Oceanologia (2017) xxx, xxx—xxx

Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.journals.elsevier.com/oceanologia/

OCEANOLOGIA

ORIGINAL RESEARCH ARTICLE

A fuzzy KNN-based model for significant wave height prediction in large lakes

Q1 Mohammad Reza Nikoo [a,*], Reza Kerachian [b,1], Mohammad Reza Alizadeh [a,2]

[a] *School of Engineering, Shiraz University, Shiraz, Iran*
Q2 [b] *College of Engineering, University of Tehran, Tehran, Iran*

Received 20 May 2017; accepted 28 September 2017

KEYWORDS
Significant wave height prediction;
Fuzzy *K*-nearest neighbor;
Bayesian networks;
Support vector regression;
Regression tree induction

Summary Some algorithms based on fuzzy set theory (FST) such as fuzzy inference system (FIS) and adaptive-network-based fuzzy inference system (ANFIS) have been successfully applied to significant wave height (SWH) prediction. In this paper, perhaps for the first time, the fuzzy *K*-nearest neighbor (FKNN) algorithm is utilized to develop a fuzzy wave height prediction model for large lakes, where the fetch length depends on the wind direction. As fetch length (or wind direction) can affect the wave height in lakes, this variable is also considered as one of the inputs of the prediction model.

The results of the FKNN model are compared with those of some soft computing techniques such as Bayesian networks (BNs), regression tree induction (named M5P), and support vector regression (SVR). The developed FKNN model is used for SWH prediction in the western part of Lake Superior in North America. The results show that the FKNN and M5P model can outperform the other soft computing techniques.

1. Introduction

Significant wave height (SWH) is an important hydrodynamic variable for design and operation of coastal and offshore structures. A successful prediction model should be able to accurately predict the wave parameters in different conditions. Review of related literature about wave parameters estimation models shows that the application of data mining

* Corresponding author at: School of Engineering, Shiraz University, Shiraz, Iran. Tel.: +98 713 6133497; fax: +98 711 6473161.
E-mail addresses: nikoo@shirazu.ac.ir (M.R. Nikoo), kerachian@ut.ac.ir (R. Kerachian), alizadeh.mohamadreza@yahoo.com (M.R. Alizadeh).
[1] Tel.: +98 912 5339529.
[2] Tel.: +98 917 3061374.
Peer review under the responsibility of Institute of Oceanology of the Polish Academy of Sciences.

Production and hosting by Elsevier

https://doi.org/10.1016/j.oceano.2017.09.003
0078-3234/

and artificial intelligence models is a promising alternative to effort demanding and time consuming numerical or physical wave estimation tools (Balouchi et al., 2015; Etemad-Shahidi and Mahjoobi, 2009; Malekmohamadi et al., 2008; Nikoo et al., 2015; Özger, 2010).

In the past decade, in order to tackle the limitations of numerical and empirical models and eliminate the prior knowledge requirement about interactions among inputs, parameters, and outputs in SWH prediction problems, different soft computing techniques have successfully been applied. Artificial neural network (ANN), fuzzy inference systems (FIS), adaptive-network-based fuzzy inference system (ANFIS), Bayesian network (BN) and genetic programming (GP) are the most common methods, which have been used for SWH prediction. For example, FIS and ANFIS methods have been used by Özger and Şen (2007), Mahjoobi et al. (2008), Zanaganeh et al. (2009), Malekmohamadi et al. (2011), Patil et al. (2012), Karimi et al. (2013), Nikoo et al. (2014) and Kazemi Elaki et al. (2016) for estimating wave parameters. Altunkaynak and Wang (2012) developed a model based on Geno Kalman Filtering (GKF) for predicting SWH in the Lake Okeechobee, Florida. They compared the results of the GKF and an ANN and concluded that the GKF can provide better predictions. Nikoo and Kerachian (2017) developed an Artificial Immune Recognition System (AIRS) and five data mining techniques (i.e. ANN, SVR, BN and Rough Set Theory (RST)) for SWH predictions with different time lags. The AIRS model was applied for the SWH prediction in the Lake Superior in North America and its results were compared with those of other models. The results showed that AIRS and ANN models outperform other data-driven models in forecasting extreme SWHs. In a more recent study (i.e. Berbić et al., 2017), the real-time estimation of wave height within the successive 30 min up to 5.5 h time span was achieved using ANN and support vector machine (SVM) and later they used the least error-prone model and investigated the effect of taking into account the wind velocity on the accuracy of wave height prediction.

In this paper, fuzzy K-nearest neighbor (FKNN), which was initially introduced by Keller et al. (1985), is utilized for SWH prediction. The FKNN is a branch of supervised learning methods, which particularly account for the similarity between the training sample data and the test data. The main advantage of using the FKNN is that despite its deterministic form (KNN), the unbiased weighting of instances is considered in the decision rule regardless of its distance to the pattern to be classified, and thus the fuzzy form has a greater accuracy (Derrac et al., 2016). One disadvantage of the FKNN is that it is computationally expensive. Since this model should be run for all data set, it is time consuming and requires large memory size to store all training data.

Reviewing the literature shows that FKNN has been successfully applied in different studies. For example, Xiao and Wang (2010) utilized an FKNN-based machine learning approach to provide a prediction system for protein quaternary structural type. Chen et al. (2011) developed a bankruptcy prediction model based on an adaptive FKNN algorithm. They optimized the neighborhood size K and the fuzzy strength parameter m using a Particle Swarm Optimization (PSO) approach. Wang and Xiao (2011) used FKNN to predict and classify the risk types of Human Papillomavirus (HPV) based on the bioinformatics analysis. Somari et al. (2016) utilized KNN and FKNN in particle contamination detection. Fredj and Ouni (2017) applied the crisp and fuzzy KNN algorithm for Timit phoneme recognition and their results showed that FKNN could provide better recognition.

The developed FKNN model for predicting SWH are compared with three soft computing models (i.e. SVR, BN, and M5P), which have been widely used in this context. All models are utilized for SWH prediction in Lake Superior and their results are compared using a set of statistical error indices, namely root mean square relative error (RMSRE), root mean relative error (RMRE), correlation coefficient (CC), bias and scatter index (SI). The differences between the used soft computing models have been visually plotted to make a better comparison of the models' performances.

2. Methods

As details of M5P, SVR and BN models are presented in Etemad-Shahidi and Mahjoobi (2009), Malekmohamadi et al. (2011), Etemad-Shahidi and Ghaemi (2011) and Abed-Elmdoust and Kerachian (2012), in the following section, the main characteristics of the FKNN algorithm are presented.

Classification algorithms are mainly used for measuring the similarity of a set of objects based on some measures of distance. The K-nearest neighbor (KNN) algorithm is one the oldest pattern classifier methods with no preprocessing requirement (Cover and Hart, 1967). The decision rule of common classification algorithms such as M5P, SVR and BN assume equal weights for object membership utilities, neglecting different patterns of similarity. Taking the advantage of fuzzy set theory (Zadeh, 1965), the FKNN has been shown to not only have a lower error in classifying the objects but also it has a greater confidence measure of the classification (Keller et al., 1985). The FKNN provides a more realistic vector of membership for the objects and it also accounts for the degree of object membership to the classes of objects. In this algorithm, a class is assigned to the most common class considering its K-nearest neighbors. FKNN assigns fuzzy membership to the samples and helps decision makers for fuzzy decisions (Chen et al., 2011; Keller et al., 1985).

The purpose of FKNN algorithm is to divide (cluster) the sample vectors $X=\{x_1, x_2, \ldots, x_n\} \subset R^n$ into c $(1<c<n)$ fuzzy subsets. For $i = 1, 2, \ldots, c$ and $j = 1, 2, \ldots, n$ the fuzzy membership matrix is shown by U, where U_{ij} is the fuzzy membership degree of x_j in class i. In a non-fuzzy version of the algorithm, the jth object is assigned to the ith class which has the greatest U_{ij} in comparison with the fuzzy membership degree of all the other classes. The matrix U has two constraints as follows (Keller et al., 1985):

$$\sum_{i=1}^{c} u_{ij} = 1, \quad "j, \tag{1}$$

$$u_{ij}\tilde{A}\,[0,1], \quad 0<\sum_{j=1}^{n} u_{ij}<n. \tag{2}$$

The first constraint (Eq. (1)), ensures that all the objects' membership degrees are obtained across all the classes ($i = 1, 2, \ldots, c$) and the summation of all the membership degrees

Please cite this article in press as: Nikoo, M.R., et al., A fuzzy KNN-based model for significant wave height prediction in large lakes. Oceanologia (2017), https://doi.org/10.1016/j.oceano.2017.09.003