International Conference on Hydroinformatics

2014

River Flows Prediction By Ensemble Model

Milan Cisty

Celar Lubomir

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

RIVER FLOWS PREDICTION BY ENSEMBLE MODEL

MILAN CISTY, LUBOMIR CELAR
Department of Land and Water resources Management, Faculty of Civil Engineering, Slovak University of Technology, 813 68 Bratislava, Slovak Republic
milan.cisty@stuba.sk, lubomir.celar@stuba.sk

This work presents the application of a data-driven model for stream flow predictions. A methodology was investigated in which ensemble modeling by data-driven models was applied, and harmony search was used to optimize the ensemble structure. The proposed ensemble provides a better degree of precision in the prediction task, which was evaluated as a case study in comparison with the ensemble components, although they were powerful algorithms themselves. For this reason the proposed methodology could be considered as a potential tool in flood predictions and predictions tasks in general.

INTRODUCTION

This work presents the application of a data-driven model for stream flow predictions. In the past it was usual to search for a model optimized in some way, e.g., to find the "best" model. Nowadays, it is accepted in the hydrology modeling community that there is no best model which is superior under all circumstances [1]. The recognition of this fact has led to the application of an ensemble of models being simultaneously considered. Many researchers have shown that by combining the output of many predictors, more accurate predictions can be produced than what could be obtained from any of the individual predictors [2-4].

Ensemble approach has been adapted also in the hydrology field. A boosting application is presented by [5], where the authors demonstrated the advantages of an improved version of boosting, namely, AdaBoost.RT, which is compared to other learning methods for several benchmarking problems, and two problems involving river flow forecasting. In a recent study [5], the authors investigate the potential usage of bagging and boosting in building classification and regression tree ensembles to refine the accuracy of streamflow predictions. They report that the bagged model performs slightly better than the boosted model in the testing phase. An ensemble neural network (ENN) designed to monthly inflow forecasting was applied in [7] to the Daecheong dam in Korea. The ENN combined the outputs of the members of a neural network employing the bagging method. The overall results showed that the ENN outperformed a simple artificial neural network (ANN) among the three rainfall-runoff models. Cannon and Whitfield [8] studied the use of ensemble neural network modeling in streamflow forecasting. Boucher et al. [9] used bagged multi-layer perceptrons for the purpose of a 1-day-ahead streamflow forecasting on three watersheds.

In general, the ensemble methods are usually composed of weak predictors, e.g., decision trees or neural networks commonly used as base predictors. On the opposite, a major goal of the analysis in this study is to evaluate ensembles composed of various strong machine learning algorithms. The final prediction by the proposed ensemble is accomplished by weighted summation of the results of the individual learners. The specification of these weights is proposed to be solved with the help of the harmony search optimization methodology [10].

In the following part of the paper data used and the methods involved in this study are briefly explained, together with the ensemble methodologies used. In the "Results and Discussion" part, the settings of the experimental computations are described and the results evaluated. Finally, the "Conclusion" part of the paper summarizes the main achievements and conclusions of the work and proposes ideas for future work in this area.

DESCRIPTION OF CASE STUDY AND PREPARATION OF DATA

Ensemble modeling by data-driven methods was applied for the two-day ahead prediction of flows on the Hron River in Slovakia in Banska Bystrica gauging station. The watershed of this river is a sub-basin of the Danube River. This task was accomplished by using data observed in the period from 1.1.1984 to 31.12.2000. Specifically, the average daily flow [$m^3.s^{-1}$], the average daily temperatures [°C] and the daily rainfall depths [mm] were used. Each row in the input file for this task includes the input data of the flows from the three measuring stations, the temperatures from five meteorological stations, and the precipitation from 51 stations. All the input data were included in the input data set from 1, 2, 3 and 4 days before the date of the predicted flow.

The period from 1996 to 2000 includes many situations with high flows and floods, which was the reason for its selection as the testing period. The rest of the years (1984-1996) were used for the training.

METHODOLOGY

The proposed ensemble methodology for predicting the river flows is divided into four equally important steps (Figure 1). The final model is predicted using the weighted average of the base learners in which these weights are used.

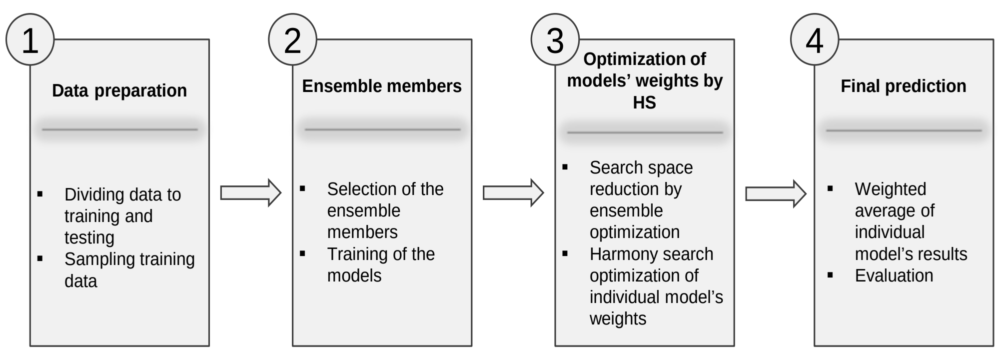

Figure 1. Proposed steps for the development of ensemble predictions of river flows

Ensemble Members

The authors of the paper wanted to answer the question as to whether improvements in performance are obtained by ensemble modeling of river flow predictions in comparison with each of the ensemble members' performances, in a case where these members are already powerful algorithms with good performances. The authors of this article are aware of some degree of subjectivity in the choice of the strong algorithms which were included in the proposed ensemble, but some supporting information in the data mining community exists [11, 12]. Following algorithms were used in this study: Support Vector Machines (SVM), Multilayer perceptron (MLP), Random Forest (RF), Multiple-linear regression (MLR), Generalized linear model with an elastic-net (GLMNET), Multivariate adaptive regression splines (MARS), Boosted linear models (B_GLM), Gradient Boosting with Smooth Components and Gradient boosting machines (B_GAM) .

Usual grid search combined with a repeated cross-validation methodology was used for finding the parameters of all the models included in the ensemble. A description of the selected algorithms is neither possible nor useful in this paper. It could be found in the relevant literature from machine learning, e.g. [4].

Ensemble Design

In this section the computation procedures which are necessary for obtaining the ensemble model are described. The ensemble model is proposed to have the following structure:

$$P_{ensemble}^{t} = \sum_{i=1}^{n} \beta_i * P_i^t \quad (1)$$

where βi are the weights of the models of which the ensemble consists, and P_i^t are the predictions by these models in time t. In this study two-day ahead flow predictions by ensemble modeling are evaluated; the flows of which are computed n times for each day, where n is the number of models. The weights of every model in the ensemble are proposed to be found by the harmony search methodology [10].

One harmony consists of n members, where n is the number of models. In the case of this work there are nine models present in the ensemble. All values of the weights βi are restricted to the interval $\langle 0, 1 \rangle$. The problem solved should be defined by the objective function, which is proposed in this paper to have the following form:

$$O_f = 1 - \left(1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} (O_i - \bar{O})^2}\right) + \left|1 - \sum_{i=1}^{n} \beta_i\right| \quad (2)$$

$$0 \leq \beta_i \leq 1 \quad (3)$$

where O_i are the observed flows, N is the number of such data, and $\bar{O}$ is their average value. The last component of the objective function (as an absolute value) forces the sum of the ensemble members` weights β_i to be equal to 1. This objective function is proposed to be minimized.

One of the main issues which must be carefully considered is what exactly has to be data, which will serve as inputs to the harmony search optimization objective function (2). There are two possibilities evaluated in this study as to how to obtain such data. The first possibility is achieved using the following steps:

1. The training data and repeated cross validation are used for finding the proper parameters of each model
2. Every model (ensemble member) is trained with the values which were found in step 1 with all training data
3. The values of the predicted flows are computed by the models from step 2 from all the training data for each ensemble member. The number of rows of resulted input matrix for HS $P_{R,C}$ is equal to the number of the rows of training data (535 in this study) and the number of columns $C = n+1$ (n is the number of models, and one extra column is the observed data). In this work $n = 9$.

The problem of obtaining data $P_{R,C}$ by this methodology, if it is used for calculating ensemble weights, is that in this approach there is no mechanism which avoids over-fitting of the final ensemble. Over-fitting or a lack of generalization means that the weights of the models obtained could work well on the training data, but poorly on the testing set. Due to this problem, the authors also proposed a second option, which will be compared to the previous one:

1. The training data and cross validation are used for finding the proper parameters of each model
2. When these parameters are obtained, the $k-1$ folds (in the case of a k-fold cross validation) are used for training with the best parameters, and 1 fold is computed by the model obtained as a test
3. This is repeated k times for every model included in the ensemble
4. Because the r-repeated cross-validation was proposed in this work, steps 2 and 3 are repeated r times
5. The computed values from all such testing folds from the cross validation are used as the input matrix for the optimization by HS, which is proposed for searching the weights of each model in the final ensemble
6. Consequently the inputs to the HS are de facto testing data, although from the training set (the results from the testing folds in the cross validation). When n is the number of models in the ensemble, N is the number of data in the training set, and r is the number of repeats of the cross validation, the number of rows of this input matrix $P_{R,C}$ is $R = N*r$, and the number of columns $C = n+1$ (one column is the observed data). In this work $n = 9$, $k = 10$, $N = 535$ (the data was reduced by the sampling!) and $r = 5$.

The ensemble models obtained from these two approaches are hereinafter identified as EHS1 for the first case and EHS2 for the second.

RESULTS

In (Table 1) the root mean square error, correlation coefficient, and Nash-Sutcliffe efficiency are evaluated for the ensemble members and the proposed ensembles. The identification of the models from their abbreviations in the heading of this table is possible. Two ensemble optimization approaches, which are identified as EHS1 and EHS2, are evaluated in this table and were described hereinbefore.

Table 1. Evaluation of the computations by r and NSE and the final values of the model weights in the ensembles

	GBM	B_GLM	RF	MLP	MARS	MLR	SVM	B_GAM	GLMNET	EHS1	EHS2
NSE	0.806	0.783	0.808	0.676	0.593	0.376	0.800	0.787	0.782	0.759	**0.825**
r	0.898	0.885	0.900	0.832	0.802	0.724	0.896	0.888	0.884	0.874	**0.909**
RMSE	13.575	14.371	13.519	17.548	19.661	24.355	13.788	14.219	14.410	9.684	**8.247**
Weights EHS1	0.128	0.011	0.190	0.549	0.021	0.022	0.032	0.003	0.045		
Weights EHS2	**0.134**	**0.056**	**0.379**	**0.034**	**0.083**	**0.021**	**0.218**	**0.029**	**0.046**		

Regarding ensembles EHS1 and EHS2, it can be clearly seen that the hypothesis about the poor performance of the above-mentioned first proposition for obtaining matrix $P_{R,C}$ was confirmed. Ensemble model EHS1 performed well on the training data (with an NSE equal to 0.82, when an NSE of 0.79 was achieved by the best ensemble component, which was the GBM model), but on the testing set, which is evaluated in (Table 1), the ensemble EHS1 gives worst results than most of the ensemble members. The ensemble approach to modeling is worth applying only in a case where the ensemble performs better than any of its members. If one considers the weights of the multilayer perceptron in ensemble EHS1, it is presumably inappropriately high (MLP are generally less precise models), which means that this model is over fitted and that the poor generalization is a consequence of the approach used for the development of the EHS1 model. To the contrary, according to (Table 1), in which the testing data are evaluated, the results with a good generalization were achieved by ensemble EHS2. From now on, we will only speak about this second model.

Column nine of (Table 1) with the evaluation of the ensemble members could also be seen as a case study of the evaluation of these models. The models are ordered from best to worst, so they can be ranked and compared with each other. However, when the weights of the models for the EHS2 ensemble in (Table 1) are considered, it can be seen that this order does not imply that the weights will also be ordered in the same way as precision. An efficient ensemble should consist of predictors that are not only sufficiently precise, but also diverse, i.e., ones that if make wrong predictions they make them at different parts of the input space, e.g., which are not highly correlated.

From the conjoint consideration of the (Table 1) (weights of models for the EHS2), it can be seen that, after optimization of the weights, the best three models, the GBM, RF, and SVM, are included in the proposed ensemble with the highest contribution (their weights are the highest). But the next best model, the boosted GAM (B_GAM), is included in the ensemble with a relatively small weight. That is because this model is highly correlated with the three best models mentioned and also with the GLMNET model. A similar case could also be observed with some other members of the ensemble. From this phenomenon it could be evaluated that the optimization procedure which was proposed in this paper is searching for the best weights not only from the point of view of the best performance of the models, but also is considering the diversity of the models as well, which is, as was mentioned, not less important. The authors assume that this is mainly due to the procedure by which was obtained matrix PR,C for model EHS2. As could be expected, the smallest contribution to the EHS2 ensemble has its least precise member – the multi-linear regression (MLR).

CONCLUSION

A new methodology of the river flow predictions was investigated in which ensemble modeling by data-driven models was applied and the harmony search was used to optimize the ensemble's structure. The authors were trying to evaluate in the case study presented (two-day's ahead prediction of river flows), whether an ensemble paradigm would also bring some gain in cases when strong algorithms are used as ensemble members. Although the improvement in precision was not relatively as high as in the case when the ensemble consists of weak learners, it was proved that the ensemble model worked better than any of its constituents.

According to the so-called "no free lunch" theorem, it is never clear in advance which machine learning algorithm suits best for a particular task. For this reason it is usually necessary to try more algorithms. Instead of selecting and using only the best algorithm, it is better to compose ensemble predictions based on all of these already tuned algorithms. Forming an ensemble usually brings an improvement in precision as was also confirmed for the case study in this paper (the results are in Table 1), and ensemble prediction is relatively easy to accomplish when tuned algorithms for a particular task are already available.

ACKNOWLEDGEMENTS

This work was supported by the Slovak Research and Development Agency under Contract No. APVV-0496-10 and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/1044/11 and 1/0908/11.

REFERENCES

[1] Duan Q., Ajami N.K., Gao X. and Sorooshian S., "Multi-model ensemble hydrologic prediction using Bayesian model averaging", *Adv. Water Resour*., Vol. 30, (2007), pp 1371–1386.

[2] Breiman L., "Bagging predictors", Machine Learning, Vol. 24, No. 2, (1996), pp. 123-140.

[3] Wheway V., “Variance reduction trends on ‘boosted’ classifiers”, *Journal of Applied Mathematics and Decision Sciences*, Vol. 8, No. 3, (2004), pp. 141-154.

[4] Hastie T., Tibshirani R. and Friedman J., “The Elements of Statistical Learning”, 2nd edition, (2009).

[5] Shrestha D.L. and Solomatine D.P., “Experiments with AdaBoostRT, an improved boosting scheme for regression”, *Neural Computation*, Vol. 18, No.7, (2006), pp 1678-1710.

[6] Erdal I.H. and Karakurt O., “Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms”, *Journal of Hydrology*, Vol. 477, (2013), pp 119-128.

[7] Jeong D.I. and Kim Y.-O., “Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction”, *Hydrological processes*, Vol. 19, No. 19, (2005), pp 3819-3835.

[8] Cannon A.J. and Whitfield P.H., “Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models,” *Journal of Hydrology*, Vol. 259, (2002), pp 136-151.

[9] Boucher M.A., Laliberté J.P. and Anctil F., “An experiment on the evolution of an ensemble of neural networks for streamflow forecasting”, *Hydrol. Earth Syst. Sci*, Vol. 14, No. 3, (2010), pp 603-612.

[10] Geem, Z.W., Kim J.H. and Loganathan G.V., “A new heuristic optimization algorithm: harmony search”, *Simulation*, Vol. 76, No. 2, (2001), pp 60-68.

[11] Rich C. and Niculescu-Mizil A., “An empirical comparison of supervised learning algorithms”, Proceedings of the 23rd international conference on Machine Learning, *ACM*, (2006), pp. 161-168.

[12] Rich C., Karampatziakis N. and Yessenalina, A., “An empirical evaluation of supervised learning in high dimensions”, *Proceedings of the 25th international conference on Machine Learning*, ACM, (2008), pp. 96-103.