Applied Ocean Research 30 (2008) 172–177

Contents lists available at ScienceDirect

Applied Ocean Research

journal homepage: www.elsevier.com/locate/apor

An alternative approach for the prediction of significant wave heights based on classification and regression trees

J. Mahjoobi [a,b], A. Etemad-Shahidi [b,*]

[a] Ministry of Energy, Water Research Institute, Hakimieh, Tehran, P.O.Box: 16765-313, Iran

[b] School of Civil Engineering, Iran University of Science and Technology, Narmak, Tehran, P.O. Box 16765-163, Iran

ARTICLE INFO

Article history:
Received 23 March 2008
Received in revised form
29 October 2008
Accepted 1 November 2008
Available online 11 December 2008

Keywords:
Wave prediction
Data mining
Machine learning
Classification
Decision trees
Regression trees
C5 algorithm
CART

ABSTRACT

In this study, the performances of classification and regression trees for the prediction of significant wave heights were investigated. The data set used in this study is comprised of 5 years of wave and wind data gathered from a deep water location in Lake Michigan. Training and testing data include wind speed and wind direction as the input variables and significant wave heights (H_s) as the output variable. To build the classification trees, a C5 algorithm was invoked. Then, significant wave heights for the whole data set were grouped into wave height bins of 0.25 m and a class was assigned to each bin. For evaluation of the developed model, the index of each predicted class was compared with that of the observed data. The CART algorithm was employed for building and evaluating regression trees. Results of decision trees were then compared with those of artificial neural networks (ANNs). The error statistics of decision trees and ANNs were nearly similar. Results indicate that the decision tree, as an efficient novel approach with an acceptable range of error, can be used successfully for prediction of H_s. It is argued that the advantage of decision trees is that, in contrast to neural networks, they represent rules.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Waves, the most significant maritime phenomenon, due to their complicated and stochastic behavior are known as one of the most difficult subject in coastal and maritime engineering practice. The effect of waves on coastal and marine activities urges us to identify the wave characteristics. Different approaches such as field measurements, theoretical studies and numerical simulations have been used for this purpose. Coastal and offshore engineers generally use these approaches to identify wave climate and extreme wave characteristics as well as annual attributes of waves. Different methods such as empirical, numerical and soft computing approaches have been proposed for significant wave height prediction.

Soft computing techniques such as artificial neural networks have been widely used to predict wave parameters [e.g. [1–5]]. A review of neural network applications in ocean engineering is given in [6]. Recently, other soft computing techniques such as the Fuzzy Inference System (FIS) and the Adaptive-Network-based Fuzzy Inference System (ANFIS) have been used to develop wave prediction models (e.g. [7,8]). These studies have shown that the wind speed is the most important parameter in wave parameter's prediction. Recently, Mahjoobi et al. [9] compared different soft computing methods such as Artificial Neural Networks (ANNs), the Fuzzy Inference System (FIS) and the Adaptive-Network-based Fuzzy Inference System (ANFIS) to hindcast wave parameters. Their results showed that the performances of these methods are nearly the same. Furthermore, using sensitivity analysis, they showed that the wind speed and direction are the most important parameters for wave hindcasting.

As mentioned before, the prediction of significant wave height that is essentially an uncertain and random process is not easy to accomplish by using deterministic equations. Therefore, it is ideally suited to decision trees since they are primarily aimed at the recognition of a random pattern in a given set of input values. Decision trees are helpful in predicting the value of the output of a system from its corresponding random inputs as the application of decision trees does not require knowledge of the underlying physical process as a precondition. Examples of decision tree applications include potential profit analysis of new drugs in pharmaceutical companies [10], medical diagnosis [11] and risk management analysis in petroleum pipeline construction [12]. However, to the authors' knowledge this method has not been applied in wave prediction. In this work, the performances of classification and regression trees as

[*] Corresponding author. Tel.: +98 21 7391 3170; fax: +98 21 77454053.
E-mail addresses: jmahjoobi@gmail.com (J. Mahjoobi), etemad@iust.ac.ir (A. Etemad-Shahidi).

0141-1187/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apor.2008.11.001

a new soft computing method for prediction of significant wave height were investigated. For building a classification tree, the C5.0 algorithm [13] is selected due to its speed, small memory requirement and boosting and cross-validation features. Since C5.0 can generate rules that have a straightforward interpretation, it is also quite robust in problems such as missing data and large numbers of fields. The CART algorithm [14] was employed for building and evaluating regression trees. CART builds classification and regression trees for predicting continuous (regression) and categorical predictor variables (classification). CART analysis has a number of advantages over other classification methods, including multivariate logistic regression [14]. First, it is inherently non-parametric. In other words, no assumption is made regarding the underlying distribution of values of the predictor variables. Thus, CART can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or non-ordinal structure. In this work, results of decision trees were also compared with those of artificial neural networks (ANNs) using statistical error measures. This paper is prepared as follows: Section 2 describes the decision trees, C5 and CART algorithms. Section 3 describes the study area and data set. Section 4 gives the description of results and statistical error analysis and Section 5 covers the summary and conclusions.

2. Decision trees

The patterns and relationships in data can be found using machine learning, statistical analysis, and other data mining techniques. Activities to discover hidden knowledge contained in data sets have been attempted by researchers in different disciplines for a long time. Through a variety of techniques, data mining identifies nuggets of information in bodies of data. Data mining extracts information in such a way that it can be used in areas such as decision support, prediction, forecast, and estimation. Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. Both disciplines have been applied in pattern recognition and classification. Decision trees are data mining methodologies applied in many real-world applications as a powerful solution to classification and prediction problems [15]. A decision tree is an arrangement of tests that prescribes an appropriate test at every step in an analysis. A decision tree is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a classification or decision. In general, decision trees represent a disjunction of conjunctions of constraints on the attribute-values of instances. Each path from the tree root to a leaf corresponds to a conjunction of attribute tests and the tree itself to a disjunction of these conjunctions. More specifically, decision trees classify instances by sorting them down the tree from the root node to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute [15,16]. The application of decision trees to classification was popularized in machine learning by Quinlan [17,18]. Quinlan's ID3 [17], is a well-known tree-growing algorithm for generating decision trees based on univariate splits. An extended version of this algorithm, called C4.5 [18] and its successor C5.0 [13] use Greedy search methods. They involve growing and pruning decision-tree structures and are typically employed in these algorithms to explore the exponential space of possible models. The algorithm basically chooses the attribute that provides the maximum degree of discrimination between classes locally. Theoretical concepts related to decision trees can be found in many text books (e.g. [15,16,19–21]).

2.1. C5 algorithm

C5.0 [13] is a commercial machine learning program developed by RuleQuest Research and is the successor of the widely used ID3 [17] and C4.5 [18] algorithms. A C5.0 decision tree is constructed using GainRatio. GainRatio is a measure incorporating entropy which measures how unordered the data set is. Entropy is denoted by the following equation:

$$Entropy(S) = \sum_{i=1}^{N} -P(S_i)\log_2(P(S_i)) \tag{1}$$

where, $P(S_i)$ is the probability of class i occurring in the data set S and N is the number of classes. Using the *Entropy*, it is possible to calculate the Information Gain (*Gain*).

Gain is a measure of the improvement in the amount of order given by:

$$Gain(S,A) = Entropy(S) - \sum_{V\in Value(A)} \frac{|S_V|}{|S|} \times Entropy(S_V)$$
$$S_V = \{s \in S | A(S) = V\} \tag{2}$$

where $|S_V|$ and $|S|$ are the number of data points in data sets S_V and S, respectively and A is an attribute.

Gain has a bias towards variables with many values that partition the data set into smaller ordered sets. In order to reduce this bias, the entropy of each variable over its m variable values is calculated as *SplitInformation*:

$$SplitInformation(S,A) = \sum_{i=1}^{m} -\frac{|S_i|}{|S|} \times \log_2\left(\frac{|S_i|}{S}\right). \tag{3}$$

Finally, *GainRatio* is calculated by dividing Gain by *SplitInformation* so that the bias towards variables with large value sets is dampened:

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}. \tag{4}$$

C5.0 builds a decision tree greedily by splitting the data on the variable that maximizes *GainRatio*.

2.2. CART algorithm

The Classification and Regression Trees (CART) method of Breiman et al. [14] generates binary decision trees. CART is a non-parametric statistical methodology developed for analyzing classification issues either from categorical or continuous dependent variables. If the dependent variable is categorical, CART produces a classification tree. When the dependent variable is continuous, it produces a regression tree. The CART tree is constructed by splitting subsets of the data set using all predictor variables to create two child nodes repeatedly, beginning with the entire data set. The best predictor is chosen using a variety of impurity or diversity measures. The goal is to produce subsets of the data which are as homogeneous as possible with respect to the target variable. In the CART algorithm for each split, each predictor is evaluated to find the best cut point (continuous predictors) or groupings of categories (nominal and ordinal predictors) based on improvement score or reduction in impurity [14]. Then the predictors are compared and the predictor with the best improvement is selected for the split. The process repeats recursively until one of the stopping rules is triggered.

Download English Version:

https://daneshyari.com/en/article/1720397

Download Persian Version:

https://daneshyari.com/article/1720397

Daneshyari.com