SDU International Journal of Technological Science
Vol. 8, No 3, December 2016
pp. 19-26

Constructional Technologies

DATA MINING PROCESS FOR RIVER SUSPENDED SEDIMENT ESTIMATION

Özlem TERZİ*, Tahsin BAYKAL
Geliş Tarihi/ Received: 25.07.2016, Kabul tarihi/Accepted: 02.12.2016

Abstract

The accurate estimation of the amount of suspended sediment of rivers is important in water resources engineering because sediment in rivers can also shorten the lifespan of dams and reservoirs. For this purpose, the models are developed to estimate suspended sediment of Kızılırmak River using the data mining process. The river flow values are used as input parameter by developing sediment models. The most appropriate model is obtained by the M5'Rules algorithm. The determination coefficient of the model is obtained as 0.66 and it is observed that the data mining process can be used to estimate suspended sediment of rivers in hydrology field.

Key Words: sediment; data mining process; M5'Rules; Kızılırmak River

AKARSULARDA ASKIDA KATI MADDE TAHMİNİ İÇİN VERİ MADENCİLİĞİ SÜRECİ

Özet

Akarsulardaki askıda katı madde miktarının tahmini, barajların ve rezervuarların ömrünü kısaltabileceği için su kaynakları mühendisliğinde önemlidir. Bu amaçla**,** veri madenciliği yöntemi kullanılarak Kızılırmak Nehri'nin askıda katı madde miktarı tahmin edilmiştir. Katı madde modelleri geliştirilirken nehir akım değerleri girdi parametresi olarak kullanılmıştır. En uygun model M5'Rules algoritması ile elde edilmiştir. Bu modelin determinasyon katsayısı 0.66 olarak elde edilmiş ve veri madenciliği yönteminin hidroloji çalışmalarında katı madde miktarını tahmin etmek için kullanılabilir olduğu görülmüştür.

Anahtar Kelimeler: Katı madde; veri madenciliği süreci; M5'Rules; Kızılırmak Nehri

1. Introduction

The sediment can be reduced the service life by filling the reservoir. Also, it increases the risk and damage of flood by raising the river bed, obstructs the entry of water intake structures, reduces the capacity of the irrigation and drainage canals and increases maintenance costs. Therefore, it must be taken into account the amount of accumulated sediment in the design of the water structures.

* Suleyman Demirel University, Technology Faculty, Civil Engineering Dept. 32260 Isparta
E-posta: ozlemterzi@sdu.edu.tr

It is used sediment transport equations or measurements made at monitoring stations to determine the amount of sediment. Although the determining suspended sediment with measurement from rivers is reliable way, it is a time-consuming expensive method and suspended sediment measurements is not do in spite of measuring river flow on several monitoring stations. Also, it is difficult to obtain the most appropriate solution from sediment equations found in the literature. The most of these equations developed in the laboratory cannot give appropriate results from each other in all natural streams (Dogan, 2009). Because of these difficulties encountered, the researchers have turned to the simple and not time-consuming data mining process being effective for solution of complex hydrological problems (Mishra et al., 2013; Hall et al., 2002; Keskin et al., 2013; Mirbagheri et al., 2010). Heng and Suetsugi (2013) simulated monthly average suspended sediment load of four catchments using artificial neural network. They said that the ANN yielded very satisfactory results.

Data mining is often defined as the process of extracting valid, previously unknown, comprehensible information from large databases in order to improve and optimize decisions (Braha and Shmilovic, 2002). In another way, data mining is defined as the identification of interesting structure, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data (Fayyad and Uthurus, 2002). Data mining is applied in a wide variety of fields for prediction. In addition, data mining has also been applied to other types of scientific data such as bioinformatical, astronomical, and medical (Li and Shue, 2004). Keskin et al. (2009) used data mining process to estimate pan evaporation for Lakes Eğirdir, Kovada, and Karacaören Dam. They showed that the REP tree model has more appropriate results than other models. Terzi (2011) developed flow models using data mining process for Kızılırmak River. The results showed that multilinear regression model gives good correlation with measurement flow values. Terzi et al. (2011) developed solar radiation models using air temperature, relative humidity, wind speed and air pressure parameters with data mining process. They said that multilayer perceptron algorithm can be used to estimate solar radiation.

The objective of this paper is to develop suspended sediment models by data mining process for Kızılırmak River, Turkey. It is used the river flow data to develop the sediment model.

2. Data Mining Process

Data mining (DM) process generally involves phases of data understanding, data preparation, modeling, evaluation and knowledge as shown in Fig. 1. DM process is a hybrid disciplinary that integrates technologies of databases, statistics, machine learning, signal processing, and high performance computing. This rapidly emerging technology is motivated by the need for new techniques to help analyze, understand or even visualize the huge amounts of stored data gathered from scientific applications. The major data mining functions that are developed in research communities include summarization, association, classification, prediction and clustering (Zhou, 2013).

Data understanding starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems and to discover first insights into the data. Data preparation covers all activities that construct the final dataset to be modeled from the initial raw data. The tasks of this phase may include data cleaning for removing noise and inconsistent data, and data transformation for extracting the embedded features (Li and Shue,

2004). Successful mining of data relies on refining tools and techniques capable of rendering large quantities of data understandable and meaningful (Mattison, 1996). The modeling phase applies various modeling techniques, determines the optimal values for parameters in models, and finds the one most suitable to meet the objectives. The evaluation phase evaluates the model found in the last stage to confirm its validity to fit the problem requirements. No matter which areas data mining is applied to, most of the efforts are directed toward the data preparation phase (Li and Shue, 2004).

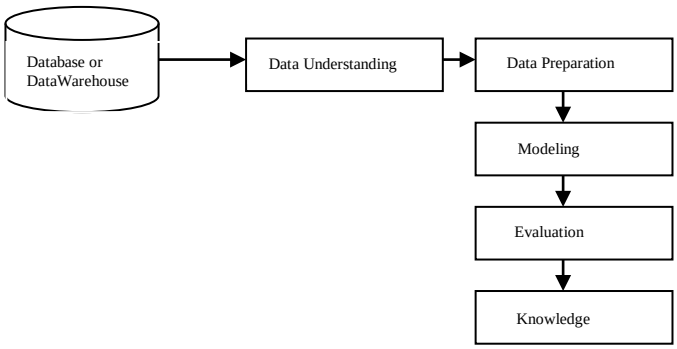

Fig. 1. Data mining process

A good relational database management system will form the core of the data repository, and adequately reflect both the data structure and the process flow, and the database design will anticipate the kind of analysis and data mining to be performed. The data repository should also support access to existing databases allowing retrieval of supporting information that can be used at various levels in the decision making process (Rupp and Wang, 2004).

Data mining is a powerful technique for extracting predictive information from large databases. The automated analysis offered by data mining goes beyond the retrospective analysis of data. Data mining tools can answer questions that are too time-consuming to resolve with methods based on first principles. In data mining, databases are searched for hidden patterns to reveal predictive information in patterns that are too complicated for human experts to identify (Hoffmann and Apostolakis, 2003).

3. River Sediment Models

There are five phases to model river sediment in the data mining process. These phases were given as follows:

a) Data understanding

The length of the Kızılırmak River which is the longest river in Turkey is 1355 km. The area of the watershed is 78 646 km^2. The average flow and rainfall are about 184 m^3/s and 446.1 mm, respectively.

The data used to develop model includes the monthly sediment and flow observations between 1972 and 1997 years. The monthly data were obtained from the General Directorate

of Electrical Power Resources Survey and Development Administration for Salur station on Kızılırmak River.

b) Data preparation

The missing flow and sediment data are examined between 1972 and 1997 years and therefore, the months of missing data are not used for modeling. Hence, the models are developed according to 290 monthly data for 1972-1997 years. It is used 80% of the data for training set and 20% of the data for testing set.

c) Modeling

In order to develop river sediment model, multilinear regression, decision table, KStar, M5'rules and multilayer perceptron algorithms are used in data mining process in Weka. These algorithms are detail explained as given in the following.

1) Multilinear Regression: Multiple linear regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the relationship between the explanatory and response variables (http://www.investopedia.com/ terms/m/mlr.asp). Linear regression is based on the assumption of a linear relationship between the dependent variable Y and its predictors X_1, X_2, …, X_n. Linear regression offers simple and easily interpretable models. However, it can result in inaccurate models that predict poorly in the presence of a nonlinear or nonadditive relationship. Due to the complexity of microarchitectural event interaction and varying event performance penalties, however, it is common for a nonlinear relationship to exist. In the linear case, the functional relationship between Y and its predictors is estimated by minimizing the residual sum of squares (http://homepages.inf.ed.ac.uk/jcavazos/SMART07/ paper_9_9.pdf).

2) Decision Table: Decision table summarizes the data set with a "decision table." In its simplest state, a decision table contains the same number of attributes as the original data set, and a new data item is assigned a category by finding the line in the decision table that matches the nonclass values of the data item. This implementation employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the data set, the algorithm reduces the likelihood of overfitting and creates a smaller, more condensed decision table (Cunningham and Holmes, 1999).

3) KStar: A nearest-neighbor classifier, this algorithm is highly effective in situations with noisy training data, provided it is supplied with a large enough training set. An important note to consider is that the algorithm calculates the distance between instances on all attributes, unlike some other methods. If only a few of the features of the given vector are relevant, then two instances with two identical values for the relevant features may find themselves spaced far apart by this algorithm (Young, 2004).

4) M5'rules: The method for generating rules from model trees, which it is called M5'Rules, is straightforward and works as follows: a tree learner (in this case model trees) is applied to the full training dataset and a pruned tree is learned. Next, the best leaf (according to some heuristic) is made into a rule and the tree is discarded. All instances covered by the rule are removed from the dataset. The process is applied recursively to the remaining instances and terminates when all instances are covered by one or more rules. This is basic separate-and-conquer strategy for learning rules; however, instead of building a single rule, as it is done usually, we build a full model tree at each stage, and make its "best" leaf into a rule. This avoids potential for over-pruning called hasty generalization. In contrast to PART, which employs the same strategy for categorical prediction, M5'Rules builds full trees instead of partially explored trees. Building partial trees leads to greater computational efficiency, and does not affect the size and accuracy of the resulting rules (Hall et al, 1999).

5) Multilayer Perceptron: The back-propagation learning algorithm is one of the most important historical developments in neural networks. It has reawakened the scientific and engineering community to the modeling and processing of many quantitative phenomena using neural networks. This learning algorithm is applied to multilayer feed-forward networks consisting of processing elements with continuous and differentiable activation functions. Such networks associated with the back-propagation learning algorithm are also called back-propagation networks. Given a training set of input-output pairs, the algorithm provides a procedure for changing the weights in a back-propagation network to classify the given input patterns correctly. For a given input-output pair, the back-propagation algorithm performs two phases of data flow. First, the input pattern is propagated from the input layer to the output layer and, as a result of this forward flow of data, it produces an actual output. Then the error signals resulting from the difference between output pattern and an actual output are back-propagated from the output layer to the previous layers for them to update their weights (Lin and Lee, 1995).

d) Evaluation

In this study, the developed models are evaluated by two criteria to estimate suspended sediment by DM process for the Kızılırmak River. The evaluation criteria are the coefficient of determination (R^2) and root mean square error (RMSE).

The coefficient of determination is based on the sediment estimating errors and calculated as,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(S_{i(sediment)} - S_{i(model)})^2}{\sum_{i=1}^{n}(S_{i(sediment)} - S_{mean})^2} \quad (1)$$

where n is the number of observed sediment data, $S_{i(sediment)}$, $S_{i(model)}$ and S_{mean} are monthly sediment measurement, the results of developed sediment model and mean of sediment measurements, respectively.

The root mean square error represents the error of model and defined as following.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(S_{i(sediment)} - S_{i(model)}\right)^2} \quad (2)$$

e) Knowledge

The river sediment models are developed by using multilinear regression, decision table, KStar, M5'rules and multilayer perceptron algorithms in data mining process. The river flow values are used as input parameter in sediment models. The R^2 and RMSE values of the developed models for training and testing data sets are given in Table 1. As seen from Table 1, M5'Rules model gives more appropriate results according to the other models. The results of the M5'Rules model are shown for training and testing sets in Fig. 2. The model comparison plot is generally around ideal line.

Table 1. RMSE and R^2 values of the developed models

Sediment Models	Training Dataset		Testing Dataset	
	RMSE	R^2	*RMSE*	R^2
Multilinear Regression	53658	0.43	53539	0.53
Decision Table	24890	0.71	38135	0.45
KStar	47574	0.50	52114	0.47
M5'Rules	53369	0.48	52983	0.66
Multilayer Perceptron	56370	0.60	53875	0.58

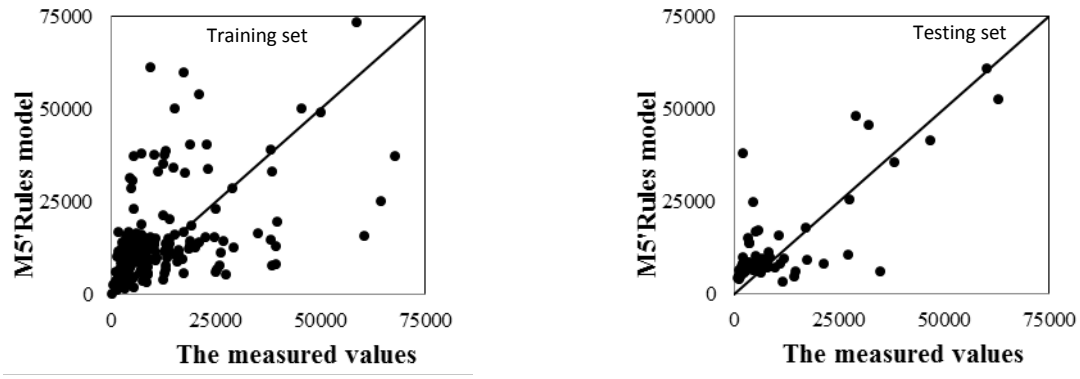

Fig. 2. Scatter plot between M5'Rules model and sediment values for training and testing sets

4. Conclusions

In this study, it is investigated the ability of data mining (DM) process to estimate monthly sediment data for Kızılırmak River which meets vital components such as irrigation, drinking water and power generation. The models by using various algorithms are developed and compared to sediment measurement values. The most appropriate algorithm determining according to the evaluation criteria is M5'Rules algorithm. The sediment values could be estimated easily from available flow data using M5'Rules algorithm. The data mining process seem better than the traditional methods and the model can be developed by adding new data.

References

Braha, D., Shmilovici, A. (2002). Data mining for improving a cleaning process in the semiconductor industry. IEEE Transactions on Semiconductor Manufacturing. vol.15, 1.

Cunningham, S. J., Holmes, G. (1999). Developing innovative applications in agriculture using data mining. Proceedings of Southeast Asia Regional Computer Confederation Conference, Singapore.

Dogan, E. (2009). Prediction of sediment concentration using artificial neural Networks. Turkish Chamber of Civil Engineers, vol. 302, pp. 4567-4582.

Fayyad, U.M., Uthurusamy, R., (2002). Evolving data mining into solutions for insights. Communications of the ACM. vol. 45(8), pp. 28-31.

Hall , M.J., Minns, A.W., Ashrafuzzaman, A.K.M. (2002). The application of data mining techniques for the regionalisation of hydrological variables. Hydrology and Earth System Sciences. vol. 6(4), pp. 685-694.

Hall, M., Holmes, G., Frank, E. (1999). Generating Rule Sets from Model Trees. Proceedings of the Twelfth Australian Joint Conference on Artificial Intelligence. pp. 1-12., Sydney, Australia .

Heng, S. and Suetsugi, T. (2013). Using Artificial Neural Network to Estimate Sediment Load in Ungauged Catchments of the Tonle Sap River Basin, Cambodia. Journal of Water Resource and Protection, vol. 5(2), pp. 111-123.

Hoffmann, D., Apostolakis, J. (2003). Crystal Structure Prediction by Data Mining. Journal of Molecular Structure. vol. 647, pp. 17-39.

http://www.investopedia.com/ terms/m/mlr.asp.

http://homepages.inf.ed.ac.uk/jcavazos/SMART07/ paper_9_9.pdf.

Keskin, M. E., Taylan, D., Kucuksille, E. U. (2013). Data Mining Process for Modelling Hydrological Time Series. Hydrology Research. vol. 44 (1), pp. 78-88.

Keskin, M.E., Terzi, Ö., Küçüksille, E.U., (2009). Data mining process for integrated evaporation model. Journal of Irrigation and Drainage Engineering. vol. 135(1), pp. 39-43.
Li, S.T., Shue, L.Y., (2004). Data mining to aid policy making in air pollution management. Expert System and Applications. vol. 27, pp. 331-340.

Lin, C.T., Lee, C.S.G. (1995). Neural fuzzy systems. Prentice Hall.

Mattison, R. (1996). Data Warehousing: Strategies, Technologies and Techniques Statistical Analysis. SPSS Inc. WhitePapers.

Mirbagheri, S. A., Nourani., V., Rajaee, T., Alikhani, A.(2010). Neuro-fuzzy models employing wavelet analysis for suspended sediment concentration prediction in rivers. Hydrological Sciences Journal – Journal des Sciences Hydrologiques. vol. 55(7), pp. 1175-1189.

Mishra, S., Dwivedi, V. K., Saravanan, C., Pathak, K. K. (2013). Pattern Discovery in Hydrological Time Series Data Mining during the Monsoon Period of the High Flood Years in Brahmaputra River Basin. International Journal of Computer Applications. vol. 67(6), pp. 7-14.

Rupp, B., Wang, J. (2004). Predictive Models For Protein Crystallization. Methods, vol. 34, pp. 390-407.

Terzi, Ö., (2011). Monthly River Flow Forecasting by Data Mining Process. Knowledge-Oriented Applications in Data Mining, K. Funatsu (Ed.). InTech.

Terzi, Ö., Küçüksille, E.U., Ergin, G., İlker, A. (2011). Estimation of Solar Radiation Using Data Mining Process. SDU International Technologic Science. vol. 3(2), pp. 29-37.

Young, A., (2004). Automatic Acronym Identification and the Creation of an Acronym Database. The Technical Report, The University of Sheffield.

Zhou, Z.-H., (2003). Three Perspectives of Data Mining. Artificial Intelligence. vol. 143(1), pp. 139–146.