Temporal Data Mining of Uncertain Water Reservoir Data

Abhinaya Mohan
Department of Computer Science
and Engineering
University of Nebraska-Lincoln
abbhey@gmail.com

Peter Revesz
Department of Computer Science
and Engineering
University of Nebraska-Lincoln
revesz@cse.unl.edu

ABSTRACT
This paper describes the challenges of data mining uncertain water reservoir data based on past human operations in order to learn from them reservoir policies that can be automated for the future operation of the water reservoirs. Records of human operations of water reservoirs often contain uncertain data. For example, the recorded amounts of water released and retained in the water reservoirs are typically uncertain, i.e., they are bounded by some minimum and maximum values. Moreover, the time of release is also uncertain, i.e., typically only monthly or weekly amounts are recorded. To increase the effectiveness of data mining of uncertain water reservoir data, *temporal data mining* with inflow and rainfall data from several prior months was used. The experiments also compared several different data classification methods for robustness in the case of uncertain data.

Categories and Subject Descriptors
H.2.8 [**Database Application**]:Data Mining.

General Terms
Measurement, Performance, Design and Standardization.

Keywords
Spatio-temporal data, uncertainty, history, water reservoir, classifiers and prediction.

1. INTRODUCTION

Water reservoir operators perform a complex task trying to balance the need to retain plenty of water for irrigation and other uses of water while preventing an overflow of the reservoir that could cause flooding of the surrounding area. As a result water reservoir operators accumulate a certain set of skills and knowledge that are not easy to express mathematically. Hence even though many water researchers studied water reservoir operations (see Section 2.1), there is currently no good automation of water reservoirs.

In this paper, we apply temporal data mining as a new approach to learn from human water reservoir operators. In theory, a data mining algorithm could learn general policies of handling the water reservoirs, and the learned policies could be automated in the future, avoiding occasional errors in human judgment and saving costs in human operators. In practice, the data mining task for water reservoirs is more complicated than for regular data mining tasks because water reservoir data is typically uncertain. For example, the recorded amounts of water released and retained in the water reservoirs are typically uncertain, i.e., they are bounded by some minimum and maximum values. Moreover, the time of release is also uncertain, i.e., typically only monthly or weekly amounts are recorded.

Uncertain data occurs not only in water reservoir operations but also in a wide variety of other applications. Hence there is an increasing interest in data mining uncertain data. We argue in this paper that the challenge of data mining uncertain data can be overcome by the temporal data mining method introduced by Revesz and Triplet [14]. As shown in Figure 1, regular data classification considers only contemporary or immediately preceding temporal values, but temporal data classification improves the accuracy by considering the feature values pertaining to some *n* time units back in time.

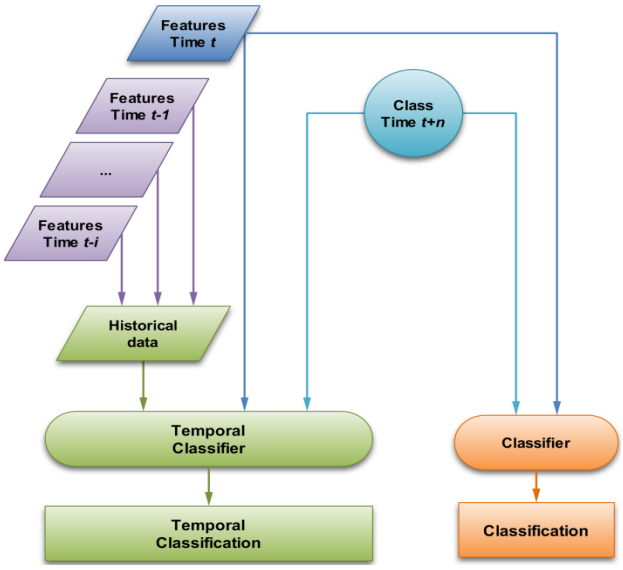

Figure 1. Temporal vs Regular Data Mining.

ACM SIGSPATIAL QUeST'12, November 6, 2012. Redondo Beach, CA, USA.

Revesz and Triplet [14] showed experimentally that temporal data classification greatly improves the performance of decision trees and SVMs (support vector machines) for weather forecasting when the input data contains usual measurement values of temperature wind direction, wind speed etc. That is, in the weather data the only uncertainty was associated with usual measurement errors.

In contrast, reservoir operational data is uncertain primarily because of human recording practices. It is traditional to record only weekly and monthly data and ranges of water release and retention values. That practice makes reservoir data uncertain and its data mining more challenging than the data mining of simple weather data. Our paper also consider more classifiers than [14], namely, we also investigate multilayer perceptron networks and Naïve-Bayes classifiers.

This paper is organized as follows. Section 2 presents previous work while Section 3 describes the data sources and their temporal enhancements. Experimental results are presented in Section 4. Finally, the derived conclusions and potential future work are summarized in Section 5.

2. PREVIOUS WORK

2.1 Current Water Reservoir Models

Alshaikh and Taher [3], Chaves and Chang [7], Gates and Alshaikh [9], Neelakantan and Pundarikanthan [13], and Simonovic [15] evaluated the efficiency of simulation optimization frameworks by incorporating data driven models with optimization algorithms. In water resources, for development of optimal policies of system operation, different methodologies are employed such as mathematical models, distributed physically-based models, and empirical models. Empirical models are currently the most frequent due to their evaluation techniques. Data-driven modeling (DDM) is to formulate a model based on existing characteristics. A data model is quantified by its ability to identify and establish the nature of connections between the features and the variables based on a given set of conditions or rules.

Data-driven modeling has been used by Abebe et al. [1] for estimating missing precipitation data, by Abrahart and See [2] and by Dawson and Wilby [8] for rainfall-runoff modeling, by Bhattacharya et al. [4] for controlling water level, by Bhattacharya and Solomatine [5] for reconstructing stage-discharge relationships, by Hall and Minns [10] for classification of hydrologically homogeneous regions, by Nageshkumar and Dhanya [12] for rainfall prediction, by Solomatine et al., [16] for the classification of surge water levels in the coastal zone, and by Solomatine et al. [17] for replicating the behavior of a hydrodynamic/hydrological river model.

In addition to these approaches, data mining was also applied by some researchers in order to learn from the expert knowledge of the reservoir operators. Neelakantan and Pundarikanthan [13] and Chandramouli and Raman [6] introduced reservoir optimization techniques using the data clustering and rule mining. In addition, Sudha et al., [18] and Taghi Sattari et al., [19] used decision trees to devise irrigation reservoir rule curves.

2.2 Classifiers

We use classifiers to classify items that are described by a set of features and a set of labels. Each classifier maps the feature space X to a set of labels Y.

A classifier is found using a training set. In the training set both the set of features and the set of labels are known. During the normal application of the classifier, only the features are known. We review below the classifiers that we have considered.

2.2.1 Multilayer Perceptron

The multilayer perceptron is an artificial neural network. The interesting feature about artificial neural networks is the fact that they have an adaptive learning technique that is employed, i.e., it changes and modifies its structure based on the input and the output values that are generated in the system during the learning phase. The multiplayer perceptron is a feed forward network that consists of several layers, namely, the input layer, the output layer, and numerous hidden layers. During normal operation or recall, the data flows from the input layer, through each of the hidden layers one-by-one, to the output layer.

Multilayer perceptron networks also use *back propagation* for learning. Back propagation is carried out in two phases. In the first phase, the network is supplied with the training set to generate the activation functions which are then propagated backwards from the output layer to the input layer to formulate the difference between the expected and the current output. This phase is followed by a second phase, when the weights or connection strengths are modified by adding to the current weight and the above calculated difference to improve the output performance. The above two steps are repeated until the multilayer perceptron's performance is satisfactory in that the outputs are close to the expected outputs for the training set.

2.2.2 Naïve Bayes

Due to the inherent property of the Naïve Bayes classifier to consider each of the features as independent entities, it can be used to design a prediction model for the water reservoir release. In addition, the Naïve Bayes classifier works upon the most-likelihood mechanism and is a type of predictive modeling where a model is created or chosen to predict the probability of an outcome. These models typically consist of using detection of the future values based on training them with a training set. In general, one advantage of the Naïve Bayes classifier is that it performs well with a relatively small amount of training data.

2.2.3 Decision Tree

The decision tree is yet another predictive modeling-based classifier. Decision trees are one of the earliest classifiers. Decision trees provide a diagrammatic illustration of the results and an explanation for arriving at them. A decision tree is essentially composed of the following:

- Internal nodes where various simple conditions, i.e., attributes compared with constants can be tested.
- Branches which corresponds to the values of the attributes.
- Finally, the leaves that assign a class to the input data.

Revesz and Triplet [16] used the temporal data classification to design a simplified version of weather forecast where the forecast used only two classes, namely warm or cold temperature.

We propose to extend the idea of temporal data classification to devise a more extensive data mining technique with a larger number of classes. We use the extended temporal data mining algorithms propose to predict the amount of water that needs to be released from a reservoir. The performance of each of the classifiers using the temporal data classification method is compared to identify which combination gives the most accurate prediction.

water release data have been obtained for 36 years, or 432 months collected by the Public Works Department (PWD) of Tamil Nadu State, which has been regularly monitoring these variables. Data collection was also carried out by off-site measurements which consist of sampled data over a period of time or data converted from manual recordings into a set of samples. These collected data were provided to us by the PWD of Tamil Nadu State.

Table 1. Data representation depicting the sample instances in the Reservoir Dataset.

MONTH	YEAR	RAINFALL(R)	INFLOW(I)	STORAGE(S)	RELEASE(L)
Aug	81	251_500	20001_30000	60001_70000	20001_30000
Sep	81	501_more	160001_more	60001_70000	120001_140000

Table 2. Temporal data representation of the Reservoir dataset

M	R_{M-2}	I_{M-2}	S_{M-2}	L_{M-2}	R_{M-1}	I_{M-1}	S_{M-1}	L_{M-1}	R	I	S	L
Aug	251_ 500	20001_ 30000	60001_ 70000	20001_ 30000	501_ more	160001_ more	60001_ 70000	120001_ 140000	101_150	70001_ 80000	80001_ 100000	**70001_ 80000**
Sep	501_ More	160001_ more	60001_ 70000	120001_ 140000	101_ 150	70001_ 80000	80001_ 100000	70001_ 80000	101_150	70001_ 80000	80001_ 100000	**70001_ 80000**

3. THE DATA SOURCE

3.1 Data Base

For testing the effectiveness of the various methods, we considered as a case study the Cauvery river basin in South India. The Cauvery River extends over a length of about 1200 km, and the watershed extends over an area of more than 80 square km. The major reservoir of this river basin is the Mettur reservoir, in Tamil Nadu State. A hydrological database was developed after collating the data observations over a period of time. Some of these observations were carried out manually, while other observations were recorded using sophisticated sensors of stream and rain gauges. The release from the reservoir is a decision variable dependent on the current storage, the inflow, and to certain a extent the rainfall in the watershed. Monthly rainfall, inflow, storage and

The original data contained a very small interval of data that coincided with the same timelines. A total of 120 rows were found to be consistent with all the values and their respective timestamps. These numerical values were then categorized into a range of possible values (bins). We have considered an equal width binning of data. Data processing to understand reservoir operations were done on the equal width bin data.

Data is usually split into three datasets: (a) training, (b) testing and (c) validation. In the training phase, using only the training dataset, the data mining algorithm finds a classifier. The accuracy of the classifier that is found is tested using the testing data. The training can be repeated until accuracy of the classifier reaches the minimum acceptable limit. The performance of tested data mining algorithm on mapping unused data (not used for training and testing) is evaluated during the validation phase.

Since measurements may often be noisy, an attempt to maximize the fit to the training data may lead to the model capturing not only the process but also the noise - a phenomenon known as over-fitting. An over-fitted model may not perform well on a new dataset. Validation can give some guarantee that the over-fitting problem is avoided.

In the present work, a model is built using the training data and is tested with the testing data. These two datasets have identical statistical distributions as they are randomly chosen from all the available data.

4. EXPERIMENTAL RESULTS

The following are the feature variables that are taken into consideration:

- **Storage**: The amount of water in the reservoir at a given time (measured in Million Cubic Feet, Mcft).
- **Rainfall**: The amount of precipitation that actually takes place (measured in mm).
- **Inflow**: The actual amount of water that reaches the reservoir after rainfall through any water source (Mcft).

In addition, the label variable is the following:

- **Release**: The amount of water that is released from the reservoir. To have a finite number of labels, we consider only a finite number of range values as possible labels (Mcft).

An important factor that is to be considered for estimating the release is that it satisfies the mass balance equation, which is given below:

$$\textit{Storage in next month } (S_{t+1}) = \textit{Current Storage } (S_t) + \textit{Inflow } (I_t) - \textit{Release } (R_t) - \textit{Evaporation } (E_t)$$

In addition, the release is a function of storage, inflow and demand. Therefore,

$$\textit{Release} = f(\textit{storage, inflow, demand})$$

The above two equations are used to estimate the release in data-driven modeling apart from data mining.

4.1 Preprocessing

The data obtained was not in the form that was easy for data mining. We had to simplify the number of bins and make sure that they have equal width. Different size bins or too small bins may result in over-fitting. The simplified input data set was a relation with rows of the form (Month, Year, Rainfall, Inflow, Storage, and Release). Some sample rows are shown in Table 1 (basic data) and Table 2 (limited history data).

Weka 3.6, a user-friendly data mining tool, was used to implement the classification for reservoir operation. In the data pre-processing stage, all the data sets have been linearly converted into intervals to lower the chance of over-fitting. In the present study, the release is identified as the label (also called the dependent or decision variable). Data mining results in a classifier that finds the release value as an interval or range.

In Weka 3.6, the numerical data for the labels (release) were converted to nominal values as required by its data mining algorithms.

4.2 Discussion

As already mentioned, the data was split into two sets; the training data and the testing data. The training data were used to obtain some classifier. The classifier's performance was evaluated on the testing data. In our experiments for Table 3, we used 70% of the available data for training, and 30% of the data for testing. We used the k-fold cross validation strategy (for the present study the value of k is found to be optimal at 5 folds) implemented in Weka 3.6. In this type of validation, the random sampling of the training and the testing data is repeated 5 times.

Experiments show that the reservoir release classifier is more accurate when history is taken into consideration as shown in Table 2. Considering each month as a separate instance (as in Table 1) results in missing out all the details that can be harvested if the data pertaining to earlier months are also considered in predictions. Temporal data mining aims to identify future release values based several previous values.

Consider the following motivational example for the use of temporal data. Suppose the current month's storage, rainfall and inflow were usual, but the previous month's release was low. Then the current release should be higher than average to provide enough water down the river. Temporal data mining can capture this scenario, but regular data mining, which only looks at the current values could lead to less than optimal result. Therefore, our temporal data, as shown in Table 2, looked back two additional months beside the current month.

We have measured the performance accuracy of the classifiers using the Root Mean Square (RMS) error measure, extended in the following way.

Let $(\mathbf{a_1, a_2, a_3, a_4\ldots})$ be the set of actual values measured as readings. RMS also uses predicted values. We take the middle of the interval values as the predicted value **p.** We now have a set of predicted values $(\mathbf{p_1, p_2, p_3, p_4\ldots})$ corresponding to each possible release interval.

Further, let N_c be the total number of labels (i.e., the total number of possible water release range values), and suppose that the classifier is trained using **n** instances. Then we calculate the RMS value as follows:

$$RMS\ error\ = \sqrt{\frac{\sum_{i=1}^{n}(a_i\ - p_i)^2}{N_c - 1}}$$

Table 3. Performance Measures of Different Classifier Methods

Classifier method	*Total No of Instances to be classified*	*Correctly Classified instances*	*Root mean square error*
Naïve bayes (Training)	84	70	0.121
Naïve bayes (Testing)	36	13	0.2330
Multilayer Perceptron (Training)	84	74	0.1138
Multilayer Perceptron (Testing)	36	15	0.2284
Decision trees (Training)	84	82	0.0088
Decision trees (Testing)	36	10	0.2548

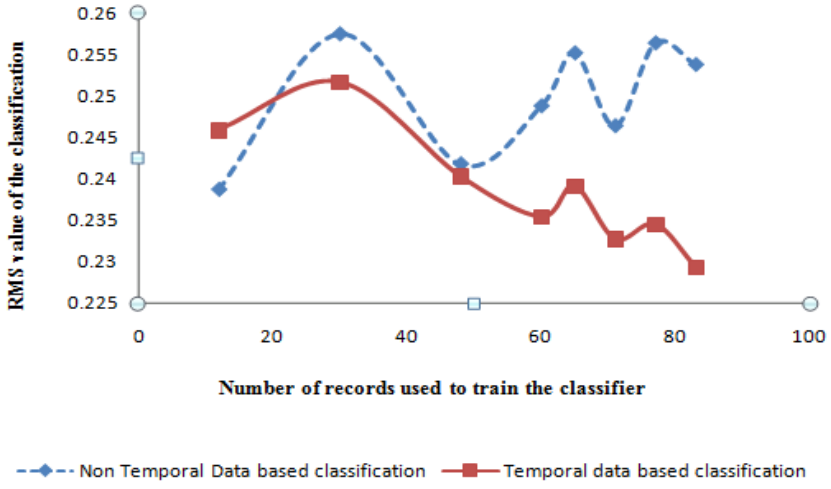

Fig. 2 Comparison of regular and temporal classification using the Naïve Bayes classifier

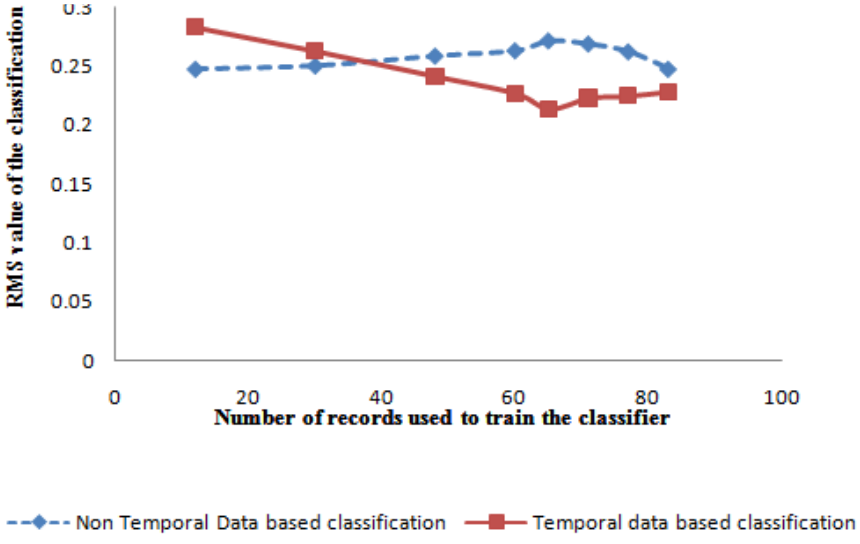

Fig. 3 Comparison of regular and temporal classification using the Multilayer Perceptron Classifier

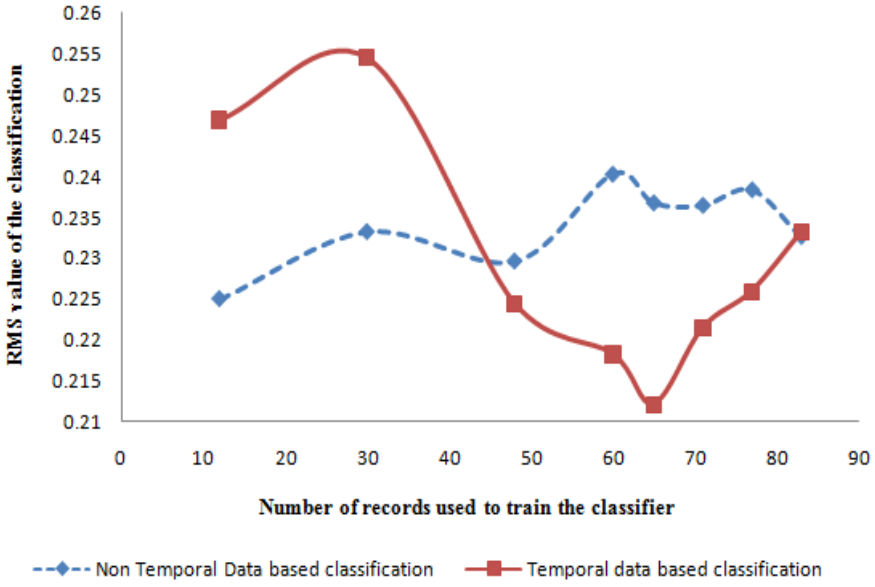

Fig. 4 Comparison of regular and temporal classification using the Decision Tree classifier

This aggregated RMS error value is the measure of the degree of performance, i.e. the degree to which the classifier correctly predicts the required value. The performance of each of the classifiers using the temporal data is listed in Table 3. As shown in Table 3, the Multilayer Perceptron classifier had the most accurate performance on the testing data because it had the lowest RMS error value, namely 0.2284.

Figures 2, 3 and 4 depict the performance comparison using the Naïve Bayes, the Multilayer Perceptron and the Decision Tree classifiers. All of these indicate a visible difference in the RMS values between using regular vs. temporal data mining. Though the regular data presents a comparatively lower RMS error to begin with, this cannot be taken as a valid state to measure the performance because only very few percentages of data (between 10 and 25 percent) were considered. This is of little significance because the dataset consists of 120 instances and a model built using 10% of the data, i.e., 12 instances, cannot be substantiated. For larger training data, the temporal data mining shows better results than regular data mining in terms of decreased RMS error values.

Interestingly, all the classifiers display an improvement, strongly backing up the theory that the history of the reservoir data does contain valid information to help design a valid classifier for regulating water release from the reservoir.

5. CONCLUSION AND FUTURE WORK

We have studied uncertain spatio-temporal data mining for the purpose of deriving operations for water quantity release. We have extended the algorithm suggested by Revesz and Triplet [14] to model the reservoir operation data using temporal data mining. Experiments show that the uncertain temporal data mining approach is an effective and efficient method. In particular, we overcome the challenge of dealing with complicated reservoir operational strategies using instead a simple approach of data mining to learn the expert knowledge of human operators of the water reservoir. Our temporal data mining showed a significant improvement over regular data mining.

As experimentally demonstrated, the proposed data-mining approach also delivers good performance when trained with relatively few instances, which is in contrast to the normal belief that a large training data is required for accuracy. The proposed model has been proved to be effective in predicting the water release quantity, which can lead to development of automated water reservoir operations, thus providing a cost-effective management of water reservoirs. Our study also suggests a small training data requirement for similar data mining problems, not only other water reservoirs, but in general where the labels are a large set of interval values.

5.1 Future Work

The data mining modeling here is carried out based on the data from only a single reservoir on the Cauvery River. However, in reality, there may be a system of reservoirs that affect each other. That is, the release of one reservoir may add to the inflow, beside rainfall, to the inflow of other downstream reservoirs. A more comprehensive modeling needs to consider a system of reservoirs with their spatial relationships.

Further experiments may answer the question whether the accuracy can be improved if the training data is larger or contains more historical data in each row.

Implementing the exhaustive approach to include all the related watershed data, such as, temperature and evaporation rates, could also improve the accuracy of the water reservoir release model.

6. NOTES

Disclaimer: Since this research paper was completed and submitted to *QUeST'12*, the second author, Peter Revesz, was awarded an *AAAS Science & Technology Policy Fellowship* and as part of the fellowship program took a leave of absence from the University of Nebraska-Lincoln to serve as a grants Program Manager in the U.S. Air Force Office of Scientific Research (AFOSR). The views and opinions expressed in this publication are those of the authors and do not necessarily reflect the official policy or position of any agency of the U.S. government.

Acknowledgement: Our heartfelt thanks to the Public Works Department in Tamil Nadu, India for enabling us to use the Mettur Water dam data.

7. REFERENCES

[1] Abebe, A.J., Solomatine, D. P., and Venneker, R.G.W. (2000) Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events, *Hydraulic Science Journal*, Vol. 45 (3), pp 425-436.

[2] Abrahart, R. J. and See, L. (2000). Comparing neural network and auto regressive moving average techniques for the provision of continuous river flow forecasting two contrasting catchments. *Hydrological Processes, Vol.* 14, pp. 2157-2172.

[3] Alshaikh, A. and Taher, S. (1995). Optimal design of irrigation canal network under uncertainty using response surface method. *Water Int.*, 20, 155–162.

[4] Bhattacharya, B ., Lobbrecht, A.H., and Solomatine D.P. (2003). Neural networks and reinforcement learning in control of water systems, *Journal of Water Resources Planning and Management*, 129(6), 458-465.

[5] Bhattacharya, B., and Solomatine, D.P. (2005). Neural networks and M5 model trees in modeling waterlevel-discharge relationship, *Neuro Computing Journal*, 63, 381-396.

[6] Chandramouli. V., H. Raman. (2001). "Multi-reservoir modeling with dynamic programming and neural network." *J. Water Resource Plan Management,* 127 (2), pp. 89–98.

[7] Chaves, F.J., and Chang (2008). "Intelligent reservoir operation system based on evolving artificial neural networks", *Advances in Water Resources*, 31, pp. 926-936.

[8] Dawson, C.W. and Wilby, R. (1998). An artificial neural network approach to rainfall- runoff modeling. *Hydrological Sciences J*., 43(1), 47-66.

[9] Gates T. K., and Alshaikh, A. A. (1993). ''Stochastic design of hydraulic structures in irrigation canal networks.'' *Irrigation and Drain. Engineering*, 119(2), 346–363.

[10] Hall, M. J. and Minns, A.W. (1999). The classification of hydrologically homogeneous regions. *Hydrological Sciences J*, 44, 693-704.

[11] Kantardzic, M., "Data mining: Concepts, models, methods and algorithms."*Journal of Computer Information Science Engineering*, Vol., No. 1, pp.393, 2003.

[12] Nageshkumar, D. and Dhanya, C. T. 2009. Data Mining and its Applications for Modeling Rainfall Extremes, *ISH Journal of Hydraulic Engineering*, Vol. 15(1),pp.25-51.

[13] Neelakantan T. R., N. V. Pundarikanthan. (2000). "Neural network-based simulation- optimization model for reservoir operation". *J. Water Resource Planning and Management*, 126(2) 2, pp. 57-64.

[14] Revesz, P. and T. Triplet (2011), "Temporal Data Classification Using Linear Classifiers," *Information Systems*, vol. 36, no. 1, pp. 30–41.

[15] Simonovic , S. P., 1999. "Prototype Virtual Database Development for Management of Floods in the Red River basin", Canadian Civil Engineer, Vol. 16, No.6. pp.12-15.

[16] Solomatine, D.P., Rojas, C., Velickov, S., and Wust, H. (2000) Chaos theory in predicting surge water levels in the North Sea*, Proceedings of the 4th International Conference on Hydroinformatics*, Iowa, USA.

[17] Solomatine, D . P ., Tores L.A. Avila. (1996) Neural network approximation of a hydrodynamic model in optimizing reservoir operations. *Proceedings of the Hydro Informatics Conference*, pp. 201-206.

[18] Sudha, V, Ambujam NK, Venugopal K. 2006. "A data mining approach for deriving irrigation reservoir operating rules". *Conference on Water Observation and Information System for Decision Support.*

[19] Taghi Sattari, M., HalitApaydin, Fazli Ozturk & Nazife Baykal (2012): Application of a data mining approach to derive operating rules for the Eleviyan irrigation reservoir, *Lake and Reservoir Management*, 28:2, 142-152.

[20] Damle C. and Yalcin A (2007). ―Flood prediction using time series data mining. *Journal of Hydrology,* 333, 305–316.

[21] Choy K. Y. and Chan C.W. (2003). ―Modelling of river discharges using neural networks derived from support vector regression. *IEEE International Conference on Fuzzy Systems*, The University of Hong Kong, Hong Kong, China.

[22] Elshorbagy, A., Jutla, A. & Kells, J. (2007). ―Simulation of the Hydrological Processes on Reconstructed Watersheds Using System Dynamics. *Hydrological Sciences Journal*, 52(3), 538–562.

[23] Trafalis, T.B., Richman, M.B., White, A., and Santosa, B. (2002). ―Data Mining Techniques for Improved WSR-88D Rainfall Estimation.*Computers & Industrial Engineering,* 43, 775–786.

[24] Sahoo, G.B., Schlawdowa, S.G., and Reuter, J.E. (2009). ―Forecasting Stream Water Temperature Using Regression Analysis, Artificial Neural Network, and Chaotic Non-Linear Dynamic Models. *Journal of Hydrology*, 378, 325–342.

[25] Frawley, W., Piatetsky-Shapiro G., and Matheus C.. *Knowledge discovery in databases: An overview.* In Piatetsky-Shapiro, G. and Frawley, W. editors, *Knowledge Discovery in Databases*, MIT Press, Cambridge, MA, 1991.

[26] Teschl, R., Randeu, W.L., and Teschl, F. (2007). ―Improving Weather Radar Estimates of Rainfall Using Feed-Forward Neural Networks. *Neural Networks,* 20, 519–527.

[27] Kanevski, M., A. Pozdnukhov, S. Canu, M. Maignan, P. Wong, S. Shibli, Support Vector Machines for Classification and Mapping of Reservoir Data, A chapter from "Soft computing for reservoir characterization and modeling", Springer-Verlag, pp. 531-558, 2001.

[28] Li Deren ,Di Kaichang, and Li Deyi 1997. A Framework of Spatial data mining and knowledge discovery. *In proc Int. Workshop on Image Analysis and Information Fusion (IAIF'97)*, Adelaide, Australia, Nov.

[29] Wade, T.D., Byrns, P.J.,Steiner, J.F.and Bondy ,J.1994, 'Finding temporal patterns- a set based approach '.*Artificial Intelligence in Medicine.* (6):263-271.

[30] Berger, G. and Tuzhilin, A. 1998. 'Discovering unexpected patterns in temporal data using temporal logic'. In *Temporal Databases-Research and Practice.* O. Etzion, S. Jajodia and S. Sripada (eds.), Lecture Notes in Computer Science 1399, Springer-Verlag, Berlin. 281-309.