Chapter 19
Prediction of the Quality of Fresh Water in a Basin

Sira M. Allende
Universidad de La Habana, Cuba

Carlos N. Bouza
Universidad de La Habana, Cuba

Daniel C. Chen
Smith and King College, UK

Agustín Santiago
Universidad Autónoma de Guerrero, Mexico

Jose Maclovio Sautto
Universidad Autónoma de Guerrero, Mexico

ABSTRACT

Derivatives play an important role in social and economic studies. They describe the behavior of conditional expectations. Once a phenomena is characterized by parametric specifications, the conditional expectation \( m(x) \) may be modeled by a regression function. Then, derivatives may be computed by fitting the regression function. In applications, parametric estimators are commonly used, because of the unknnowledge of other more effective methods. The validity of a regression fitting approach depends on the knowledge of certain aspects related with the true functional form. In this paper, we develop a study on the usage of soft computing methods for providing an alternative to the use of non-parametric regression. We develop our modeling including neural networks and rough sets approaches. The studied problem is the eutrophication due to the growth of the population of algae. Real life data is provided by a study on a fresh water basin. They are used for developing a comparison of different approaches. A methodology is recommended for implementing a monitoring system of the water quality.

INTRODUCTION

Commonly, environment the studies of the quality of fresh water consider the level of eutrophication of a source. It is well-known that eutrophication plays a key role in evaluating water quality, nutrition status, organic pollution extent etc. The need of monitoring Chl-a leads to sampling field water, which is costly and time-consuming. Modernly, reliable data are obtained by using remote sensing. They should allow obtaining the necessary information on variables close related with Chl-a. Mathematically, this information permits to model the relationship of variables with Chl-a. Obtaining a mathematically based method will reduce costs and maintaining an adequate level of accuracy of the prediction on Chl-a.

DOI: 10.4018/978-1-5225-0997-4.ch019
The objectives of this study are accomplished by comparing the behavior of various classification techniques for establishing the quality of the water in the sample sites. A derivative model is used for predicting Chl-a. The data used for fixing an equation are derived by Allende et al. (2016) are used for prediction. The data were divided into a training sample and a validation sample. The true values of Chl-a allowed classifying the records as satisfying “is a source of potable water”, “is a source usable only for agriculture and similar proposes” and “is source of highly contaminated water”.

We derived a procedure based on the statistical ideas of classifying using the minimum distance of the variables to the mean vector of the class or to the nearest neighbor. A Neuronal Network was used with the same purposes. Bayesian principles supported establishing when the network was well trained. The validation sample was analyzed and the misclassification probabilities were computed for each method. Section 2 is concerned with the aspects of classification using statistical methods or Neural Networks. Section 2 and 3 presents issues on Rough Sets. In section 4 an evaluation of the behavior of the methods was made using a large set of data on contamination in a basin.

CLASSIFICATION

Classification is one of the most used techniques of the multivariate static tool-box. It may be described roughly as a technique concerned with assigning data cases (i.e. observations) to one of a fixed number of possible classes, see Sharma (1996). Classifying is present when dealing with the needs grouping. It has been developed firstly by statisticians. Machine learning researchers have given new insights both from the theoretical and computation issues of classification. They consider that classification is a process that permits learning the existence of patterns. It ends fixing how cases must be classified different predetermined classes of data. See discussion on this problem in Brooks (2010), Hastie-Tibshirani-Friedman (2001). Once the model is built, it can be used to classify new data. Hence, the goal of Classification is sorting observations into two or more labeled classes. The emphasis of it is on deriving a rule that can be used to optimally assign new objects to the classes.

When classifying we assign a membership to each data item (case). Each record is related to a variable of interest \(Y\), which allows classifying, and a set of covariates \(X\). Statisticians denominate the covariables “independent variables” and computer scientists “input variables”. A functional relationship between \(Y\) and \(X\) should be described. Take a case (observation) \(k\). We evaluate, where \(x_k\), is the vector of attributes measured in case \(k\) and \(\bar{w}\) represents a vector of parameters which \(y_k = y_k(x_k, \bar{w})\) weights the importance of each coordinate of \(x_k\).

Let us consider that we have two classes \(c_1\) and \(c_2\). The outcome of the classification may be modeled the dichotomous variable

\[
y_k = \begin{cases} 
1 & \text{if case } k \in c_1 \\
0 & \text{if case } k \in c_2 
\end{cases}
\]

The statistical community uses a decision theory framework for establishing the normative that leads to decide looking for the optimal class membership. An approach is considering as objective deriving a