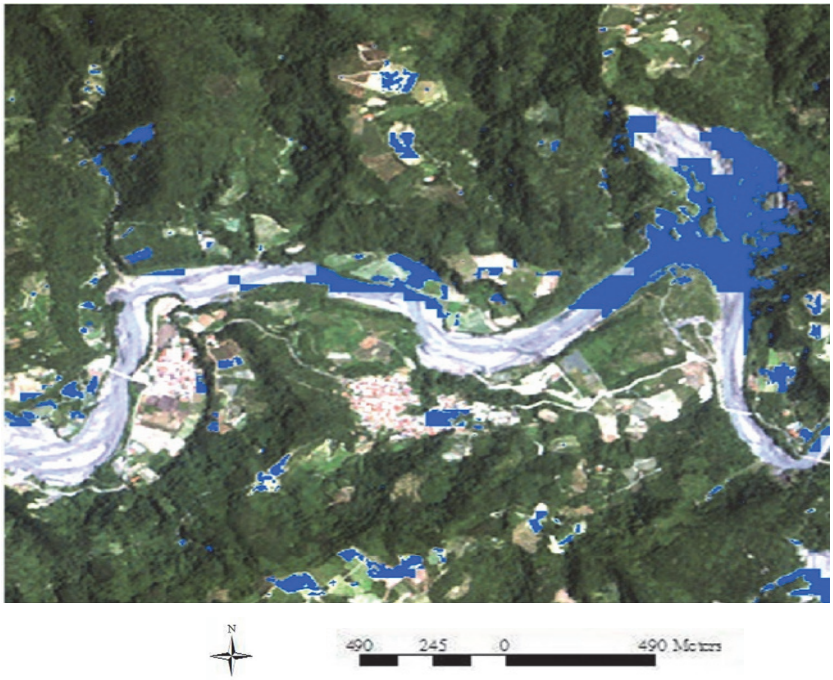

Fig.10 Constrain BFO with threshold method

Table 5 Error matrix of K-means clustering case

	Kappa：0.54	Ground Truth non-landslide	landslide	sum	User accuracy
Predicted class	non-landslide	28	3	31	90.3%
	landslide	10	17	27	63.0%
	sum	38	20	58	
	Producer accuracy	73.7%	85.0%		Overall accuracy = 77.6%

Table 6 Error matrix of BFA clustering case

	Kappa：0.61	Ground Truth non-landslide	Ground Truth landslide	Ground Truth sum	User accuracy
Predicted class	non-landslide	29	2	31	93.6%
	landslide	9	18	27	66.7%
	sum	38	20	58	
	Producer accuracy	76.3%	90.0%		Overall accuracy = 81.0%

Table 7 Error matrix of constrained BFA clustering case

	Kappa：0.85	Ground Truth non-landslide	Ground Truth landslide	Ground Truth sum	User accuracy
Predicted class	non-landslide	35	1	36	97.2%
	landslide	3	19	22	86.4%
	sum	38	20	58	
	Producer accuracy	92.1%	95.0%		Overall accuracy = 93.1%

The advantage on using constrained clustering

Clustering is a well-known approach on automatic unsupervised data analysis. Although the given a collection of data instances, the clustering task can find a solution on grouping of the physical data, aggregate them into sets (the clusters) such that the similarity among each group can be studied and analyzed.

Considering a better approach, a new solution of semi-supervised clustering algorithms, called Constrained Clustering, has emerged into the original clustering algorithm in this study. The new algorithm employs some given domain knowledge which allows scientists or researchers to guide the clustering process efficiently. This information could be in the shape of a set of pairwise relations among data elements (called constraints). It presents the superior behavior on the two data samples linked to each of these constraints should/should not be put in the same group. Consequently, under the Constrained Clustering approach, the prior knowledge that was unused in traditional clustering algorithms is used to enhance the way to group different distribution of data. It makes the final clustering approach more accurate and meaningful which render more visualized on the range of data.

Conclusion

Remote sensing image classification is one of the important methods that have been studied and proven to be effective. In general, the landslide usually occurs in remote places where it is very difficult to reach for in-situ investigation. This study selects only a few samples to generate the susceptibility map with an innovative classification approach for landslide investigation. In this study, we employed a few in-situ survey data incorporated with respect to the corresponding image data to study the landslide area characteristics. Among data mining approaches of image classification, constrained clustering is a class of semi-supervised which is typically considering a better solution for Data clustering algorithm than the traditional clustering outcomes. Consequently, the study is decided to use three unsupervised parallel approaches to understand differences between the generated susceptibility maps. The outcomes are compared and discussed. The results of BFA clustering and K-means are not good enough and much misclassification is observed. The accuracy rate of BFA is about 81%. Also, there are many misjudge place on the thematic map such as the bedrock area. The best one is that rendered by constrained BFA clustering. It indicates that considering limited site survey samples with constrained BFA clustering approaches yields an overall accuracy of 93.1%.

Acknowledgment

National Science Council, Taiwan, (Research Project 104-2119-M-275-002- and 105-2119-M-275-001-) sponsored this work.

References

[1] Akgun, A. (2012). “An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm.” *Computers and Geosciences*, 38(1), 23-34.

[2] Basu, S., Banerjee, A., and Mooney, R.J. (2004). “Active semi-supervision for pairwise constrained clustering.” *Proceedings of the 2004 SIAM International Conference on Data Mining*, 333-344.

[3] Bilenko, M., Basu, S., and Mooney, R.J. (2004). “Integrating constraints and metric learning in semi-supervised clustering.” *Proceedings of the 21st International Conference on Machine Learning*, 81-88.

[4] Bishop, C.M. (2006). “Pattern Recognition and Machine Learning.” *Pattern Recognition*, 4, 738.

[5] Chang, S.H., and Wan, S. (2015). “A novel study on ant-based clustering for paddy rice image classification.” *Arabian Journal of Geosciences*, 8(8), 6305-6316

[6] Cura, T. (2012). “A particle swarm optimization approach to clustering.” *Expert Systems with Applications*, 39, 1582-1588.

[7] Das, S., Biswas, A., Dasgupta, S., and Abraham, A. (2009). "Bacterial Foraging Optimization Algorithm: Theoretical Foundations, Analysis, and Applications." *Foundations of Computational Intelligence*, 3, 23-55.

[8] Dasgupta, S., Biswas, A., Das, S., Panigrahi, B.K., and Abraham, A. (2009). "A micro-bacterial foraging algorithm for high-dimensional optimization." *Proceedings of 2009 IEEE Congress on Evolutionary Computation*, 785-792.

[9] Davidson, I., and Ravi, S.S. (2007). "The complexity of non-hierarchical clustering with instance and cluster level constraints." *Data Mining and Knowledge Discovery*, 14(1), 25-61.

[10] Fan, C.Y., Fan, P.S., Chan, T.Y., and Chang, SH. (2012). "Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals." *Expert Systems with Applications*, 39, 8844-8851.

[11] Higham, D.J., Kalna, G., and Kibble, M. (2007). "Spectral clustering and its use in bioinformatics." *Journal of Computational and Applied Mathematics*, 204, 25-37.

[12] Jain, A.K. (2010). "Data clustering: 50 years beyond K-means." *Pattern Recognition Letters*, 31(8), 651-666

[13] Kim, D.H., Abraham, A., and Cho, J.H. (2007). "A hybrid genetic algorithm and bacterial foraging approach for global optimization." *Information Sciences*, 177(18), 3918-3937.

[14] Kirschbaum, D.B., Adler, R., Hong, Y., and Lerner-Lam, A. (2009). "Evaluation of a preliminary satellite-based landslide hazard algorithm using global landslide inventories." *Natural Hazards and Earth System Science*, 9, 673-686.

[15] Liu, Y.L. (2010). *The self-organization map on detecting landslide area*, Master thesis, National Chung Hsing University, Taiwan, ROC. (in Chinese).

[16] Pradhan, B., Lee, S., and Buchroithner, M.F. (2010a). "Remote sensing and GIS-based landslide susceptibility analysis and its cross-validation in three test areas using a frequency ratio model." *Photogrammetrie - Fernerkundung - Geoinformation*, l, 17-32.

[17] Pradhan, B., Lee, S., and Buchroithner, M.F. (2010b). "A GIS-based back-propagation neural network model and its cross application and validation for landslide susceptibility analyses." *Computers, Environment and Urban Systems*, 34(3), 216-235.

[18] Pradhan, B. (2013). "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS." *Computers & Geosciences*, 51, 350-365.

[19] Pradhan, B. Mansor, S., Pirasteh, S., and Buchroithner, M. (2011). "Landslide hazard and risk analyses at a landslide prone catchment area using statistical based geospatial model." *International Journal of Remote Sensing*, 32(14), 4075-4087.

[20] Basu, S., Davidson, I., and Wagstaff, K. (2008). *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, ISBN 978-1-58488-996-0.

[21] Trelea, I.C. (2003). "The particle swarm optimization algorithm: Convergence analysis and parameter selection." *Information Processing Letters*, 85(6), 317-325.

[22] Wagstaff, K., and Cardie, C. (2000). "Clustering with instance-level constraints." *Proceedings of the Seventeenth International Conference on Machine Learning*, 1103-1110.

[23] Wagsta, K., Cardie, C., Rogers, S., and Schroedl, S. (2001). "Constrained k-means clustering with background knowledge." *Proceedings of the Eighteenth International Conference on Machine Learning*, 577-584.

[24] Wan, S. (2009). "A spatial decision support system for extracting the core factors and thresholds for landslide susceptibility map." *Engineering Geology*, 108(3-4), 237-251.

[25] Wan, S., Lei, T.C., and Chou, T.Y. (2012). "A landslide expert system: image classification through integration of data mining approaches for multi-category analysis." *International Journal of Geographical Information Science*, 26(4), 747-770.

[26] Wan, S. (2013). "Entropy-based particle swarm optimization with clustering analysis on landslide susceptibility mapping." *Environmental Earth Sciences*, 68(5), 1349-1366.

[27] Wan, S., Yen, J.Y., Lin, C.Y., and Chou, T.Y. (2015). "Construction of knowledge-based spatial decision support system for landslide mapping using fuzzy clustering and KPSO analysis." *Arabian Journal of Geosciences*, 8(2), 1041-1055.

[28] Wan, S., Lei T-C., and Chou, T-Y. (2014). "Optimized object-based image classification: development of landslide knowledge decision support system." Arabian Journal of Geosciences, 7(5), 2059-2070.

[29] Wagstaff, K., R.Caruana, and I. Davidson, Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC 2008.

[30] Yilmaz, I. (2009). "Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat-Turkey)." *Computers & Geosciences*, 35(6), 1125-1138.

2018 年 03 月 05 日　收稿
2018 年 06 月 30 日　修正
2018 年 07 月 10 日　接受
(本文開放討論至 2018 年 09 月 31 日)