Advances in Computer and Communications, 2022, 3(2), 74-76
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Ice Jam Prediction for Sukhona River Based on KNN and Decision Tree Algorithm

Yuxuan Cui

Lomonosov Moscow State University, Moscow, 119991, Russian Federation.

How to cite this paper: Yuxuan Cui. (2022) Ice Jam Prediction for Sukhona River Based on KNN and Decision Tree Algorithm. *Advances in Computer and Communication*, 3(2), 74-76.
DOI: 10.26855/acc.2022.12.004

Received: October 28, 2022
Accepted: November 25, 2022
Published: December 30, 2022

***Corresponding author:** Yuxuan Cui, Lomonosov Moscow State University, Moscow, 119991, Russian Federation.

Abstract

Prediction of ice jam is very important for reduction and prevention of ice jam floods in cold regions. This article focuses on the assessment of the possibility of predicting ice jam on the Sukhona River in Russia based on selected most significant hydrological and meteorological features. The maximum water level during the ice drift and ice-jam induced water level rising are the main determinants. The optimal prediction model is developed based on KNN algorithm with decision tree algorithm. The model built by the KNN algorithm was found to perform best and accurately found all blockage years. The research in this paper provides help to establish ice jam prediction in the Veliky Ustyug region. The knn method has a recall of 1 in the studied river segment, which predicts the occurrence of ice jam more accurately than other prediction methods. This implies that the chosen forecast factor is highly reliable.

Keywords

Ice jam, KNN, Decision Tree Method, multiple linear regression mode

1. Introduction

The Sukhona River is a tributary of the North Dvina River in the European part of Russia. It is 558 km long and originates from Lake Kubena in the Vologda Region and flows northeastward. In the upper reaches of the river, the channel is gentle and the river bank is wide, while in the lower reaches the valley is steep and narrow. The spring flood of the Sukhona River lasts for about three months, then the water level continues to drop and enters the flat water period in July. The ice thickness in the downstream severe winter averages 1 m. It is prone to flooding caused by ice jam during the spring drift ice period [1].

The studied river segment is located in the lower Sukhona River, north of the city of Kalikino and south of the city of Kotlas. This river section has warm and short summers and long, severe winters with continuous snow cover and an average temperature of -14°C in January [2, 3]. In winter the temperature drops sharply due to the invasion of cold air masses from the Arctic, with the lowest temperature dropping to -46°C. At the same time the channel of the lower Sukhona River narrows, the slope of the riverbed increases, and in the city of Velikiy Ustyug the width of the river is 500 m. The narrow channel and rapidly decreasing temperatures caused severe ice jam in this section of the river [4].

2. Method

In the research, the metric KNN method was chosen as the algorithm for predicting the appearance of ice plugs based on the assumption of compactness and proximity of similar objects. The advantages of this method are resistance to outliers, ease of implementation, interpretability, and ability to handle small data. The prediction results are also compared with decision tree algorithms models [5, 6].

Figure 1. Location of the Sukhona River basin and study reaches.

2.1 Prognostic factor analysis

Factors affecting ice jams include the maximum water level prior to the ice phenomenon, the cumulative number of days from September 1 until the temperature is less than zero, the duration of ice phenomena, the maximum ice thickness, and the river flow during spring floods. In this article, data from three meteorological stations and five hydrological stations in the study river segment were collected from 1936 to 2020. The obtained features are presented in Table 1.

In machine learning it is common to divide the data into training and test sets to evaluate the model prediction effect. In this article, data from 2000 to 2018 are designated as the test set, and the training set is set to data from 1960 to 1999. The target variables were set to 0 (no overwash) and 1 (overwash).

Table 1. List of hydrological and meteorological features

Feature number	Feature name	City	Characteristic type, measuring unit
1	Maximum water level before freeze-up	Berezovaya Slobodka	Hydrological feature, cm
2		Kalikino	
3	Duration of drift ice	Velikiy Ustyug	Hydrological feature, day
4		Totma	
5	Duration of ice cover	Berezovaya Slobodka	Hydrological feature, day
6	Maximum ice thickness	Velikiy Ustyug	Hydrological feature, cm
7		Kotlas	
8	Features of temperature regime during the freezing period	Velikiy Ustyug	Meteorological sign, temperature transition through 0 ° C, number of days from September 1, day
9		Nyuksenitsa	
10		Nikolsk	

2.2 Feature transformation

Metric classification algorithms are very sensitive to the size of the data. The initial features can belong to and vary over different ranges, contributing differently to the metric [7]. Imbalance between feature values can lead to instability and reduce the quality of the model [8]. To avoid this situation, features must be normalized. In this task, the MaxAbsScaler tool is used for this purpose. According to this normalization, each feature is scaled by its maximum absolute value, thus transforming the range of variation of each feature into a range of (-1; 1).

2.3 Comparison of KNN algorithm model and decision tree algorithm model

KNN algorithm and decision tree algorithm are one of the most basic and simple machine learning algorithms for classification problems. KNN algorithm finds the training sample that is most similar to the test sample and classifies it. To choose the appropriate parameter values, we use cross-validation method to find the k-values. The cross-validation method prevents overlearning by dividing the training set into n equal subsets, training the model with n-1 subsets, and testing it with the remaining subsets. The process is repeated n times and the average of the n results is taken. The GridSearchCV function in the sklearn library is called in python to find the best parameters. The best model obtained by GridSearchCV in the knn algorithm is k=5 with a uniform weight function. Sample accuracy of 81% is achieved under this model.

The decision tree algorithm combines logical conditions into a tree structure with a hierarchical structure consisting of a root node, decision nodes and terminal nodes. The same GridSearchCV function is used to find the optimal parameters, for the decision tree algorithm we get the best model parameters for a maximum tree depth of 3 and a maximum number of features of 3. but the accuracy of this model is only 43.8%.

3. Result and discussion

For the machine learning approach, we evaluate the best model obtained with more metrics, besides the most common accuracy metrics, precision and recall are also important. KNN model metric is Minkowski distance with a k value of 5 and a recall of 1 for a model with uniform weights, which means that the model correctly finds all years of ice congestion. The decision tree model, on the other hand, does not perform well, which is caused by the fact that fewer years of observation have changed the sensitivity of the model to outliers.

4. Conclusions

(1) According to the analysis of historical book data, the water level in the Grand Usk section of the Sukhona River exceeding 700 cm is associated with the formation of ice jam floods. The incidence of ice jam floods from 1936 to 2018 was 64%, occurring every three to four years. Of concern is the decreasing trend of ice thickness in this river section due to global climate effects.

(2) In this paper, we compare KNN model and decision tree model. The KNN algorithm achieves better results, and the most suitable model for this area is obtained by cross-validation with K=5, the weight is uniform, and the distance metric is Minkowski distance. The accuracy of this model is 81% and the recall rate is 1, which means that the model correctly finds all ice blockages.

References

[1] Buzin V.A. Ice jams and jamming floods on rivers, 2004.

[2] Buzin V.A. Dangerous hydrological phenomena, 2008.

[3] Sazonov A.A. Probable scenarios and calculation of the characteristics of flooding of residential areas based on a set of mathematical models, 2021.

[4] Du, W. Application of support vector regression in prediction model using genetic algorithm optimized. Journal of Physics: Conference Series. IOP Publishing, 2021.

[5] Beltaos, S. Assessing Ice-Jam Flood Risk: Methodology and Limitations, 2021.

[6] Vanantwerp R.L. Ice jam flooding: causes and possible solutions, 1994.

[7] Yang, Z. The research on the RS Dynamie Monitoring Mode Of The Yellow River Icicle Hazard And The Ice Regime Information Extraction Model, 2006.

[8] Hu, W. Research on product recommendation system based on deep learning. Фундаментальные и прикладные научные исследования: актуальные вопросы, достижения и инновации. 2022.