8

Monthly River Flow Forecasting by Data Mining Process

Özlem Terzi
Suleyman Demirel University
Turkey

1. Introduction

In design, plan, project, construction, maintenance, and especially management of water resources, surface water input and output must be calculated based on measurements. One of the priority parameter is surface flow in the studies. The flow data measured in the past is required for design of the water structure be built in the future and calculation of natural disasters such as flood and drought according to the pre-specified risk level (Şen, 2003).
Stochastic models and artificial intelligence techniques (artificial neural networks, fuzzy logic and adaptive neuro fuzzy inference systems etc.) on flow predicting are commonly used by many researchers while data mining (DM) process is not yet widely used in the hydrology area. Russo et al. (2006) fitted a stochastic rainfall model to rainfall radar data in order to produce a realistic representation of the distribution of rainfall in space and time. The results show that the model, calibrated on the study area, is able to forecast satisfactorily the rain field in space and time. Archer & Fowler (2008) investigated the links between climate and runoff for eight gauging stations in the Jhelum catchment but then concentrated on seasonal forecasting of spring and summer inflows to Mangla Dam. They are used precipitation and temperature variables to forecast summer season flows at stations upstream from the reservoir with a lead time of up to three months based on multiple linear regression models. The analysis demonstrates that good forecasts within 15% of observed flows for 92% of years can be achieved for summer season flows from April to September. For spring flows from April to June, excellent forecasts can be provided within 15% of observed flows for 83% of years. Lin & Chen (2004) used the radial basis function network (RBFN) to construct a rainfall-runoff model, and presented the fully supervised learning algorithm for the parametric estimation of the network. The proposed methodology has been applied to an actual reservoir watershed to forecast the one- to three-hour ahead runoff. The result shows that the RBFN can be successfully applied to build the relation of rainfall and runoff. Rajurkar et al. (2004) presented an approach for modeling daily flows during flood events using ANN. They showed that the approach produces reasonably satisfactory results for data of catchments from different geographical locations. Nayak et al. (2004) suggested that performance of ANFIS model is capable of preserving the statistical properties of the time series and it is viable for modeling river flow series. Keskin et al. (2006) developed a flow prediction model, based on the adaptive neural-based fuzzy inference system (ANFIS) coupled with stochastic hydrological models. An ANFIS is applied to river flow prediction in Dim Stream in the southern part of Turkey. Synthetic

series, generated through autoregressive moving-average models, are then used to train data sets of the ANFIS. They showed that the extension of input and output data sets in the training stage improves the accuracy of forecasting by using ANFIS. Jacquin & Shamseldin (2006) explored the application of Takagi–Sugeno fuzzy inference systems to rainfall–runoff modeling. The models developed intend to describe the non-linear relationship between rainfall as input and runoff as output to the real system using a system based approach. They showed that fuzzy inference systems are a suitable alternative to the traditional methods for modeling the non-linear relationship between rainfall and runoff.

Knowledge discovery uses data mining and machine learning techniques that have evolved through a synergy in artificial intelligence, computer science, statistics, and other related fields. Although there are technical differences, the terms 'machine learning', 'data mining', and 'knowledge discovery and data mining (KDD)' are often used interchangeably (Goodwin et al., 2003).

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 & Shmilovici, 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 & Uthurusamy, 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 & Shue, 2004). Keskin et al. (2009) developed pan evaporation models using data mining process for Lakes Eğirdir, Kovada, and Karacaören Dam and formed an integrated evaporation model by aggregation of daily pan evaporation of these lakes for the Lakes District in the southern part of Turkey. They showed that the REP tree model has better agreement with measured daily pan evaporation than other models.

This chapter is organized as follows: Section 2 briefly defines the DM process. Section 3 describes to construct a model to forecast river flow using data mining process. This model is developed to be accomplished in Kızılırmak River which is longest river in Turkey. Section 4 includes conclusions of the chapter.

2. Data Mining process

Data mining (DM) process generally involves phases of data understanding, data preparation, modeling, evaluation and knowledge as shown in Figure 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, 2003).

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 & Shue, 2004). Successful mining of data relies on refining tools and techniques capable of rendering large quantities of data understandable and meaningful (Mattison, 2000). The

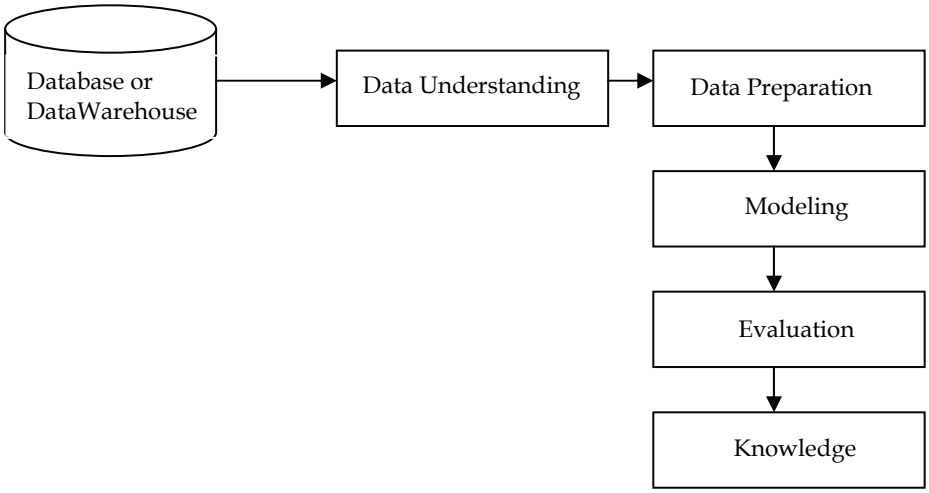

Fig. 1. Data mining process

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 & Shue, 2004).

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 & 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 & Apostolakis, 2003).

3. River flow models

A data mining process generally has five phases. In this chapter, these phases were considered as follows.

3.1 Data understanding

The DM process is applied to the Kızılırmak River in northern part of Turkey to forecast river flow. The length of the Kızılırmak River which is 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 flow and rainfall observations between 1975 and 2005 years i.e. a total of 322 months in this chapter. The monthly flow

data were obtained from the General Directorate of Electrical Power Resources Survey and Development Administration for Yamula (1501), Söğütlühan (1535) ve Bulakbaşı (1539) stations. The monthly rainfall data were obtained from Turkish State Meteorological Service for Kayseri, Sivas and Zara stations.

3.2 Data preparation

In some months between 1975 and 2005 years, missing rainfall and flow data are determined. Therefore, the months of missing data are not used for modeling. Hence, the models are developed according to 322 monthly data for 1975-2005 years. It is used 80% of the data for training set and 20% of the data for testing set.

3.3 Modeling

In order to develop river flow model, multilinear regression, multilayer perceptron, radial basis function (RBF) network, decision table, REP tree and KStar algorithms are used in data mining process in Weka. Detailed explanations of these algorithms are given in the following.

Multilinear Regression

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable (http://www. statsoft.com/textbook/stathome.html). Linear regression is based on the assumption of a linear relationship between the dependent variable Y and its predictors X1, X2, …, Xn. 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).

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 & Lee, 1995).

Radial Basis Function Network

A radial basis function network is a two-layer network whose output neurons form a linear combination of the basis functions computed by the hidden neurons. The basis functions in the hidden layer produce a localized response to the input. That is, each hidden neuron has

a localized receptive field. The basis function can be viewed as the activation function in the hidden layer. The basis function used is a Gaussian function (Fu, 1994).

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 & Holmes, 1999).

REP Tree

The decision tree tool of REP tree in Weka was employed for formulating the resource access patterns for the considered applications that are common in the target execution environment. The REP tree procedure builds a decision tree using information gain as the splitting criterion, and uses reduced-error pruning for pruning. This procedure is also characterized by lower computational overhead compared to other decision-tree-based classification methods as a result of its efficient pruning mechanism (Rajan et al., 2006).

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).

3.4 Evaluation

The monthly flow data of Yamula (1501) and Bulakbaşı (1539) station and monthly rainfall data of Kayseri, Sivas and Zara rainfall measurement stations are used to forecast monthly flow for Söğütlühan (1535) station. The cross-correlations between input and output parameters are calculated for the selection of the input parameters. According to cross-correlations, three different models have been developed to forecast for flow values of Söğütlühan station (Table 1). In the first model (M1), rainfall data of Kayseri, Sivas and Zara stations and flow data of Yamula and Bulakbaşı stations are used as input parameters.

Model	Input parameters
M1	Flow (1501, 1539), Rainfall (Kayseri, Sivas, Zara)
M2	Flow (1501, 1539), Rainfall (Sivas, Zara)
M3	Flow (1501, 1539)

Table 1. Construction of the developed river flow models

While input parameters of second model (M2) are rainfall data of Sivas and Zara stations and flow data of Yamula and Bulakbaşı stations, them of other model (M3) are only flow data of Yamula and Bulakbaşı stations.

Two criteria are used to evaluate the adequacy of each model: the coefficient of determination (R^2) and the root mean square error (RMSE).

The coefficient of determination based on the flow forecasting errors is calculated as,

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

where n is the number of observed data, $F_{i(\text{flow})}$, $F_{i(\text{model})}$ and F_{mean} are monthly flow measurement, the results of developed flow model and mean flow measurements, respectively.
The root mean square error represents the error of model and defined as,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_{i(flow)} - F_{i(\text{mod}el)})^2} \quad (2)$$

where parameters have been defined above.

3.5 Knowledge

The best-fit DM algorithm is determined according to coefficients of determination (R^2) using Eq. (1) and RMSE values using Eq. (2) for testing data set. The results of statistical analyses of the models are given in Table 2. As seen from Table 2, the best-fit DM algorithm is determined as multilinear regression algorithm for M1, M2 and M3 models. Although M1, M2 and M3 models has same the R^2 values (0.981) using multilinear regression algorithm for testing data set, M3 model has the lowest RMSE among all models. The M3 models based on flow data gives generally best R^2 values. In case of a limited parameter, the M3 model based on only flow data has advantage because it uses fewer input parameters. Hence, the M3 model developed by using multilinear regression algorithm is selected for flow forecasting of Kızılırmak River among the developed models. The results of the developed model are also indistinguishable (mean, standard deviation etc.) from measured flow values.
Figure 2 shows comparison plot of flow values forecasted with M3 model and measured for testing data set. The comparison plot of the model is around 45° straight lines which imply that there are no bias effects. It is apparent a close relationship between forecasted and measured flow values. The results suggest that the monthly flow could be easily forecasted from flow and rainfall data using DM algorithms.

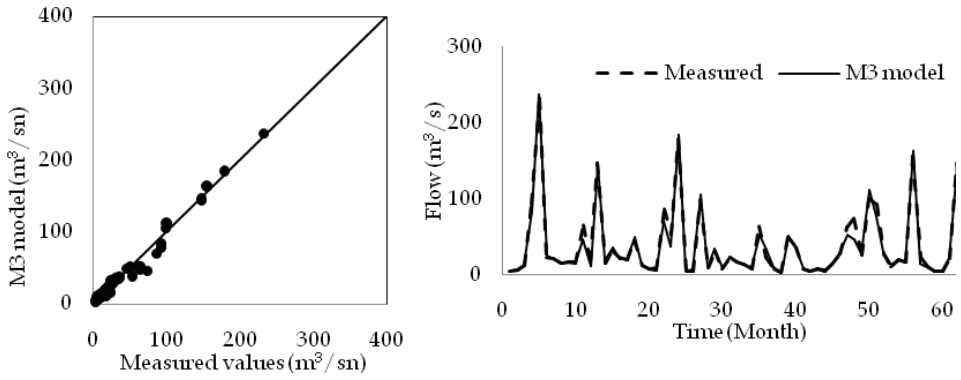

Fig. 2. Comparison of flow values of measured and the M3 model developed multilinear regression algorithm

Models	*Algorithms*	*Testing data set*					
		Mean (m^3/sn)	*Std. Devia.*	*Skewness*	*Kurtosis*	*RMSE*	R^2
M1 (rainfall data of Kayseri, Sivas and Zara stations- flow data of Yamula and Bulakbaşı stations)	Measured flow	38.59	47.18	2.20	4.93	-	-
	Multilinear Regression	38.09	47.08	2.41	5.95	6.53	0.981
	Multilayer Perceptron	41.11	47.04	2.70	9.01	8.85	0.964
	RBF Network	39.19	34.40	0.79	-1.37	36.37	0.397
	Decision Table	34.91	38.20	2.17	4.74	27.11	0.665
	REP tree	38.35	45.85	2.42	6.81	12.08	0.933
	KStar	31.69	36.79	1.96	3.07	18.41	0.845
M2 (rainfall data of Sivas and Zara stations- flow data of Yamula and Bulakbaşı stations)	Multilinear Regression	38.09	47.08	2.41	5.95	6.53	0.981
	Multilayer Perceptron	42.06	48.16	2.78	9.44	9.77	0.956
	RBF Network	40.74	37.52	0.95	-1.08	34.89	0.445
	Decision Table	34.91	38.20	2.17	4.74	27.11	0.665
	REP tree	38.28	45.89	2.42	6.79	12.09	0.933
	KStar	32.72	37.82	1.99	3.49	17.80	0.855
M3 (flow data of Yamula and Bulakbaşı stations)	Multilinear Regression	37.81	46.94	2.44	6.10	6.51	0.981
	Multilayer Perceptron	40.45	45.93	2.56	6.66	7.68	0.973
	RBF Network	34.46	39.84	2.00	2.17	22.05	0.778
	Decision Table	34.91	38.20	2.17	4.74	27.11	0.665
	REP tree	38.30	45.87	2.42	6.80	12.09	0.933
	KStar	33.22	39.49	2.12	4.21	12.23	0.932

Table 2. The descriptive statistics of developed models

4. Conclusions

Determination of the flow is of great importance especially in many issues such as flood control, use and operation of water, determination of settlement and energy production. This chapter indicates the ability of data mining (DM) process to forecast monthly flow data. The DM process has been applied to Kızılırmak River which meets vital components such as irrigation, drinking water and power generation. The various models based on rainfall and flow data are developed and compared to measurement flow data. The most appropriate algorithm is determined according to the model performance criteria for testing data set. The comparisons show that there is a better agreement between monthly flow data and the results of the multilinear regression algorithm in data mining process than others for M1, M2 and M3 models. The M3 model based on only flow data gives better performance than M1 and M2 models. It is shown that the M3 model is superior among all models. The performance of the developed models suggests that the flow could be successfully forecasted from available flow and rainfall data using DM process, and the models are used in water resources planning and management. Finally, DM process can be used for forecasting flow in which measurement system has failed or to forecast missing monthly flow data in hydrological modeling studies.

5. References

Archer, D.R. & Fowler, H.J. (2008). Using meteorological data to forecast seasonal runoff on the River Jhelum, Pakistan. *Journal of Hydrology*, 361, 10-23

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

Cunningham, S. J. & Holmes, G. (1999). Developing innovative applications in agriculture using data mining. *Proc. Southeast Asia Regional Computer Confederation Conf.*, Singapore, Dept. of Computer Science, Univ. of Waikato, Hamilton, New Zealand

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

Fu, L. (1994). *Neural Networks in Computer Intelligence*. McGraw-Hill International Editions.

Goodwin, L.; VanDyne, M.; Lin, S. & Talbert, S. (2003). Data mining issues and opportunities for building nursing knowledge. *Journal of Biomedical Informatics*, 36, 379-388

Hoffmann, D. & Apostolakis, J. (2003). Crystal structure prediction by data mining. *J. Mol. Struct.*, 647,1–3, 17–39

Jacquin, A.P. & Shamseldin, A.Y. (2006). Development of rainfall–runoff models using Takagi–Sugeno fuzzy inference systems. *Journal of Hydrology*, 329, 154–173

Keskin, M.E.; Taylan, D. & Terzi, Ö. (2006). Adaptive neural-based fuzzy inference system (ANFIS) approach for modelling hydrological time series. *Hydrological Sciences–Journal–des Sciences Hydrologiques*, 51, 4, 588-598

Keskin, M.E.; Terzi, Ö. & Küçüksille, E.U. (2009). Data mining process for integrated evaporation model. *Journal of Irrigation and Drainage Engineering*, 135, 1, 39-43

Li, S.T. & Shue, L.Y. (2004). Data mining to aid policy making in air pollution management. *Expert System and Applications*, 27, 331-340

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

Lin, G.F. & Chen, L.H. (2004). A non-linear rainfall-runoff model using radial basis function network. *Journal of Hydrology*, 289, 1–8

Nayak, P.C.; Sudheer, K.P.; Rangan, D.M. & Ramasastri, K.S. (2004). A neuro-fuzzy computing technique for modeling hydrological time series. *Journal of Hydrology*, 291, 1-2, 52-66

Mattison, R. (2000). *Data Warehousing: Strategies, Technologies and Techniques Statistical Analysis*, SPSS Inc. White Papers

Rajan, D.; Poellabauer, C. & Chawla, N.(2006). Resource Access pattern mining for dynamic energy management. *Proc. Workshop on Autonomic Computing: A New Challenge for Machine Learning*, Berlin, Germany

Rajurkar, M.P.; Kothyari, U.C. & Chaube, U.C. (2004). Modeling of the daily rainfall-runoff relationship with artificial neural network. *Journal of Hydrology*, 285, 96–113

Rupp, B. & Wang, J. (2004). Predictive models for protein crystallization. *Methods*, 34, 3, 390–407

Russo, F.; Lombardo, F.; Napolitano, F. & Gorgucci, E. (2006). Rainfall stochastic modeling for runoff forecasting. *Physics and Chemistry of the Earth*, 31, 1252–1261

Şen, Z. (2003). *Water Science and Methods*. Water Foundation Publications, ISBN:975-6455-02-0, Istanbul

Young, A. (2004). *Automatic acronym identification and the creation of an acronym database*. Technical Rep., Univ. of Sheffield, Sheffield, U.K.

Zhou, Z.H. (2003). Three perspectives of data mining. *Artif. Intell.*, 143, 1, 139–146

Knowledge-Oriented Applications in Data Mining
Edited by Prof. Kimito Funatsu

ISBN 978-953-307-154-1
Hard cover, 442 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by 'Data Mining' address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821