Use of classification algorithms for the ice jams forecasting problem

Natalia Semenova[1*], *Alexey* Sazonov[1,2], *Inna* Krylenko[1,2], and *Natalia* Frolova[1]

[1] Lomonosov Moscow State University, GSP-1, Leninskie Gory, 119991, Moscow, Russia
[2] Water Problems Institute of the Russian Academy of Science, Gubkina st., 3, 119333, Moscow, Russia

Abstract. In the research the prediction of occurrence of ice jam based on the K Nearest Neighbor method was considered by example of the city of Velikiy Ustyug, located at the confluence of the Sukhona and Yug Rivers. A forecast accuracy of 82% was achieved based on selected most significant hydrological and meteorological features.

1 Introduction

Floods gain a lead among natural disasters both in terms of area of distribution and damage caused for Russia. Flooding can be caused by snow cover melting, a large amount of precipitation, the effects of surges, a breakthrough of a dam, etc. For northern rivers, including rivers of the European part of Russia, ice jams often cause floods.

The goal of this research is developing a methodology for predicting the occurrence of ice jam based on the machine learning method. The place of confluence of the Sukhona and Yug Rivers, where the city of Velikiy Ustyug is located, was chosen as the object of study. The probability of the ice jams formation in this area is 43.5% according to statistics. Their occurrence leads to an increase of water level and flooding of residential and utility buildings.

2 Data and methods

Over the past two decades, there has been a huge leap in the development of computer technology and machine learning, which has allowed the application of various machine learning algorithms to a large number of applied problems, including the prediction of flood characteristics. There are many different algorithms used for classification problem in machine learning theory. The following algorithms can be noted among them: probabilistic, metric, logical algorithms and neural networks. Each of these methods uses its own mathematical apparatus, which effectiveness directly depends on the task. For example, probabilistic methods and neural networks require a sufficiently large number of observations to obtain statistically significant results. In the absence thereof, metric and logical approaches are usually used.

[*] Corresponding author: snkone132@mail.ru

In the research, the metric KNN method (K Nearest Neighbor [2, 3]), based on the hypothesis of compactness and proximity of similar objects, was chosen as an algorithm for predicting of an ice jam emergence. The advantages of this method are resistant to outliers, ease of implementation, interpretability and the ability to work with small data.

2.1 Source data for forecasting

Data from the following 6 hydrological gauges was used for the features constructing: Velikiy Ustyug, Kotlas, Totma, Kalikino v., Berezovaya Slobodka v., Podosinovets set.; and weather data was used from 3 weather stations: Nyuksenitsa v., Nikolsk and Velikiy Ustyug. The choice of weather stations was due to the fact that two of them are located exactly on the Sukhona River (the village of Nyuksenitsa) and the Yug River (the city of Nikolsk), and the weather station in Velikiy Ustyug is precisely at the confluence of these two rivers.

Two groups of features were constructed from the initial data (hydrological for each of 6 gauging stations (42 in total) and meteorological for 3 stations (21 in total), as it is shown in Table.

Table. List of hydrological and meteorological features.

Feature number	Feature name	Characteristic type, measuring unit
1	Water level before freeze-up	Hydrological feature, cm
2	Water level of the first ice phenomena appearance	Hydrological feature, cm
3	Duration of freezing	Hydrological feature, day
4	Duration of autumn ice drift	Hydrological feature, day
5	Maximum ice thickness	Hydrological feature, cm
6	Number of days before opening	Hydrological feature, number of days from February 1, day
7	Water discharge a day before opening	Hydrological sign, $м^3/s$
8	Features of temperature regime during the freezing period	Meteorological sign, temperature transition through 0 ° C, number of days from September 1, day
9	Features of the temperature regime during the break up period	Meteorological sign, temperature transition through 0 ° C, number of days from February 1, day
10	Sum of negative values of air temperature during the cold period	Meteorological sign, ^oC
11	Sum of positive temperature values for the cold period	Meteorological sign, ^oC
12	Total precipitation	Meteorological sign, mm
13	Total amount of solid precipitation	Meteorological sign, mm
14	The number of days with a positive temperature for the cold period	Meteorological sign, days

The obtained features form the base of observations from 1960 till 2016 years. Each year of observations is characterized by a vector consisting of meteorological and hydrological features (feature vector) that have numerical values.

Two scenarios for forecasting were identified:

- the presence of jam-induced rise of the water level at the confluence of the Sukhona and Yug Rivers in the area of Velikiy Ustyug;
- absence of jam-induced rise of the water level in the area of Velikiy Ustyug.

These scenarios define the classes of the classification problem. In the future, an increase of the number of observations years will expand the number of scenarios. For example, it will be possible to predict an ice jam power, determined by the maximum height of the ice jam level. The following classes of ice jams are distinguished: small (less than 1.5 m) medium (1.5-2.5 m), powerful (2.5-3.5 m), catastrophic (rise more than 3.5 meters) [4, 5].

2.2 Forecasting Methodology

The K Nearest Neighbors method refers to metric classifiers. It is based on the assumption that similar, having compatible features objects will belong to the same class. To determine the concept of "proximity" of two objects, a distance binary function ϱ (**x, y**) (metric) is introduced, as a function of two feature vectors $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ of these observations. The following metrics are considered as classical metrics in machine learning problems:

- Euclidean metric

$$\varrho(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

- Manhattan distance

$$\varrho(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \quad (2)$$

- Chebyshev metric

$$\varrho(\mathbf{x}, \mathbf{y}) = \max_{i=1\dots n} |x_i - y_i| \quad (3)$$

The algorithm calculates the distances from the classified object to each known observation, finds k objects with the minimum distance to the classified observation and according to the “vote” procedure – the classified object considered belonging to that class, which contains most of the neighbors of this element.

With such a classification, the distance plays a role only in the selection of k close objects, but it does not matter for voting procedure. One of the possible modifications of this algorithm is weighted voting. With this modification, the distance to the new record is also taken into account. Thus, the closer the neighbor is to the classified object, the higher coefficient its vote gets.

Accuracy (the proportion of correct answers) was used as the main metric for assessing of the model quality. The recall metric was also additionally considered. Recall shows which part of the relevant objects were selected by the classifier. In our case, which part of the ice jam years was correctly recognized by the algorithm.

2.3 Selection of the algorithm optimal parameters

The operation result for the nearest neighbors algorithm for each specific task depends on the choice of model parameters: the number of neighbors k, the distance metric between objects ϱ (**x, y**), and the presence of voting weights. The model quality is calculated for each set of parameters and the best combination of values is selected to choose the best model parameters.

In order to avoid the effect of retraining when selecting hyperparameters, the K-fold cross validation method is used to evaluate the algorithm quality. The training sample is randomly divided into K parts. The algorithm is sequentially trained on a subsample consisting of K-1 parts, and the quality of the model calculated for suspended part (E_i). The final quality of algorithm (E) for a given set of hyperparameters is considered as the average between the quality values obtained by cross-validation.

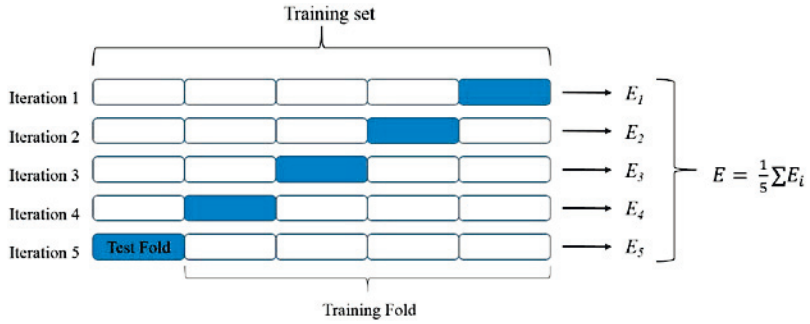

Fig. Illustration of the cross-validation method for the case K = 5.

3 Construction of the (forecast) model

The initial data were divided into training (1960-1999) and test (2000-2016) samples. Only the training sample was involved for selecting of the model parameters. The test sample was used for subsequent validation of the resulting model. In order to improve the quality of the forecast, additional data processing was carried out.

3.1 Data preparation

3.1.1 Feature selection

Since the observation duration is 57 years, and there are too many features, in order to prevent the occurrence of excess noise and retraining, the 10 most significant features were selected (1 – Berezovaya Slobodka village, Kalikino village; 3 – Berezovaya Slobodka village; 4 – Velikiy Ustyug, the city of Totma; 5 – Kotlas, Velikiy Ustyug; 8 – the village of Nyuksenitsa, Nikolsk, Velikiy Ustyug; Table). The influence of each feature on the occurrence of ice jam was studied using the WoE analysis (Weight of Evidence [6]), which allows to capture both linear and nonlinear dependencies.

3.1.2 Feature transformation

Metric classification algorithms are very sensitive to data scale. Initial features may belong to and vary in different ranges, making a different contribution to the metric. An imbalance

between the feature values can cause instability of the model and degrade its quality. For example, features with large numerical values may become predominant, while the contribution of features with small values to the metric will be very small, and such features will simply not be taken into account when building the model. In order to avoid such a situation, features must be normalized. In this task, MaxAbsScaler tool was used for this purpose. According to this normalization, each feature is scaled by its maximum absolute value, thereby the range of change of each feature is transformed into the range [-1; 1].

3.2 Model building and model verification

The model parameters were selected using cross-validation for the training sample after data preprocessing. The best parameters appear to be Euclidean metric with 5 neighbors without weights. The average quality score for the training sample was 79%.

The corresponding model was trained on the training sample after selecting the optimal parameters, and the final model was checked. The number of correctly classified years was 82%.

4 Conclusions

Model verification gives 82% forecast accuracy, and it is worth to note that the model mistakes concerns only the years when there was no ice jam, the years of its presence were given correctly. For all the largest floods of the ice jam genesis of the 21st century (2013, 2016), the model showed the correct result.

The study was supported by the RFBR grant No. 18-05-60021 Arctic.

References

1. S.A. Agafonova, N.L. Frolova,Wat. Res., **34**(2), 123 – 131 (2007).
2. D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining.* MIT Press, Cambridge (2001)
3. Z. Zhang, Ann. Transl. Med., **4**(11), 1-7 (2016)
4. V. Buzin, *Gaps and ice jams on the rivers of Russia* (2015)
5. R. Donchenko, *Ice regime of the USSR rivers*, p. 248. (1987)
6. D. Weed., Risk Anal., **25**(6), 1545 – 1557 (2005).