Analysis of secondary-factor combinations of landslides using improved association rule algorithms: a case study of Kitakyushu in Japan

Jiaying Li, Wei-Dong Wang, Zheng Han and Guangqi Chen

School of Civil Engineering, Central South University, Changsha, Hunan, China; MOE Key Laboratory of Engineering Structures of Heavy-haul Railway, Central South University, Changsha, Hunan, China; School of Engineering, Kyushu University, Fukuoka, Japan

ABSTRACT
Landslide analysis prevents landslides from threatening residents’ safety and property, and the predominant method is susceptibility assessment which is cumbersome and time-consuming. The association rule algorithm (ARA) is proposed to mine the correlation between the factors and landslides simply and rapidly. The original ARA cannot reflect the scope of landslides which is non-negligible for landslide analysis and is thus improved to mine the frequent secondary-factor combinations (SFCs). Firstly, eight factors are selected using the out-of-bag error and chi-squared ($\chi^2$) test. The accuracy of the factor selection is further verified employing landslide susceptibility assessment which is predicted using 30% of study grid data selected randomly as the training data. The improved ARA employs the area of historical landslides to mine the frequent SFCs, and the results are then verified by the frequency ratio and $\chi^2$ test. It is concluded that the frequent SFCs are: (21, 41), (21, 74), (34, 41), (34, 74), (41, 74), (21, 41, 74), and (34, 41, 74), and the area with the SFCs needs special protection. The present study provides a valuable reference for the primary prevention of landslides.

ARTICLE HISTORY
Received 1 December 2020
Accepted 21 June 2021

KEYWORDS
Landslide prevention; data mining; improved ARA; SFC; frequent combinations

1. Introduction
The occurrence of geo-hazards leads to casualties, property damage, and environmental issues (Wang et al. 2019; Li et al. 2021). The prediction and prevention of geo-hazards is the focus of current scholars (Metternicht et al. 2005; Ma et al. 2019; Li et al. 2020). However, geo-hazards, such as landslides, are difficult to predict accurately even using current advanced technology due to complex natural and human factors, such as real-time rainfall and mining, and environmental compound elements, such as geological condition and climatic condition. Analyzing and predicting
landslides influenced by multi-factors is thus a difficult problem in scientific research (Chen et al. 2020). Anthropogenic activities, such as deforestation, engineering construction, and improper land use, and natural environments, such as heavy rainfalls and earthquakes, result in slope instability and reshape the topography with complex dynamics (Confuorto et al. 2017; Cebulski et al. 2020; Gomes et al. 2020). Nevertheless, the importance of influencing factors is difficult to determine. The common methods are field surveys and aerial photograph interpretation (Eker and Aydı n 2021). Meanwhile, the machine learning algorithm (MLA) is an alternative method because of the low cost and time cost (Sameen et al. 2020). It is broadly divided into supervised, unsupervised, and reinforcement learning (Leem and Kim 2020). In recent years, MLA has been popularly used for biomedicine (Sung et al. 2020; Qin et al. 2021), software information technology (González et al. 2020; Singh and Singh 2020), and ecological environment (Ge et al. 2020; Obsie et al. 2020; Wang, Zhou, et al. 2020). Currently, the landslide susceptibility assessment has been predicted using MLA, such as support vector machine (SVM) (Hu et al. 2020; Saha and Saha 2020; Wang, Feng, et al. 2020), deep learning algorithms (Bui et al. 2020; Dao et al. 2020), artificial neural network ensemble (Bragagnolo et al. 2020; Fang et al. 2020), optimized machine learning methods (Chen et al. 2020; Chen and Chen 2021), and optimized intelligence models (Chen, Chen, Peng, et al. 2021; Chen and Li 2020; Zhao and Chen 2020). Hypothesis tests, such as chi-squared ($\chi^2$) test, and statistical analysis methods are proposed to select the significant factors (Pourghasemi, Kornejad, et al. 2020; Sahin et al. 2020; Sameen et al. 2020; Wang, Kariminejad, et al. 2020). To date, the methods have been successfully utilized in many study fields, namely text categorization (Yang 1997), credit risk assessments (Attigeri et al. 2017), and landslide susceptibility mapping (Sahin et al. 2020). The statistical analysis methods primarily include discriminant analysis (Dao et al. 2020), cluster analysis (Melchiorre et al. 2008), and correlation analysis (Wistuba et al. 2021), etc. The statistical analysis methods commonly used for the selection of landslide-influencing factors are multicollinearity analysis (Du et al. 2020), accuracy analysis in the random forest (RF) model (Sun et al. 2020), and recursive feature elimination (Sun et al. 2020). Relevant literature proposes various models for susceptibility assessment of landslides using the selected factors (Pham et al. 2016; Huang and Zhao 2018). Although the assessment results are refined and accurate, the process of assessment using conventional statistical methods is laborious, cumbersome, and time-consuming. The methods are thus not convenient enough to apply to the susceptibility assessment with excessive data and cumbersome processes.

An alternative approach needs to be developed due to the limitation of susceptibility assessment of landslides. Researchers tend to pay attention to the factor combination prone to landslides (Pourghasemi, Kariminejad, et al. 2020; Yao et al. 2019), which requires the methods to extract useful information from large amounts of data quickly and accurately. Data mining is an effective method to extract knowledge from complicated data (Ouyang et al. 2011; Witten et al. 2011; Sameen et al. 2020).
Association rule algorithm (ARA) is an effective data mining algorithm to analyze the landslide factors (Tsai et al. 2013).

ARA is a process of discovering associations among items or itemsets (Bagui et al. 2020). A classic algorithm for discovering frequent itemsets and association rules is the Apriori algorithm which requires scanning the database multiple times (Agrawal and Srikant 2000; Xie et al. 2019). It is applied in many research fields such as engineering applications (Guo et al. 2014; Singh et al. 2018) and data management (Cheng et al. 2015). The Frequent Pattern (FP)-Growth algorithm is another well-known algorithm for discovering frequent itemsets in a concise form (Pei and Yin 1970) and only scans the dataset twice without candidate itemsets (Bagui et al. 2020). Due to its efficiency, it is applied widely to electric management (Wang and Cheng 2018) and network detection (Hidayanto et al. 2017). The Eclat algorithm represents the vertical database without traversing the database repeatedly (Zaki 2000). It helps understand the associations between items and performs better to long patterns (Das et al. 2018). It is applied to various studies, namely transport management (Zheng and Wang 2014; Das et al. 2018) and energy consumption (Liu et al. 2020).

Mining the deep information of landslides, namely the frequent secondary-factor combinations (SFCs), using ARA is simpler and more rapid for the preliminary prevention of landslides, and it also further analyzes the association between causative factors and landslides. However, the original ARA is difficult to apply to landslide analysis because it mines the frequent itemsets by counting the occurrence number of landslides. It cannot reflect the scope of landslides which is a non-negligible parameter for analysis. The original ARA thus needs to be improved by learning from the

![Flowchart of the methodology.](image-url)
area of historical landslide to apply to the study issue. Few studies in the current literature have improved the original ARA.

In the present study, the influencing factors of landslides are selected using out-of-bag (OOB) error and $\chi^2$ test. The receiver operating characteristic (ROC) curve is used for the verification of susceptibility assessment of landslides obtained using the RF model, deep belief network (DBN) model, and SVM model, which further verifies the accuracy of factor selection. The Apriori algorithm, FP-Growth algorithm, and Eclat algorithm are improved to mine the frequent SFCs. The association rules between the SFCs with landslides are then verified using the frequency ratio (FR) and $\chi^2$ test.

2. Methodologies

2.1. Methodological flow

The flowchart consisting of five main steps are as follows: (a) preparing the data of influencing factors and historical landslides; (b) selecting the influencing factors and determining their importance to establish a factor system; (c) evaluating the landslide susceptibility using various models based on the selected factors and verifying the assessment results to further prove the accuracy of the factor selection; (d) taking the area of historical landslides as a parameter to optimize the original ARA and mining the frequent SFC; (e) verifying the frequent SFCs. The methodologies in the present study are executed and presented in Figure 1.

2.2. Selection of influencing factor

2.2.1. OOB error and $\chi^2$ test

The OOB error is an index of feature selection (Arora and Kaur 2020; Wang, Kariminejad, et al. 2020). Not only can it be used to obtain the significance of features, but also determine the optimal number of features. Another index of feature selection is Gini importance, but it is difficult to determine the optimal number of features using the index. The OOB is thus used to determine the importance of the influencing factors and the optimal factor number in the present study.

It assumes that the total number of OOB data is $O$ which is classified using the RF classifier. The number of classification errors $X$ is obtained because of the known classification of OOB data, and the OOB error is obtained using the ratio of $X$ to $O$. In this study, the influencing factors are ranked according to their OOB scores obtained using the RF algorithm, and the least important factors are eliminated based on the recursive feature elimination. According to OOB errors of different factor sets in the elimination process, the optimal factor set is selected as the factor system.

The $\chi^2$ test is a hypothesis testing method for variable classification based on the chi-squared distribution. It is a well-established technique for measuring independence and determining whether the variables are related (Doğan et al. 2021). It assumes that the actual values are unrelated to theoretical values, and the $\chi^2$ can be calculated by Eq. 1.
\[ \chi^2 = \sum \frac{(A-T)^2}{T} \]  

where \( A \) is the actual value; and \( T \) is the theoretical value.

### 2.2.2. Susceptibility assessment model

The RF model, DBN model, and SVM model have extensively been applied in susceptibility assessment, and it provides a solid foundation for the assessment. In the present study, the landslide region is set to 1, while the non-landslide region is set to 0. 30% of study data are selected randomly as the training data to predict the landslide susceptibility of each grid in the study area. The ROC curve is a commonly used method to verify the assessment results of landslide susceptibility (Chen, Lei, et al. 2021; Chen, Chen, Janizadeh, et al. 2021). In the present study, the landslide susceptibility is thus evaluated using the three models, and the accuracies of the assessment results are verified using the ROC curve. The more accurate the assessment results are, the higher the accuracy of factor selection is. RF model is an advanced integrated learning algorithm based on the ensemble of unpruned classification trees which are created by bootstrap sampling and random feature selection, and the results are obtained by a majority voting of the various classification trees (Xie et al. 2019).

DBN is an efficient unsupervised learning algorithm in deep learning and a probabilistic generative model composed of restricted Boltzmann machines (RBM). RBM consists of the visual layer and hidden layer. There is no connection between the neurons in each layer. A DBN is structured by several RBMs. The hidden layer of the previous RBM is the visual layer of the next RBM, while the output of the previous RBM is the input of the next RBM.

SVM is performed by many kernels function, such as linear function, polynomial function, and radial basis function (RBF). The main parameters are the penalty parameter \((c)\) and the kernel function parameter \((g)\). In the present study, the RBF is chosen as the kernel function, and the optimal \(c\) and \(g\) are found using the grid method and cross-validation.

### 2.3. Improved association rule learning

Association rule learning is a common algorithm for discovering strong rules hidden in a large database and is used for mining the frequent SFC of landslides in the study area.

#### 2.3.1. Original ARA

The dataset of the original ARA is \{TID: itemset\} in which the TID is the thing identifier, and the itemset is the content of the TID. The two parameters are the data basis to find the association rules and frequent itemsets. There are two sub-problems in the original ARA: (a) finding out the frequent datasets whose supports are greater than the specified minimum support, and (b) determining the strong association rule based on frequent itemsets and the minimum confidence. The support and confidence are obtained respectively by Eq. 2 and Eq. 3.
Support \((A, B) = P(A \& B)\), 

\[
\text{Confidence}(A \Rightarrow B) = P(A|B) = \frac{P(A \& B)}{P(A)},
\]

where \(P(A \& B)\) is the probability of \(A\) and \(B\) concurrently; \(P(A|B)\) denotes the probability of \(B\) given \(A\); and \(P(A)\) is the probability of \(A\).

Apriori algorithm is an ARA for Boolean mining based on a recursive algorithm based on the two key steps, namely connection step pruning (Wang et al. 2018). The FP-Growth algorithm compresses the data of frequent itemsets to a frequent pattern tree and retains the itemset association information. There is no need for the candidate set and only a need to traverse the database twice. The Eclat algorithm is a depth-first-search algorithm based on the set intersection. It is applied to sequential and parallel issues with the characteristic of local reinforcement. Its inverted theory considers the item and transaction ID as the key and value, respectively. The detailed steps of the Apriori algorithm, FP-Growth algorithm, and Eclat algorithm are proposed as Figure 2.

The original ARA mines the itemsets meeting the requirements of support and confidence from a considerable amount of itemsets by counting the number of items.
It only can mine frequent SFCs by using the occurrence number of landslides and is difficult to apply to the present study.

2.3.2. Improved ARA
In the present study, the scope of landslides is a non-negligible parameter for association rule analysis. However, the two parameters of the original ARA cannot accurately reflect the parameter. The characteristic is introduced, and the improved ARA mines the frequent itemsets based on \{TID: itemset\, characteristic\} in which the characteristic is a continuous variable. The frequent itemsets are mined based on the accumulation of the corresponding characteristic rather than the occurrence number of the itemsets.

The characteristic, namely the area of historical landslides in this paper, is introduced in the improved ARA. The support and confidence are optimized using the area of historical landslides (Eqs. 4, 5).

\[
\text{Support}(A, B) = \frac{\text{Area}(A \& B)}{\sum \text{Area}},
\]

\[
\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A, B)}{\text{Area}(A) / \sum \text{Area}} = \frac{\text{Area}(A \& B)}{\text{Area}(A)},
\]

in which \(\text{Area}(A \& B)\) is the area of historical landslides with the secondary-factor \(A\) and \(B\); \(\text{Area}(A)\) is the area of historical landslides with the secondary-factor \(A\); and \(\sum \text{Area}\) is the area of historical landslides.

For the improved Apriori algorithm, after the datasets are scanned, the candidate itemsets are generated by accumulating the landslide area. The frequent SFC is mined and then connected and pruned based on the support in Eq. 4. The confidence used to generate rule is renewed using Eq. 5. For the FP-Growth algorithm, the root nodes created in the frequent item table also include the characteristic accumulation, when building the FP-Trees. For the Eclat algorithm, the support is calculated based on the characteristic accumulation rather than the length of the TID set, and there is the same improvement in the prior theory as the Apriori algorithm when the candidate itemsets are generated.

On the other hand, the FR is used to prove the association between the frequent SFCs obtained using improved ARA and landslides, and FR is obtained by Eq. 6.

\[
FR = \frac{P(LF_i)}{P(F_i)} = \frac{A_{LF_i}/A_{F_i}}{A_L/A} = \frac{A_{LF_i} \cdot A}{A_{F_i} \cdot A_L},
\]

where \(A_{LF_i}\) is the area of historical landslide with the SFC \(F_i\); \(A_{F_i}\) is the area with SFC \(F_i\); \(A_L\) is the area of landslides; and \(A\) is the area of the study area.
3. Data collection and preparation

3.1. Study area

Kitakyushu is located in the northern Kyushu Island, Japan (Figure 3). It indicates the north latitude range of $33^\circ58' - 33^\circ43'$ and the east longitude range of $130^\circ40' - 131^\circ01'$, with an area of 488.78 km². The terrain tilts from north to south with a relative altitude of 954 m. According to the geological characteristics and terrain genesis of the study area, it can be mainly divided into four regions, namely the southern mountain region, central plain region, northeastern mountain region, and northwestern hilly region. The terrain is smooth and characterized by the overburden soil layer thickness of about 1.30-1.76 m. According to the Ministry of Land, Infrastructure, Transport and Tourism of Japan, the geological condition of the study area is complex with an active geological tectonic movement. The geological formations are mainly sedimentary rock and igneous rocks, and the landfill area is more than 5% of the study area.

The study area is warm and humid throughout the year with an average annual temperature of 16.2°C and average annual precipitation of 1265 mm (Sun et al. 2011). In the northern Kitakyushu, the area has a typical Sea of Japan climate, while the climate in the eastern region belongs to the Seto Inland Sea Climate which is warm and dry. The precipitation significantly varies, concentrated during the rainy season and typhoon season. Meanwhile, the study area is located in the Pacific Rim Volcanic Seismic Zone at the junction of Eurasian and Pacific plates with frequent crustal movement. There are thus frequently occurring landslides induced by rainfall and earthquake, and most of the landslides are shallow landslides with a sliding surface depth of less than 6 m. The data of historical landslides is obtained from the

Figure 3. Location of the study area: (a) Fukuoka in Japan; (b) Kitakyushu in Fukuoka; and (c) historical landslide area in the study area.
Figure 4. Maps of the various factors.
Bureau of Land Policy in the Ministry of Land, Infrastructure, Transport and Tourism of Japan, and the geological environment of landslide-prone regions is complex, with active geological tectonic movements such as earthquakes. The historical landslides from 1992 to 2011 are shown in Figure 3.

3.2. Case influencing factors

Landslides are typical multi-factor complex geo-hazards, and their mechanisms are complicated with various induced factors. The Kitakyushu is considered as the study area, and the data in the study area is obtained using the field environment and related literature. The digital elevation model (DEM) data at a resolution of 10 m are provided by the Geospatial Information Authority of Japan. The geology conditions, such as lithology, surface information, and runoff, are provided by the Land and Water Resources Bureau in the Ministry of Land, Infrastructure, Transport and Tourism of Japan. The present study establishes a factor system consisting of ten factors, namely soil thickness (ST), cumulative runoff (CR), distance from road (DRO), topography, elevation, distance from construction line (DCL), slope, distance from railway (DRA), lithology, and distance from river (DRI). There are two qualitative factors and eight quantitative factors in the established factor system. The factors in

Figure 5. Importance of factors in the various factor selections based on the recursive feature elimination.
the present study are divided into four levels to avoid too many factor levels leading to the excessively great computational amount, and there are 40 secondary-factors. Four quantitative factors, namely the ST, CR, elevation, and slope, are reclassified using the natural break method. However, four additional quantitative factors, namely the DRO, DCL, DRA, and DRI, are unsuitable to be reclassified by the same method because the factors cannot affect the entire study area and their impacts disappear beyond a short distance from them. Therefore, the four factors are reclassified within a certain distance according to the actual influence scope of the factors. The maps of the various factors are thus shown in Figure 4.

4. Results

4.1. Selection and verification of influencing factor

A significant characteristic of the RF model is the OOB which can calculate the feature importance. Based on the recursive feature elimination, the factor importance, namely the OOB score, in the various factor selections and the OOB errors of various factor selection are presented respectively in Figure 5 and Figure 6.

As shown in Figure 5 and Figure 6, the OOB error of eight influencing factors is minimum, and the factors selected in the present study are thus DCL, topography, DRO, slope, DRI, ST, CR, and DRA.

The \( \chi^2 \) test can be used to verify the significance of factors affecting landslides, and the \( \chi^2 \) of the eight factors are obtained in Table 1 using Statistical Product and Service Solutions (SPSS) and compared with the test critical value. The eight \( \chi^2 \) are greater than the critical value (\( k = 3.84 \)) which proves the accuracy of factor selection.

4.2. Landslide susceptibility assessment

The landslide susceptibility assessment is employed to further verify the accuracy of factor selection. The assessment results of landslide susceptibility are obtained using the RF model, DBN model, and SVM model. The results are then classified using the
natural breakpoint method into five levels, namely very low susceptibility, low susceptibility, medium susceptibility, high susceptibility, and very high susceptibility, and the level maps are obtained using the three models in Figure 7. The ROC curves of the assessment results using the three models are obtained employing the historical landslide data as a reference to verify the accuracy (Figure 8).

As can be seen from Figure 7, the level distributions of landslide susceptibility maps obtained by the three models are very similar. The high-susceptibility areas are mainly distributed in the south-central region, while the low-susceptibility areas are distributed in the northern region. The area under the curve (AUC) is in the range of 0.5-1, and an AUC of 1 indicates perfect prediction, while an AUC of 0.5 indicates useless prediction. The AUCs of the RF model, DBN model, and SVM model are respectively 0.909, 0.878, 0.809. The three AUCs of the three models are greater than 0.8, and it indicates that the performance of assessment results is excellent and results in high accuracy. The highest accuracy is recorded by the RF model, followed by the DBN model and SVM model. It can be concluded that the RF model has better accuracy than the other models. Meanwhile, the results further prove the accuracy of the factor selection.

4.3. Mining and verification of the frequent SFC

The secondary-factors of the study area are coded and shown in Table 2.

The improved ARA is executed in Python, and two parameters, namely the minimum support and confidence need to be set in the algorithm. Two methods are usually applied, namely the trial and error method and using other parameters to replace the parameters (Zhang et al. 2017). However, the substitute parameter still needs to be set if using the second method, and the issue is not fundamentally addressed. The former approach is thus employed in the present study. The minimum support and confidence are selected respectively as 60% and 70% using the trial and error method,

Table 1. $\chi^2$ of the selected factors.

| Factors | DCL | Topography | DRO | Slope | DRI | ST | CR | DRA |
|---------|-----|------------|-----|-------|-----|----|----|-----|
| $\chi^2$ | 135.9 | 3326.7 | 285.1 | 332.6 | 392.5 | 1484.5 | 446.9 | 997.5 |

Figure 7. Level maps of landslide susceptibility assessment using (a) RF model; (b) DBN model; and (c) SVM model.
and the two parameters are applied to the improved ARA. The frequent SFCs are obtained as follows: (21, 41), (21, 74), (34, 41), (34, 74), (41, 74), (21, 41, 74) and (34, 41, 74). The average confidences of the frequent SFCs are shown in Table 3.

Meanwhile, the FR and $\chi^2$ are used to verify the association between the frequent SFCs and landslides and are presented in Table 4.
The greater the FR is, the greater the probability of landslides is. As can be seen from the table above, the FRs are greater than one, and it indicates the frequent SFCs are prone to landslides. The SFCs are thus sorted: (34, 41), (34, 41, 74), (21, 41, 74), (21, 41), (41, 74), (34, 74), and (21, 74). The $\chi^2$ is greater than the critical value, which denotes that the frequent SFCs are prone to landslides. The SFCs are sorted according to the $\chi^2$: (34, 41), (41, 74), (34, 41, 74), (21, 41), (21, 41, 74), (34, 74), and (21, 74). All FRs and $\chi^2$ are correspondingly greater than one and critical value, which indicates a tight relationship between the SFCs and landslides. The most frequent SFC is (34, 41), namely the distance from road $> 100$ m and the topography of the mountain, and the area with the frequent SFCs needs special protection.

4.4. Comparison with original ARA

The dataset of the original ARA is \{TID: itemset\}, and the number of itemsets is the number of historical landslides. The TID is the identifier of the landslides, and the itemset is the secondary-factor with the largest area in the corresponding landslide. However, there is no combination meeting the requirement of the minimum support of improved ARA in the original ARA. The minimum support is thus set to 20% using the trial and error method, while the minimum confidence is set to 40%. As a result, there are three SFCs, namely (21, 41); (21, 74); and (41, 74), and their confidences are 48.59%; 41.90%; and 50.70%, respectively.

It is concluded that even if the minimum support is set to 20%, the maximum confidences of the three SFCs are only about 50%. It denotes that the results of data mining by taking the number of landslides as the research objectives are inaccurate enough for the study area, and the improved ARA is more applied to the study area than the original ARA.

5. Discussions

According to the relevant literature (Xie et al. 2019), the geo-hazards including landslides are analyzed using the data statistics and research reports. However, much of the literature on the analysis of geo-hazards pays particular attention to susceptibility assessment. The MLAs, such as the RF model and SVM model, are the most commonly used approach (Chang et al. 2019; Merghadi et al. 2020). In recent years, deep learning algorithms, such as the DBN model, begin to outperform previous traditional methods and develop rapidly (Dou et al. 2020; Wang, He, et al. 2020). The models have been the focus of studies on landslide prevention. Although the susceptibility

| Combination | FR  | $\chi^2$ |
|-------------|-----|---------|
| (21, 41)    | 5.20| 37.00   |
| (21, 74)    | 3.71| 4.29    |
| (34, 41)    | 5.97| 46.80   |
| (34, 74)    | 4.95| 22.99   |
| (41, 74)    | 5.16| 43.88   |
| (21, 41, 74)| 5.23| 30.97   |
| (34, 41, 74)| 5.91| 39.81   |
assessment results are refined and accurate, the process of assessment is laborious, cumbersome, and time-consuming. An alternative approach thus needs to be developed. Mining the deep information of the landslides using the data mining algorithm is simpler and more rapid and also valuable for the preliminary prevention of landslides.

Current studies have investigated the triggering factors and threshold analysis of landslides employing the data mining methods and generated the association rules between triggering factors and deformation (Ma et al. 2017; Miao et al. 2021). Meanwhile, researchers have determined the cause-and-effect relationships between factors and landslide movement by identifying the contribution of each parameter to landslides employing association rule mining (Ma et al. 2017). However, the original ARA only can be used for discrete problems, and the problems involving the continuous variable cannot be solved. The method of the previous studies cannot apply to this study because it cannot accurately reflect the scope of landslides which is a non-negligible parameter, and few current studies have improved the original ARA. The present study thus introduces the area of historical landslides to improve the original ARA to apply to the analysis of landslide factors.

The landslide information can be mined by employing the improved association rule analysis. Finding the frequent SFCs and discovering the association rules between the landslides and the influencing factors are particularly useful for landslide prevention. It provides a novel insight into the improvement of the ARA and a valuable reference for the primary prevention of landslides. However, the proposed method mines the association rules based on the scope of landslides, and it thus only addresses the issue in terms of space and has certain spatiotemporal limitations. Future research should concentrate on the investigation of extracting more valuable landslide information that optimizes the analysis of landslide prevention and rescue.

6. Conclusions

In the present study, the ARAs, namely the Apriori algorithm, FP-growth algorithm, and Eclat algorithm, are improved for mining the frequent SFCs of landslides. There are few studies on optimizing ARA in the same method and employing the improved ARA to landslide analysis. The conclusions are obtained as follows:

1. The influencing factors of landslides in the study area are selected using the OOB error and \( \chi^2 \) test. The factors are considered as the evaluation indices to evaluate the landslide susceptibility using the RF model, DBN model, and SVM model, which further verifies the accuracy of the factor selection.

2. The ARA is improved by introducing a continuous variable, namely the area of historical landslides, to apply to the present study, and the frequent SFCs are mined. It is proved that the frequent SFCs are: (21, 41), (21, 74), (34, 41), (34, 74), (41, 74), (21, 41, 74), and (34, 41, 74), and the association between the SFCs and landslides is verified using their FRs and \( \chi^2 \) test. Their FRs are corresponding 5.20, 3.71, 5.97, 4.95, 5.16, 5.23, and 5.91, and their \( \chi^2 \) are corresponding
37.00, 4.29, 46.80, 22.99, 43.88, 30.97, and 39.81 which are all greater than the critical value.

3. The SFCs are sorted according to the FRs: (34, 41), (34, 41, 74), (21, 41, 74), (21, 41), (41, 74), (34, 74), and (21, 74), and they are also sorted according to $\chi^2$: (34, 41), (41, 74), (34, 41, 74), (21, 41), (21, 41, 74), (34, 74), and (21, 74). The most frequent SFC is (34, 41), namely the distance from road $>100$ m and the topography of the mountain, and the area with the frequent SFCs needs special protection.

4. The results obtained employing the original ARAs are inaccurate enough for the study area, and the improved ARA has more widespread applicability than the original ARA. The improved ARA provides a valuable reference for the primary prevention of landslides.

Disclosure statement

There are no financial competing interests.

Funding

This work was supported by the National Natural Science Foundation of China under Grant No. 51478483 and No. 41702310 and the China Scholarship Council.

Data availability statement

The data that support the findings of this study are available from the corresponding authors, upon reasonable request.

References

Agrawal R, Srikant R. 2000. Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference Very Large Data Bases (VLDB), 1215. Available from: https://rakesh.agrawal-family.com/papers/vldb94apriori.pdf.

Arora N, Kaur PD. 2020. A Bolasso based consistent feature selection enabled random forest classification algorithm: an application to credit risk assessment. Appl Soft Comput. 86: 105936.

Attigeri G, Pai M, Pai R. 2017. Credit risk assessment using machine learning algorithms. Adv Sci Lett. 23(4):3649–3653.

Bagui S, Devulapalli K, Coffey J. 2020. A heuristic approach for load balancing the FP-growth algorithm on MapReduce. Array. 7:100035.

Bragagnolo L, Silva RVd, Grzybowski JMV. 2020. Artificial neural network ensembles applied to the mapping of landslide susceptibility. CATENA. 184:104240.

Bui DT, Tsangaratos P, Nguyen V-T, Liem NV, Trinh PT. 2020. Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. CATENA. 188:104426.

Cebulska J, Pasierb B, Wieczorek D, Zielinski A. 2020. Reconstruction of landslide movements using digital elevation model and electrical resistivity tomography analysis in the Polish outer carpathians. CATENA. 195:104758.
Chang K-T, Merghadi A, Yunus AP, Pham BT, Dou J. 2019. Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. Sci Rep. 9(1):12296.

Chen X, Chen W. 2021. GIS-based landslide susceptibility assessment using optimized hybrid machine learning methods. CATENA. 196:104833. https://doi.org/10.1080/10106049.2021.1892212

Chen Y, Chen W, Janizadeh S, Bhunia GS, Bera A, Pham QB, Linh NTT, Balogun A-L, Wang X. 2021. Deep learning and boosting framework for piping erosion susceptibility modeling: spatial evaluation of agricultural areas in the semi-arid region. Geocarto Int. 1–27.

Chen W, Chen X, Peng J, Panahi M, Lee S. 2021. Landslide susceptibility modeling based on ANFIS with teaching-learning-based optimization and Satin bowerbird optimizer. Geosci Front. 12(1):93–107.

Chen W, Chen Y, Tsangaratos P, Ilia I, Wang X. 2020. Combining evolutionary algorithms and machine learning models in landslide susceptibility assessments. Remote Sensing. 12(23):3854.

Chen W, Lei X, Chakrabortty R, Chandra Pal S, Sahana M, Janizadeh S. 2021. Evaluation of different boosting ensemble machine learning models and novel deep learning and boosting framework for head-cut gully erosion susceptibility. J Environ Manage. 284:112015.

Chen W, Li Y. 2020. GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models. CATENA. 195:104777.

Cheng X, Su S, Xu S, Li Z. 2015. DP-Apriori: a differentially private frequent itemset mining algorithm based on transaction splitting. Comput Secur. 50:74–90.

Confuorto P, Di Martire D, Centolanza G, Iglesias R, Mallorqui JJ, Novellino A, Plank S, Ramondini M, Thuro K, Calcatera D. 2017. Post-failure evolution analysis of a rainfall-triggered landslide by multi-temporal interferometry SAR approaches integrated with geotechnical analysis. Remote Sens Environ. 188:51–72.

Dao DV, Jaafari A, Bayat M, Mafi-Gholami D, Qi C, Moayedi H, Phong TV, Ly H-B, Le T-T, Trinh PT, et al. 2020. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. CATENA. 188:104451.

Das S, Dutta A, Jabayer M, Bibeka A, Wu L. 2018. Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: Exploration using ‘Eclat’ association rules to promote safety. Int J Transp Sci Technol. 7(2):114–123.

Doğan O, Taspınar S, Bera AK. 2021. A Bayesian robust chi-squared test for testing simple hypotheses. Journal of Econometrics. 222(2):933–958.

Dou J, Yunus AP, Merghadi A, Shirzadi A, Nguyen H, Hussain Y, Avtar R, Chen Y, Pham BT, Yamagishi H. 2020. Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. Sci Total Environ. 720:137320.

Du J, Glade T, Woldai T, Chai B, Zeng B. 2020. Landslide susceptibility assessment based on an incomplete landslide inventory in the Jilong Valley, Tibet, Chinese Himalayas. Eng Geol. 270:105572.

Eker R, Aydin A. 2021. Long-term retrospective investigation of a large, deep-seated, and slow-moving landslide using InSAR time series, historical aerial photographs, and UAV data: The case of Devrek landslide (NW Turkey). CATENA. 196:104895.

Fang Z, Wang Y, Peng L, Hong H. 2020. Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. Comput Geosci. 139:104470.

Ge G, Shi Z, Zhu Y, Yang X, Hao Y. 2020. Land use/cover classification in an arid desert-oasis mosaic landscape of China using remote sensed imagery: performance assessment of four machine learning algorithms. Global Ecol Conserv. 22:e00971.

Gomes PIA, Aththanayake U, Deng W, Li A, Zhao W, Jayathilaka T. 2020. Ecological fragmentation two years after a major landslide: correlations between vegetation indices and geo-environmental factors. Ecol Eng. 153:105914.
González S, García S, Del Ser J, Rokach L, Herrera F. 2020. A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities. Informat Fusion. 64:205–237.

Guo Z, Chi D, Wu J, Zhang W. 2014. A new wind speed forecasting strategy based on the chaotic time series modelling technique and the Apriori algorithm. Energy Convers Manage. 84:140–151.

Hidayanto BC, Muhammad RF, Kusumawardani RP, Syafaat A. 2017. Network intrusion detection systems analysis using frequent item set mining algorithm FP-max and Apriori. Procedia Comput Sci. 124:751–758.

Hu Q, Zhou Y, Wang S, Wang F. 2020. Machine learning and fractal theory models for landslide susceptibility mapping: case study from the Jinsha River Basin. Geomorphology. 351:106975.

Huang Y, Zhao L. 2018. Review on landslide susceptibility mapping using support vector machines. CATENA. 165:520–529.

Leem J, Kim H. 2020. Action-specialized expert ensemble trading system with extended discrete action space using deep reinforcement learning. Plos One. 15(7):e0236178.

Li J, Wang W, Han Z. 2021. A variable weight combination model for prediction on landslide displacement using AR model, LSTM model, and SVM model: a case study of the Xinming landslide in China. Environ Earth Sci. 80(10):386.

Li J, Wang W, Han Z, Li Y, Chen G. 2020. Exploring the impact of multitemporal DEM data on the susceptibility mapping of landslides. Applied Sciences. 10(7):2518.

Liu Y, Hu X, Luo X, Zhou Y, Wang D, Farah S. 2020. Identifying the most significant input parameters for predicting district heating load using an association rule algorithm. J Cleaner Prod. 275:122984.

Ma R, Cui C, Ma M, Chen A. 2019. Performance-based design of bridge structures under vehicle-induced fire accidents: basic framework and a case study. Eng Struct. 197:109390.

Ma J, Tang H, Hu X, Bobet A, Zhang M, Zhu T, Song Y, Ez Eldin MAM. 2017. Identification of causal factors for the Majiagou landslide using modern data mining methods. Landslides. 14(1):311–322.

Melchiorre C, Matteucci M, Azzoni A, Zanchi A. 2008. Artificial neural networks and cluster analysis in landslide susceptibility zonation. Geomorphology. 94(3–4):379–400.

Merghadi A, Yunus AP, Dou J, Whiteley J, ThaiPham B, Bui DT, Avtar R, Abderrahmane B. 2020. Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. Earth Sci Rev. 207:103225.

Metternicht G, Hurni L, Gogu R. 2005. Remote sensing of landslides: an analysis of the potential contribution to geo-spatial systems for hazard assessment in mountainous environments. Remote Sens Environ. 98(2–3):284–303.

Miao F, Wu Y, Li L, Liao K, Xue Y. 2021. Triggering factors and threshold analysis of baishuihe landslide based on the data mining methods. Nat Hazards. 105(3):2677–2696.

Obsie EY, Qu H, Drummond F. 2020. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. Comput Electron Agric. 178:105778.

Ouyang Y, Luo SM, Cui LH, Wang Q, Zhang JE. 2011. Estimation of real-time N load in surface water using dynamic data-driven application system. Ecol Eng. 37(4):616–621.

Pei J, Yin Y. 1970. Mining frequent patterns without candidate generation. Available from: http://www.cse.msu.edu/~cse960/Papers/MineFqPatteren-HPY-SIGMOD2000.pdf.

Pham BT, Pradhan B, Tien Bui D, Prakash I, Dholakia MB. 2016. A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environ Model Software. 84:240–250.

Pourghasemi HR, Kariminejad N, Gayen A, Komac M. 2020. Statistical functions used for spatial modelling due to assessment of landslide distribution and landscape-interaction factors in Iran. Geosci Front. 11(4):1257–1269.

Pourghasemi HR, Kornejad A, Kerle N, Shabani F. 2020. Investigating the effects of different landslide positioning techniques, landslide partitioning approaches, and presence-absence balances on landslide susceptibility mapping. CATENA. 187:104346.
Qin X, Liu M, Zhang L, Liu G. 2021. Structural protein fold recognition based on secondary structure and evolutionary information using machine learning algorithms. Comput Biol Chem. 91:107456.

Saha A, Saha S. 2020. Comparing the efficiency of weight of evidence, support vector machine and their ensemble approaches in landslide susceptibility modelling: a study on Kuruseong region of Darjeeling Himalaya, India. Remote Sens Appl: Soc Environ. 19:100323.

Sahin EK, Colkesen I, Acmali SS, Akgun A, Aydinoglu AC. 2020. Developing comprehensive geocomputation tools for landslide susceptibility mapping: LSM tool pack. Comput Geosci. 144:104592.

Sameen MI, Sarkar R, Pradhan B, Drukpa D, Alamri AM, Park H-J. 2020. Landslide spatial modelling using unsupervised factor optimisation and regularised greedy forests. Comput Geosci. 134:104336.

Singh S, Garg R, Mishra PK. 2018. Performance optimization of MapReduce-based Apriori algorithm on Hadoop cluster. Comput Electr Eng. 67:348–364.

Singh J, Singh J. 2020. Detection of malicious software by analyzing the behavioral artifacts using machine learning algorithms. Inf Softw Technol. 121:106273.

Sun J, Wang L., Long P., Chen G. 2011. An assessment method for reginal susceptibility of landslides under coupling condition of earthquake and rainfall. Chinese Journal of Rock Mechanics and Engineering. 30(4):752–760.

Sun D, Wen H, Wang D, Xu J. 2020. A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology. 362:107201.

Sung SM, Kang YJ, Cho HJ, Kim NR, Lee SM, Choi BK, Cho G. 2020. Prediction of early neurological deterioration in acute minor ischemic stroke by machine learning algorithms. Clin Neurol Neurosurg. 195:105892.

Tsai F, Lai J-S, Chen WW, Lin T-H. 2013. Analysis of topographic and vegetative factors with data mining for landslide verification. Ecol Eng. 61:669–677.

Wang J, Cheng Z. 2018. FP-growth based regular behaviors auditing in electric management information system. Procedia Comput Sci. 139:275–279.

Wang Y, Feng L, Li S, Ren F, Du Q. 2020. A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China. CATENA. 188:104425.

Wang W, He Z, Han Z, Li Y, Dou J, Huang J. 2020. Mapping the susceptibility to landslides based on the deep belief network: a case study in Sichuan Province. Nat Hazards. 103(3):3239–3261.

Wang W, Li J, Qu X, Han Z, Liu P. 2019. Prediction on landslide displacement using a new combination model: a case study of Qinglong landslide in China. Nat Hazards. 96(3):1121–1139.

Wang C, Pan Y, Chen J, Ouyang Y, Rao J, Jiang Q. 2020a. Indicator element selection and geochemical anomaly mapping using recursive feature elimination and random forest methods in the Jingdezhen region of Jiangxi Province, South China. Appl Geochem. 122:104760.

Wang X, Song C, Xiong W, Lv X. 2018. Evaluation of flotation working condition recognition based on an improved Apriori algorithm. IFAC-PapersOnLine. 51(21):129–134.

Wang J, Zhou Y, Xiao F. 2020. Identification of multi-element geochemical anomalies using unsupervised machine learning algorithms: a case study from Ag–Pb–Zn deposits in northwestern Zhejiang, China. Appl Geochem. 120:104679.

Wistuba M, Malik I, Gorczyca E, Słężak A. 2021. Establishing regimes of landslide activity – Analysis of landslide triggers over the previous seven decades (Western Carpathians, Poland). CATENA. 196:104888.

Witten IH, Frank E, Hall MA. 2011. Chapter 6 - Implementations: real machine learning schemes. In: Witten IH, Frank E, Hall MA, editors. Data mining: practical machine learning tools and techniques. 3rd ed. Boston: Morgan Kaufmann; p. 191–304.

Xie X, Fu G, Xue Y, Zhao Z, Chen P, Lu B, Jiang S. 2019. Risk prediction and factors risk analysis based on IFOA-GRNN and apriori algorithms: application of artificial intelligence in accident prevention. Process Safe Environment Protect. 122:169–184.
Yang Y. 1997. A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML’97); p. 412–420.
Yao W, Li C, Zuo Q, Zhan H, Criss RE. 2019. Spatiotemporal deformation characteristics and triggering factors of Baijiabao landslide in Three Gorges Reservoir region, China. Geomorphology. 343:34–47.
Zaki MJ. 2000. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng. 12(3):372–390.
Zhang Z, Pedrycz W, Huang J. 2017. Efficient frequent itemsets mining through sampling and information granulation. Eng Appl Artif Intell. 65:119–136.
Zhao X, Chen W. 2020. Optimization of computational intelligence models for landslide susceptibility evaluation. Remote Sensing. 12(14):2180.
Zheng X, Wang S. 2014. Study on the method of road transport management information data mining based on pruning Eclat algorithm and mapreduce. Proc Social Behav Sci. 138:757–766.