DSC540 Advance Machine Learning

Paper Review 2

Lavinia Wang
1473704

The uncertainties and complexities of the factors involved in causing landslides makes it difficult to analyze their influences quantitatively and to predict the probability of landslide occurrence. Researchers in *Susceptibility assessment of earthquake-induced landslides using Bayesian network: A case study in Beichuan, China* proposed a hybrid method based on Bayesian network (BN) to analyze earthquake-induced landslide-causing factors and assess their effects in order to achieve higher precision and accuracy compared to existing methods. Key techniques from landslide susceptibility assessment (LSA) modeling with BN are explored, including data acquisition and processing, BN modeling, and validation to provide a robust assessment of landslide probability.

Landslides are one of the most common secondary disasters caused by earthquakes. Analyzing the factors that cause landslides and assessing the landslide susceptibility can prevent future geological disasters and aid in decision making for the reconstruction process, which has become more important and timely following the earthquake. Three most commonly employed factor-analysis methods for landslide prediction were discussed in the paper, which are expert evaluation, data mining, and physical models.

- Expert evaluation method, which is based on a landslide inventory and historical information, can be used to evaluate and classify landslides, and determine the main contributing factors and disaster levels. The main disadvantage of this method is that it provides a qualitative assessment and it is difficult to estimate the weight of each factor quantitatively and ensure result consistency.
- Data mining method, integrated with machine learning or statistical algorithms, can be used to determine the principal components leading to landslides, calculate the weights of the factors, and construct mathematical models or rules for landslide prediction. This method provides quantitative or semiquantitative analysis of landslides, but the results are sensitive to data quality.
- The physical model method was derived from the slope displacement model. It is appropriate for assessing landslides in very small areas, whereas its capability for evaluation over a large area is insufficient.

In the paper, new methods and mechanisms was introduced by developing a Bayesian network model that can be used to analyze the factors contributing to earthquake-induced landslides and assess landslide susceptibility. The advantage of BN method is that it provides a natural way of handling missing data, allows for the combination of data with domain knowledge, facilitates learning about causal relationships between variables, and offers a method to avoid overfitting the data. BN can provide good prediction accuracy even with small sample sizes, and it can be easily combined with analytic tools to aid management.

The procedure of the study consisted of two main steps: data acquisition and processing, and LSA modeling with BN. The first step in data acquisition and processing was to collect spatial data and then integrate into the same spatial reference frame. After data processing the data were used to analyze landslide-causing factors and to assess the study area for landslide susceptibility. Taking the landslide as the root node and the landslide-causing factors as the child nodes in a tree, BN model was constructed in three steps: K2 algorithm for BN structure learning,

expectation maximization algorithm for BN parameter learning and inference engine. A posteriori probability of landslide-causing factors and the Landslide Susceptibility Index (LSI) of each pixel in the study area were calculated with the BN Junction Tree inference engine after the BN modeling. Based on a posteriori probability, the influence of landslide- causing factors was analyzed both qualitatively and quantitatively. With a coordinate transformation, landslide-susceptibility mapping based on the LSI was obtained with spatial information and the format of it can be converted to Geographic Information System (GIS) grid images. Finally, the BN model used for the LSA was validated by combining 10-fold cross-validation method with a ROC curve.

In data acquisition process, an event-based landslide inventory map between May 12, 2008 and May 19, 2008 was generated resulting in the identification of 563 earthquake-triggered landslides in the study area and a total landslide area of 35.6km^2. The landslide data were later rasterized onto a raster grid layer with a cell size of 25 m × 25 m using the cubic spline interpolation in ArcMap. A value was assigned to each pixel: 1 to indicate the presence of landslide and 0 to indicate the absence of landslide. The same grid cell size (25 m × 25 m) was used in the landslide-causing factor rasterizing process.

Different factors contribute differently to landslides, and only the key factors are required in the LSA model. Researchers then used a hybrid method combining correlation analysis, gain ratio (GR) and principal component analysis (PCA) to determine the key factors causing landslides. The first step was to detect the correlation coefficient between each pair of landslide-causing factors and calculate the GR value of the landslide-causing factor with respect to the landslide. In the second step, based on PCA, a further selection was done for the remaining landslide-causing factors of the first step. Based on the two steps above, eight landslide-causing factors: relief amplitude, lithology, distance from a fault, distance from a road, distance from a river, the Normalized Difference Vegetation Index (NDVI), land cover, and Arias intensity (AI) were selected for the LSA.

Researchers used discretization to handle continuous landslide-causing factors in BN. Among the eight selected landslide-causing factors, the attribute values of lithology and land cover were defined as discrete and were classified into predefined categories according to expert knowledge, whereas the values of relief amplitude, distance from a fault, distance from a road, distance from a river, and NDVI were defined as continuous. Supervised discretization methods and a Weka discretization filter based on a minimum description length (MDL) algorithm were used to discretize the six continuous landslide-causing factors into categories with appropriate states.

In the BN construction process, first the tree-augmented naive Bayes (TAN) BN structure was chosen by researchers because it provides better pre- diction accuracy and a better representation of the correlation among variables than other classic models such as the naive Bayes model. The initial network used for structure learning is a naive Bayes model and The K2 algorithm which is a classic search and score- based algorithm, adds a parent node to each landslide-causing factor at random with the hill climbing search algorithm, allowing all the possible structures to be described. The Bayesian information criterion (BIC) was used to choose a better structure.

Next, using the previously constructed structure, BN parameter learning was applied to calculate the conditional probability table (CPT) and update the prior probability distribution of landslide-

causing factors in the existing BN model. The expectation maximization (EM) algorithm, a maximum-likelihood estimation method, was adopted in BN parameter learning. Furthermore, Junction tree (JT), one of the more popular inference engine algorithms, was adapted for computing the marginal distributions of landslides.

The qualitative component of the BN model is a DAG with nodes representing the landslide-causing factors used in the LSA and links representing the direct causal influences between the linked nodes. The BN structure reveals the complexity of the relationships between the landslide-causing factors that form the basis of the BN model: the landslide is the root node of the BN structure, and each landslide-causing factor is directly linked to it. Most of the links between the landslide-causing factors contain two variables connected in the following order: lithology and AI.

To conclude, eight landslide-causing factors were chosen as the independent variables for BN modeling. The study shows that lithology and Arias intensity are the major factors affecting landslides in the study area. On the basis of a posteriori probability distribution, the occurrence of a landslide is highly sensitive to relief amplitudes above 116.5 m. Using a 10-fold cross-validation and a receiver operating characteristic (ROC) curve, the resulting accuracy of the BN model was determined to be 93%, which demonstrates that the model achieves a high probability of landslide detection and is a good alternative tool for landslide assessment.

From the model results, I can see that BN model works well on the chosen dataset. However, it is not clear whether other machine algorithms would achieve the same or better performance than the current method. The structure of BN tree model has a similar concept as Neural Network, so I was hoping to see if it's applicable to compare these two algorithms on the same dataset. Also, 8 out of 11 parameters were selected to build the model, which may seem insufficient to draw a conclusion on such topic (I believe domain experts should be consulted on such topic). The last thing I'm skeptical about is model generalization. The studied data discussed in the paper is an area located in China with 35.6km^2 rasterized into 25 m × 25 m grid. Researchers didn't mention the size of transformed data, i.e. how many rows in the table. If we use the model to predict somewhere else in China or locations in other continent, whether the high accuracy (93%) be maintained is questionable.