River Flow Forecastingby a Dynamic K-Nearest Neighbors Method

EhsanEbrahimi[1], MojtabaShourian[1*]

[1] Faculty of Civil, Water and Environmental Engineering, Technical and Engineering College, ShahidBeheshti University, Tehran, Iran
[*]Email: m_shourian@sbu.ac.ir

Abstract

River flow prediction is an important aspect for robust water resources planning and flood warning systems operation. Data driven approaches have been found efficient to this end. K-nearest neighbors (KNN) is a lazy learning method that can be used for this purpose. In this study, a new method for selecting neighbors named dynamic number of K nearest neighbor (DKNN) is introduced which uses an optimized distance to select a different number of neighbors for each instance of predictors instead of using a fixed K number as in the classic method. The Particle Swarm Optimization (PSO) algorithm is used for optimization process and for improvement of results. Performance of the proposed method is tested using two years of the daily inflow to the Gheshlagh reservoir in west of Iran. Results indicate that the proposed method increased the accuracy of prediction 4.6% by reducing RMSE compared to the classic KNN.

Keywords: KNN, River Flow, Prediction, Dynamic

1. Introduction

An important issue for surface water resources planning is to predict accurately the inflow to a reservoir so that planning and management of the said reservoir would be more effective and efficient [1]. Physical-based and data driven models can be used to predict the river flow [2]. Physical-based models are usually time consuming, complex and difficult to be used [3]. So to alleviate the problems, data driven approaches (DDA) are proposed vastly for river flow prediction. Although DDA do not simulate the hydrologic processes, they can be used to accurately predict the river flow and help in the process of planning. Some of the popular methods that are widely used for river flow prediction are multiple linear regressions (MLR), non-parametric regressions like K-nearest neighbors (KNN), artificial neural network (ANN), adaptive neuro-fuzzy inference systems (ANFIS) and support vector machines (SVM) [4, 5].

The KNN method is one of the widely used methods for variety of problems such as classification [5], clustering [6] and regression [7]. In this field, Galeati used KNN for predicting daily inflow and compared results to autoregressive model with exogenous input (ARX) method. The results showed that both methods have good performance but the KNN have simpler structure and due to this simplicity it's better to be used in larger scales [8]. Shamseldin and O'Connor introduced Nearest Neighbor Linear Perturbation Model (NNLPM) for river flow prediction. Their proposed model showed better accuracy in prediction of non-seasonal flows compared to simple linear mode and linear perturbation model [9]. Lall and Sharma introduced a method for resampling flow of monthly data and it showed to be effective with time series generated with autoregressive models [10]. Souza Filho and Lall used KNN to disaggregate annual flow prediction to monthly or higher resolution flows. Their method maintained space-time consistency in different places and sub-periods and showed ability to predict river flow up to 18 month ahead of time [11]. Laio et al showed that KNN performed slightly better in short term flood estimation in comparison with ANN [12]. Leander et al. used KNN method to resample extreme flood conditions. They found out that KNN models underestimate less extreme daily discharges [13]. Solomatine et al. showed that using KNN is more suited for short term inflow prediction than ANN [14]. Wu et al. tested KNN and other DDA coupled with preprocessing techniques to predict stream flow. Results showed that models performed better when fed preprocessed data [15]. Hilaire et al. used KNN to predict river water temperatures and showed that KNN can be dependable as a tool to predict water temperatures [16]. Liu et al. used and compared KNN for real time flood prediction to Kalman filter and showed KNN performs better in longer lead time prediction [17].

The power of KNN lies in the facts that it follows a simple algorithm and it can be used for linear and nonlinear problems. But the results of this method are dependent on choosing the optimum number of neighbors and spikes in error occurrence for extreme values predicts [18]. Domeniconi et al. found out that when the query

points are not uniformly disturbed finding out the optimized k is very difficult. So different applications and cases need a distinct optimal value of k and finding this out is still a challenge [19]. Liu et al. proposed a new method to alleviate the challenge of finding an optimal value of k. Their method used mutual nearest neighbor to determine the class for an unknown query. They showed that the strength of this approach is outlier neighbors affecting the accuracy of prediction can be identified and excluded from the procedure. The results showed better classification performance to the classic KNN [20].

In the present study, an improvement is proposed in the classic KNN method for river flow prediction. Instead of using a fixed number of neighbors, all neighbors within an optimized distance of the predictors are used. So, for every instance of river flow data the number of contributing neighbors may differ. The benefits of this approach are:

1. Maximum number of neighbors within the optimized distance can be used without degrading the accuracy.
2. Increased number of training vectors gives us the option to extract more information and features from data.
3. The prediction accuracy for a certain interval can be improved without changing the result of prediction for other intervals.

The distances and the weights of the neighbors are optimized by the particle swarm optimization algorithm. The results of the proposed method is compared with the classic KNN using the statistical criteria.

2. Methodology

The classic KNN method, the proposed dynamic number of neighbors, the PSO algorithm and the assessment criteria are explained in the following sections.

2.1. K-Nearest Neighbors Algorithm

KNN methods in general use and compare the similarities (in our study distances) between predictors and historical data to calculate best estimation of dependent variables of that particular set of predictors [21]. Predictors or independent variables are used as input for prediction procedure. The algorithm of a classic KNN regression can be broken into the following steps:

1. Considering a vector consisting *m* independent variables as predictors $X = \{x_1, x_2, x_3, \ldots, x_m\}$ with Y as the dependent variable.
2. Considering a training set containing the historical data with *n* vectors of $X_t = \{x_{1t}, x_{2t}, \ldots, x_{mt}\}$ and a dependent variable of Y_t associated with each vector.
3. Calculating the distance of the predictor vector with each of the *n* training vectors. The Euclidian distance function which is used in this study is [21]:

$$\Delta_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2} \quad (1)$$

4. K training vectors with the least distance to the predictor vector (K-nearest neighbors) are selected.
5. A Kernel function for each of the K selected training vectors is calculated as:

$$f_k(\Delta_k) = \frac{1/\Delta_k}{\sum_{k=1}^{K} 1/\Delta_k} \quad (2)$$

6. Estimated dependent variable is calculated as:

$$Y = \sum_{k=1}^{K} f_k(\Delta_k) \times y_k \quad (3)$$

2.2 KNN with Dynamic Number of Neighbors (DKNN)

The main difference between the proposed method of dynamic number of neighbors and the classic KNN is the way of selection of the training vectors. The steps of this method are as following:

1. Considering number of intervals and assigning a distance number to each interval (distance will be optimized using PSO):

$$\begin{cases} D_1 & i_1 < Predictor \leq i_2 \\ D_2 & i_2 < Predictor \leq i_3 \\ & \vdots \\ D_n & i_n < Predictor \leq i_{n+1} \end{cases} \quad (4)$$

D_j is the j'th interval optimized distance, i_j is the j'th interval's start point and n is the number of intervals.
2. Predictors based on their greatness are placed in one of the intervals.
3. Every training vector is examined and gets selected if its absolute distance is equal or less than the distance assigned to the interval the predictors are placed in.
4. All of the selected vectors contribute (which the number of vectors would be different for each predictor) and the estimation of dependent variable follows as the classic KNN with calculating the Euclidian distance and the Kernel function. The benefits and advantages of this method are:
1. Maximum number of neighbors within the optimized distance can be used without degrading the accuracy.
2. Increased number of training vectors gives us the option to extract more information and features from data.
3. The prediction accuracy for a certain interval can be improved without changing the result of prediction for other intervals.

2.3. Particle Swarm Optimization Algorithm

The particle swarm optimization (PSO) is one of many swarm intelligence methods used for solving optimization problems. PSO can be easily implemented and since computational and memory requirements are relatively low, it is inexpensive to compute. In PSO, a population (swarms) of candidate solutions are moved around search space using velocity and weight formulas. These formulas are effected by each of the particle's best known position in the search space as well as the entire swarm's best position in each iteration. So if a better position is found in an iteration it will guide the movements of the swarm to that position. The two main PSO equations that calculate velocity and position of each particle in the search domain in every iteration (i) are:

$$v_{i+1} = K\left[w_i \times v_i + C_1 r_1 \left(p_{best_i}{}^{-x_i}\right) + C_2 r_2 \left(G_{best_i}{}^{-x_i}\right)\right] \quad (5)$$

$$x_{i+1} = x_i + v_{i+1} \quad (6)$$

$K$is constriction factor, w_i is inertia weight, C_1 and $C_2$are cognitive and social parameters respectively, $r_1$and r_2 are uniformly disturbed random number between 0 and 1 and p_{best} is the best previously visited position of the particle while G_{best} is the best position visited across all particles.

2.4 Assessment Criteria

In this study, parameters used in each of the proposed method and techniques are optimized with PSO. To ensure that the best result is obtainedby PSO, each of the variations is optimized 10 times and best result is selected. The optimization is used to maximize the Nash-Sutcliffe coefficient in each method. Three performance criteria to assess the proposed method and subsequent techniques are used as follow:
1. Nash-Sutcliffe Error [22]:

$$NSE = 1 - \frac{\sum_{t=1}^{n}(T_t - Y_t)^2}{\sum_{t=1}^{n}(T_t - \bar{T})^2} \quad (7)$$

2. Root Mean Square Error [23]:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(T_t - Y_t)^2}{n}} \quad (8)$$

3. Correlation coefficient:

$$R = \frac{\sum_{t=1}^{n}(T_t - \bar{T})(Y_t - \bar{Y})}{\sum_{t=1}^{n}(T_t - \bar{T}) \cdot \sum_{t=1}^{n}(Y_t - \bar{Y})} \quad (9)$$

n is the number of prediction instances, $T_t$is the observed flow, $Y_t$is the model estimated value, $\bar{T}$ is mean of the observed flows and $\bar{Y}$ is mean of model estimated values in above equations. To assess the performance of the proposed method and subsequent techniques in extreme conditions and high flow and low flow values, RMSE is calculated once for validation data that are greater than the average and again for data below the average.

3. Case study

Predicting the daily inflow to the Gheshlagh Reservoir in west of Iran is the purpose of this study. This dam supplies the water demands of the Sanandaj city and is also used for recreational and sport purposes. In Figure 3, location of the Gheshlagh Dam in Iran is shown.

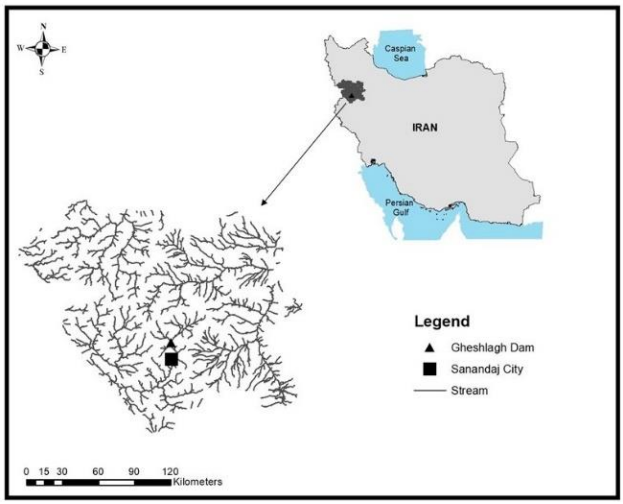

Figure 1: Location of the Gheshlagh Dam in Iran

The daily mean inflow to the reservoir for 10 years from 2007 to 2017 is used for training, calibration and evaluation. The first six years are used for training, the seventh and eighth years are used for calibration and the last two years are used to evaluating the proposed method. Predictors or independent variables are used as input for the prediction procedure. The predictors used are inflow of days prior to the day being predicted. The method proposed, work with any number of days as predictors but in this study two, three, four and five days prior are used as the predictors and best performing number of days for each variation are selected in calibration phase and used in verification to compare the results. In Figure 4, 10-years daily inflow to the Gheshlagh dam is shown.

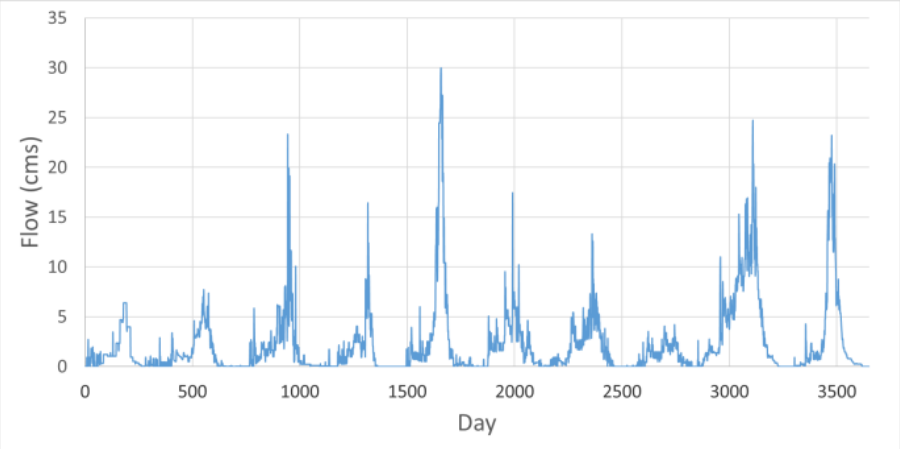

Figure 2: 10 year's daily time series of the Gheshlagh reservoir inflow

4. Results and Discussion

To assess the performance of the proposed method, Gheshlagh Dam inflow is predicted with the classic KNN and the dynamic K-nearest neighbors (DKNN). Optimum number of contributing neighbors, their weights and number of predictor are obtained by PSO using the calibration data set and are applied for the verification data set. In Figure 3, result of the DKNN is shown.

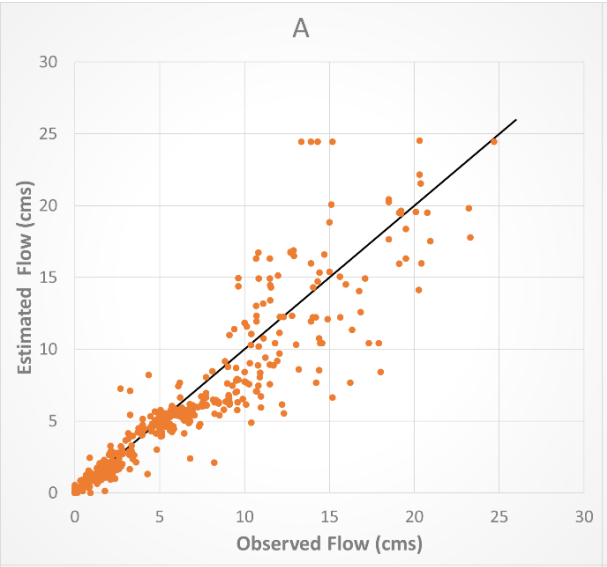

Figure 3: Observed vs. Predicted flow by the modified KNN method (DKNN)

In Figure 4, result of the DKNN is presented and compared to observed data and the classic KNN method results.In Table 1, result of various assessment criteria mentioned in section 2.5 are presented.

Table 1: Values of the assessment criteria for DKNN method and subsequent techniques

Method	Nash-Sutcliffe	Total RMSE	Above mean data RMSE	Below mean data RMSE	Correlation coefficient
KNN	0.869	1.84	3.02	0.29	0.96
DKNN	0.880	1.76	2.87	0.35	0.94

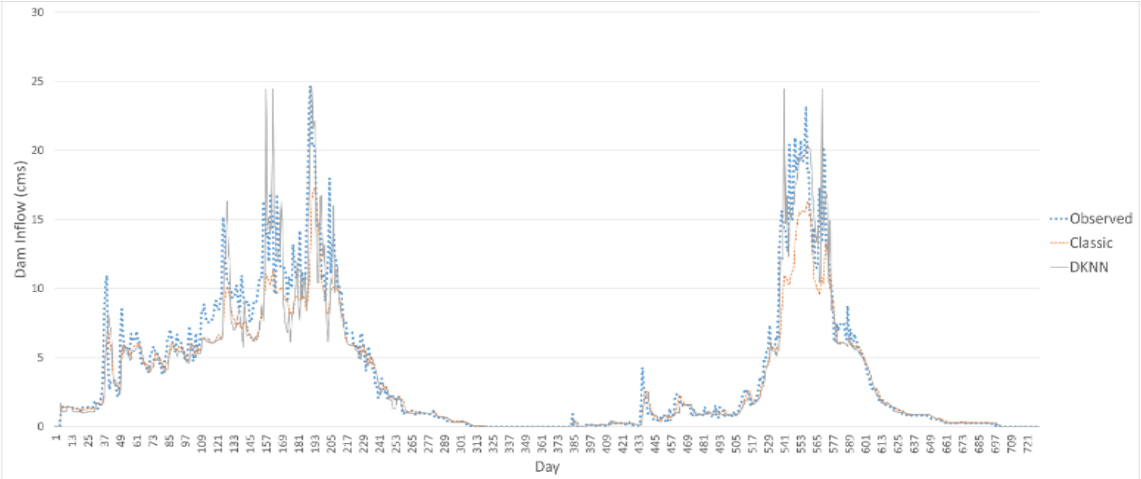

Figure 4: Time series of the observed vs. classic KNN and DKNN estimated flow

The result of the verification process shows better performance by the dynamic K number of neighbors (DKNN) method than the classic KNN. Looking at Table 1, the NS coefficient is increased from 0.869 for the classic KNN to 0.880 for DKNN and total root mean square error (RMSE) is decreased from 1.84 to 1.76 which shows 4.6% improvement in accuracy of prediction. It should be noted that in testing some other data, these two methods had similar performance but with added benefits by the DKNN method which make it superior. These benefits are the ability to use maximum number of neighbors for each prediction instance without sacrificing the performance of the whole prediction which opens up the opportunity to extract features and information from more variety of data for each prediction instance. For example, in some tests, the optimum number of neighbors for the classic KNN obtained very small, so a neighbor with a little more distance than the last selected neighbor which may have better features is neglected, while in the DKNN method, these neighbors contribute to the prediction procedure. Another benefit is the ability to change and optimize each dynamic interval without affecting other intervals.

It should be noted that considering more intervals and thus optimizing more specific weights for each of the selected training neighbors would give better results for the modified KNN method but increasing the number of intervals grows the optimization time and therefore the computational cost. In this research, the number of intervals was assumed constant and the same number was used across all variations and tests. So, the number of intervals and their start points and end points were always the same. Selection of this number was done by trial and error and the best performing number of intervals was used in analyses.

5. Concluding Remarks

In this study, a novel approach for selecting neighbors in the KNN method and three techniques for prioritizing the contributing neighbors are introduced. The new approach called (DKNN) uses different number of neighbors for each prediction based on the greatness of the predictors. The proposed method was applied to predict two-year daily inflow to the Gheshlagh dam in Iran. Result showed that the approaches using the dynamic selection had better performance than the classic KNN with added benefits. These benefits include selecting maximum number of neighbors for each prediction without sacrificing the performance of the whole procedure, increased number of selected neighbors gives the option for gaining more information and the ability to improve performance of certain intervals without changing the result of other intervals.

6. References

1. Araghinejad, S., Fayaz, N., &Hosseini-Moghari, S. M. (2018). Development of a hybrid data driven model for hydrological estimation. Water resources management, 32(11), 3737-3750.
2. Sivakumar, B., Jayawardena, A. W., & Fernando, T. M. K. G. (2002). River flow forecasting: use of phase-space reconstruction and artificial neural networks approaches. Journal of hydrology, 265(1-4), 225-245.
3. Hadi, S. J., &Tombul, M. (2018). Monthly streamflow forecasting using continuous wavelet and multi-gene genetic programming combination. Journal of hydrology, 561, 674-687.
4. Ahani, A., Shourian, M., and Rad, P. R. (2018) Performance assessment of the linear, nonlinear and nonparametric data driven models in river flow forecasting. Water Resources Management, 32(2), 383-399.
5. KhazaeePoul, A., Shourian, M. &Ebrahimi, H. (2019) A Comparative Study of MLR, KNN, ANN and ANFIS Models with Wavelet Transform in Monthly Stream Flow Prediction. Water Resources Management, Doi: 10.1007/s11269-019-02273-0.
6. Zhang, M. L., & Zhou, Z. H. (2005). A k-nearest neighbor based algorithm for multi-label classification. GrC, 5, 718-721.
7. Liu, T., Moore, A. W., & Gray, A. G. (2003, December).Efficient Exact k-NN and Nonparametric Classification in High Dimensions.In NIPS (pp. 265-272).
8. Kramer, O. (2011). Unsupervised K-nearest neighbor regression. arXiv preprint arXiv:1107.3600.
9. Galeati, G. (1990). "A comparison of parametric and non-parametric methods for runoff forecasting." Hydrological Sciences Journal 35(1): 79-94.

10. Shamseldin, A. Y., & O'Connor, K. M. (1996). A nearest neighbour linear perturbation model for river flow forecasting. Journal of Hydrology, 179(1-4), 353-375.

11. Lall, U., & Sharma, A. (1996).A nearest neighbor bootstrap for resampling hydrologic time series. Water Resources Research, 32(3), 679-693.

12. Souza Filho, F. A., &Lall, U. (2003). Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm. Water Resources Research, 39(11).

13. Laio, F., Porporato, A., Revelli, R., &Ridolfi, L. (2003).A comparison of nonlinear flood forecasting methods. Water Resources Research, 39(5).

14. Leander, R., Buishand, A., Aalders, P., & Wit, M. D. (2005). Estimation of extreme floods of the River Meuse using a stochastic weather generator and a rainfall–runoff model/Estimation des cruesextrêmes de la Meuse à l'aide d'un générateurstochastique de variables météorologiqueset d'un modèlepluie–débit. Hydrological Sciences Journal, 50(6).

15. Solomatine, D. P., Maskey, M., &Shrestha, D. L. (2008). Instance-based learning compared to other data-driven methods in hydrological forecasting. Hydrological Processes: An International Journal, 22(2), 275-287.

16. Wu, C. L., Chau, K. W., & Li, Y. S. (2009). Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resources Research, 45(8).

17. St-Hilaire, A., Ouarda, T. B., Bargaoui, Z., Daigle, A., &Bilodeau, L. (2012). Daily river water temperature forecast model with ak-nearest neighbour approach. Hydrological Processes, 26(9), 1302-1310.

18. Liu, K., Yao, C., Chen, J., Li, Z., Li, Q., & Sun, L. (2017). Comparison of three updating models for real time forecasting: A case study of flood forecasting at the middle reaches of the Huai River in East China. Stochastic Environmental Research and Risk Assessment, 31(6), 1471-1484.

19. Araghinejad, S. (2013). Data-driven modeling: using MATLAB® in water resources and environmental engineering(Vol. 67). Springer Science & Business Media.

20. Domeniconi, C., Peng, J., &Gunopulos, D. (2002).Locally adaptive metric nearest-neighbor classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(9), 1281-1285.20. Liu, H., et al. A new classification algorithm using mutual nearest neighbors.in 2010 Ninth International Conference on Grid and Cloud Computing. 2010. IEEE.

21. Karlsson, M., &Yakowitz, S. (1987).Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. Water Resources Research, 23(7), 1300-1308.

22. Nash, J. E., & Sutcliffe, J. V. (1970).River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology, 10(3), 282-290.

23. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.