Comparative analysis of machine learning algorithms in water extraction

Aimin Li¹, Meng Fan²*, Guangduo Qin²

¹ School of Geo-Science and Technology, Zhengzhou University, Zhengzhou 450001, Henan Province, China;
² School of Water Conservancy Engineering, Zhengzhou University, Zhengzhou 450001, Henan Province, China

*Corresponding author’s e-mail: 2594809931@qq.com

Abstract: There are many traditional methods available for water body extraction based on remote sensing images, such as normalised difference water index (NDWI), modified NDWI (MNDWI), and the multi-band spectrum method, but the accuracy of these methods is limited. In recent years, machine learning algorithms have developed rapidly and been applied widely. Using Landsat-8 images, models such as decision tree, logistic regression, a random forest, neural network, support vector method (SVM), and Xgboost were adopted in the present research within machine learning algorithms. Based on this, through cross validation and a grid search method, parameters were determined for each model. Moreover, the merits and demerits of several models in water body extraction were discussed and a comparative analysis was performed with three methods for determining thresholds in the traditional NDWI. The results show that the neural network has excellent performances and is a stable model, followed by the SVM and the logistic regression algorithm. Furthermore, the ensemble algorithms including the random forest and Xgboost were affected by sample distribution and the model of the decision tree returned the poorest performance.

1. Introduction

When extracting water bodies using traditional methods, a spectral water index extracted based on spectral signals of water bodies and other ground objects is mostly utilised. These methods perform well in the extraction of large-area water bodies but fail to guarantee the accuracy of extraction of small water bodies from a small range and rivers in cities with complex surroundings [1]. Based on the water index, many scholars have developed a series of methods, such as the methods based on texture features of images [2], object-oriented methods [3], and methods based on mathematical morphology. The classification accuracy of remote sensing images directly affects application and development of remote sensing technologies. These problems are difficult to solve by traditional methods alone, so a large number of scholars have been trying, improving, and exploring new methods in recent years to improve the accuracy and speed of image classification. With the development of computer information technologies, many machine learning algorithms for image classification and object identification have emerged. These methods are mainly utilised in classification of remote sensing images [4], water body identification and extraction [5], and prediction of water quality [6] in the water conservation industry.

This study involved an experiment aimed at water body extraction by using a multi-hidden-layer neural network, a decision tree, a support vector machine (SVM), and logistic regression as well as ensemble algorithms, such as the random forest and Xgboost algorithm mostly used in current research.
to analyse extraction effects. Furthermore, their accuracies were compared with those of the commonly used methods for water body extraction. The research results can provide a reference for future work on water body extraction based on remote sensing.

2. Data and pre-processing

2.1 Data
Landsat-8 data from the website (http://glovis.usgs.gov/) of the United States Geological Survey were used. Two images from the study area with a row-column number of 123-039 taken on 20 October and 4 October 2019 were separately used for establishing models, conducting a comparative experiment of the algorithms and testing applicability of the results. All original data were processed through radiometric calibration to convert the original digital number (DN) value into radiance.

2.2 Pre-processing
By combining waveband combinations 7/5/4, 7/4/3, 6/5/4, and 4/3/2 with visual interpretation, a sample dataset was selected from Landsat images for classification, in which there were 340 and 454 sample points for water bodies and other ground objects, respectively.

The characteristics, such as presence of a large correlation between multiple wavebands in the original images and similar information and structures between different wavebands, generally brought significant amounts of redundancy to the data. For this reason, principal component analysis (PCA) for dimensionality reduction was applied to remove repetitive and redundant information between various wavebands [7]. The first and second principal components in the PCA with a cumulative variance contribution of 99% were selected as classification characteristics.

Based on the PCA, four generally used texture features, i.e. contrast, autocorrelation, dissimilarity and entropy were extracted. Considering the factors including correlation between ground objects and resolution of images, in this study we set the distance to 1 and selected a 3 × 3 window with four orientations of 0°, 45°, 90°, and 135°.

After determining size and window parameters, J-M distance [8] and transformed divergence (T-D) in many extracted texture features were used for studying the separability of ground objects, to determine the characteristics ultimately used for classification. After calculation, it was found that the heterogeneity characteristics of the second component of the principal component had the best distinguishing ability. Therefore, in later classifications, the dissimilarity of the second component in the four directions together with the first component (PCA1) and the second component (PCA2), a total of six characteristics, were selected.

3. Research methods
Firstly, the performance of machine learning algorithms given different sample sizes was discussed. During this process, the optimal parameters of the models were determined and the indices, such as precision and AUC were used to evaluate performances of the algorithms in a test set. Secondly, by selecting thresholds based on water indices constructed according to spectral characteristics, water bodies and other ground objects were classified and identified. Moreover, through use of machine learning methods, such as SVM, decision tree, and random forest, water bodies were extracted. Thirdly, the extraction results were compared and analysed.

3.1 MNDWI
By using the modified NDWI (MNDWI) [9], the water bodies were binarized by selecting an appropriate threshold with which to achieve classification effects. To eliminate such influences, three methods for determining thresholds were used for comparison and discussion. The three thresholds included determining a threshold through multiple artificial tests by observing the range of greyscales of ground objects, thus determining a global threshold by Otsu algorithm. For the adaptive threshold, by setting the 3 × 3 template and using the method of taking the mean greyscale values in the template as a global
3.2 Machine learning algorithms

In this research, six machine learning algorithms were selected, all of which used the same group of sample set, and divided the total samples into a training set and a test set according to the ratio of 7:3. Furthermore, in the process of model training, the relevant parameters of the models were further trained by using 10-fold cross validation with hierarchical sampling of the training set. Finally, they were tested in the same test set and the indices, such as accuracy, recall rate, and AUC were utilized to assess the results.

(1) SVM

SVM has a simple structure and a strong generalisation ability making it easy to solve problems, with high-dimensionality, small sample sizes, and uncertainty \(^{[10]}\). In this study, the Gaussian radial basis function was selected as the kernel function. By using the grid search method in combination with 10-fold cross validation, the optimal parameters were determined as \(C = 3\) and \(\gamma = 0.003\).

(2) Decision tree

The decision tree determines the categories of the samples in the dataset by assigning the sample data to a certain leaf node. A classification and regression tree (CART) algorithm was used in this research, in which pre-pruning was utilised to avoid the overfitting problem. By fixing a single, or multiple, variables, the appropriate ranges of each parameter were determined. By using the grid search method and 10-fold cross validation, the final parameters were determined as follows: the entropy was selected as the purity index and the maximum depth was 7. The least sample size of separable leaf nodes was 8 and the minimum sample size of leaf nodes was 1.

(3) Multi-hidden-layer neural network

The neural network uses specific learning algorithms to learn from data through many learning algorithms. The input characteristics are passed to the next layer of nerve cells through a non-linear activation function and then continue to be passed down after activation of the nerve cells in this layer. That process is repeated and cycled to the output layer. The repeated superposition of these non-linear functions ensures that the neural network has sufficient non-linear fitting ability, while different activation functions can affect the output of different neural networks. By selecting a sigmoid activation function, it was determined that the neural network structure should have four layers based on multiple tests through cross validation. Except for input and output layers, the numbers of nerve cells in the two hidden layers were eight and six, respectively.

(4) Random forest

The random forest is an ensemble method specially designed for a decision tree classifier, and the selection of random attributes is further added to its training process. Using similar parameters to those used for the decision tree, the random forest model is easy to implement and shows good effects. In this research, its parameters were determined by using cross validation and grid search methods. When constructing the main parameters of the random forest, there are 10 weak estimators in the decision tree and the maximum depth was 4. Moreover, a Gini function was selected as the purity index.

(5) Xgboost

The core of Xgboost is an ensemble algorithm based on the gradient boosting decision tree (GBDT), that is, it can be used for classification or regression problems. Its modelling process is as follows: a decision tree is built and one more tree is added upon each iteration to form a strong evaluator integrating many numerical models. The accuracy is superior to that of a weak estimator and its calculation speed and performance are good. The main parameters are set as follows: the maximum depth of each tree is 3 and a weak classification estimator with 300 decision trees is established. The learning rate is set to be 0.01.

(6) Logistic regression algorithm

The logistic regression is a form of classification model. It establishes a regression formula for samples and a sigmoid function is used for classification, with simple principles governing its use.
4. Experiment and analysis

4.1 Effects of the sample size on learning algorithms

It is assumed that the samples selected by visual interpretation are reliable, namely, the various classes of the sample points are assigned to correct labels. Based on this, a small sample is randomly selected from the training set and divided into a training set and a validation set in the proportion of 7:3. By using the accuracy of the validation set of the small sample as an evaluation index, the effects of the sample size on classification effects of each algorithm are discussed, so as to judge whether the sample size selected is sufficient to achieve the purpose of the training model.

As demonstrated in Fig. 1, the accuracies of the classification algorithms in the validation set in the experiment all tend to increase with the sample size, and show a smaller error relative to the accuracy in the training set. Moreover, the accuracies gradually tend to be equal. This indicates that there is almost no underfitting of the samples and the parameters of each algorithm are well adjusted. The accuracy of the logistic regression algorithm is improved rapidly, approximating to the accuracy in the training set when the sample size is small, suggesting that there is almost no overfitting. As the sample size increases, the accuracy stabilises; however, other classification algorithms need larger samples to achieve this stability, and the accuracy fluctuates (albeit within a small range), therefore, the number of training samples selected in the experiment can meet the needs of model training.

![Fig. 1 Effects of the sample size on performance of each algorithm](image)

4.2 Comparative analysis of NDWI and machine learning algorithms

Two areas with dimensions 616 × 616 were selected on the Landsat-8 image taken on 20 October 2019. Area 1 has large-area water bodies and simple types of ground objects, without characteristic clutter arising from building and hill shadows. In Area 2, two completely different areas are selected to compare and analyse the generalisation abilities of each algorithm in the presence of complex ground objects and hill shadows.

As displayed in Fig. 2, in the classification methods that involved setting thresholds based on water indices, the Otsu algorithm obtained the best effect, however, significant salt and pepper noise signals were visible when using the other two methods. Under simple ground objects and large-area water bodies, the visual interpretation effects of each machine learning algorithm were similar, while only a slight discrepancy was found at the edge due to the influences of adjacent ground objects. As indicated by the red circles in the figure, compared with the machine learning algorithms, the traditional Otsu algorithm can be used to identify small-area scattered water bodies, but it also introduces other noise.
signals that are caused by non-water objects. Moreover, it is difficult to distinguish between them, however, the machine learning algorithms only identify certain small-area water bodies and reduce noise, which is attributed to the selection of the samples. This is because, when selecting samples, the accuracy of attributing categories of the samples is considered, which reduces the selection of small-area scattered water samples.

Fig. 2 Performance of each algorithm in Area 1

In the presence of complex ground objects and hill shadows, significant differences are found between algorithms. Fig. 3 shows that the decision tree and MNDWI based on custom thresholds and adaptive thresholds have lost the ability to distinguish water bodies in Area 2. Nevertheless, the Otsu algorithm introduces much unwanted noise due to its ability to over-identify a small range of scattered water bodies. Meanwhile, the Otsu algorithm, and the logistic regression and SVM among the machine learning algorithms trialled here, are likely to generate more misclassifications (marked in blue circles), while the random forest and Xgboost have obvious advantages over a single decision tree, but generate some missed classifications (marked in green circles). However, fewer misclassifications and less noise are produced compared with the other algorithms. In addition, the neural network generated fewer misclassifications, less noise, and rarely presented a missed classification. To summarise, in the presence of complex ground objects, the neural network has better effects, followed by the SVM, and logistic regression. The effects of the ensemble algorithms including the random forest and Xgboost are better than the decision tree, while performing moderately in more complex areas. The Otsu algorithm is superior to the other two methods for determining thresholds but generates significant noise.

Fig. 3 Performance of each algorithm in Area 2
5. Conclusions
Based on Landsat-8 images, machine learning algorithms, such as: the decision tree, logistic regression, random forest, neural network, SVM, and Xgboost were used for water body extraction and compared and analysed using three methods for determining thresholds in traditional MNDWI. The following conclusions were drawn:

(1) Several machine learning algorithms have different requirements on sample size. The neural network performs best on the images taken under various conditions and is a stable model, followed by the SVM and the logistic regression algorithm. The ensemble algorithms including the random forest and Xgboost with their excellent performance in classification of the original remote sensing image for collecting training samples are affected to a significant extent by the sample distribution and demonstrate large discrepancies in the images acquired under different conditions.

(2) The MNDWI based on the threshold determined by Otsu algorithm can identify water bodies while meeting bimodal conditions in the histogram of greyscales of surface features; however, it is generally difficult to meet bimodal conditions in different remote sensing images and areas. Therefore, the method for classifying water bodies by constructing water indices to select the threshold is unstable.

(3) Owing to the complex distribution of ground objects and many influential factors in the remote sensing image classification, it is difficult to collect small and dispersed water bodies when collecting samples in this research. This limits the performances of the models in the environment with many hill shadows and complex ground objects. In the later stage, more samples should be collected from images acquired over different areas and periods to train the models, thus making the models more applicable.

Funding
The work is supported by the Joint Funds of National Natural Science Foundation of China (Grant no. U1704125).

References
[1] Wang X. B., Xie S. P., Du J. K. Water index formulation and its effectiveness research on the complicated surface water surroundings. Journal of Remote Sensing, 2018, 22(02): 360-372.
[2] Feng J. H., Yang Y. J. Study of texture images extraction based on the grey level co-occurrence matrix. Beijing Surveying and Mapping, 2007, 3: 19-22.
[3] Li T. Study on remote sensing image comprehensive evaluation method for image classification. Beijing Surveying and Mapping, 2019, 33(02):157-160.
[4] Cui L., Zhang P., Che J. Overview of deep neural network-based classification algorithms for remote sensing images. Computer Science, 2018. 45(S1): 50-53.
[5] Wang H., Qin F. Summary of the research on water body extraction and application from remote sensing image. Science of Surveying and Mapping, 2018. 43(05): 23-32.
[6] Sun T. et al. Overview of application of deep learning in water conservancy industry. China Rural Water and Hydropower, 2018, 12: 90-93+100.
[7] Zhao, G. and A. L. Maclean. A comparison of canonical discriminant analysis and principal component analysis for spectral transformation. Photogrammetric Engineering and Remote Sensing, 2000, 66(7): 841-847.
[8] Ma N. et al. Determination on the optimum band combination of HJ-1A hyperspectral data in the case region of Dongguan based on optimum index factor and J-M distance. Remote Sensing Technology and Application, 2010, 25(3): 358-365.
[9] Xu H.Q. A study on information extraction of water body with the modified normalized difference water index (MNDWI). Journal of Remote Sensing, 2005, 5: 589-595.
[10] Pal, M. and P. M. Mather, Support vector machines for classification in remote sensing. International Journal of Remote Sensing, 2005, 26(5): 1007-1011.