A Novel Prediction Approach for Runoff Based On Hybrid HMM-SVM Model

Feng Chen[1, a], Yongqing Su[2,b], Yin Wang[3,c]

[1]Department of Electronics and Information Egineering, Tongji Univeristy, Shanghai, China,201804

[2]Department of Electronics and Information Egineering, Tongji Univeristy, Shanghai, China,201804

[3]Department of Electronics and Information Egineering, Tongji Univeristy, Shanghai, China,201804

[a]tj_chen_feng@163.com, [b]suyongqing@tongji.edu.cn, [c]13162199852@163.com

Keywords: Hidden Markov Model, Shape Based Clustering, SVM, Runoff Prediction

Abstract. This research demonstrates an application of Hidden Markov Model (HMM) and Support Vector Machine (SVM) for watershed-runoff forecasts. HMM is used for shape-based clustering by calculating log-likelihood values of each data to identify data in the data set with similar data pattern.Then we put these data into different classes based on their shapes and train their corresponding SVM model to predict the output of the system finally. The applications of daily runoff and monthly runoff are used for testing the competence of this method and experimental results demonstrate that this hybrid HMM-SVM algorithm can meet the prediction requirement and has high prediction accuracy.

INTRODUCTION

Hydrological systems are highly non-linear systems. Runoff is the result of a combination of climatic conditions and drainage area. The accuracy of runoff prediction is a key factor for reservoir operation and water resources management. However, due to the complexity of the atmospheric processes, runoff is one of the most complex and difficult elements to understand and simulate in the hydrological cycle.The effects of variations generating flow, including the watershed, topography, climatic characteristics and many of their combinations is a very complex physical process. Because of the complexity of the process, many researchers have begun to focus on runoff forecasts that only consider past runoff data[1].

Runoff time sequence one-dimensional data contains a wealth of information because each value in the time series represents the combined result that many different factors simultaneously acting. We reconstruct the chaotic time series into nonlinear dynamical systems by using reconstruction theory.

Then we use HMM classifier to deal with Multidimensional sequential inputs to output the probability of belonging to the homologous classes. After putting these data into different classes based on their shapes and training corresponding SVM models of different classess, the SVM is used to make the run-off prediction finally[2].

The prediction results show that the HMM-SVM model solves the problems of over-learning, small sample size, local minimum and high dimensionality effectively.It has a strong generalization ability and yields very satisfactory prediction results. Those data which has similar features with predicted data are obtained by dealing with large amounts of data based on HMM cluster method. Similar data are uesd as training samples as the input of SVM to reduce the amount of data to ensure data consistency and improve the accuracy of prediction[3].

The rest of this paper is organized as follows. In Section II,brief of HMM and SVM will be introduced. In Section III, the hybrid strategy will be presented. The simulation results will be compared and discussed in Section IV. We end this paper with conclusion in Section V.

BASIC THEORY

A.Hidden Markov Model

In this section, we briefly review the Hidden Markov model (HMM). Hidden Markov Model comes from the Markov process or Markov chain[4]. It is a canonical probability model of temporal or sequential data.HMM is a double embedded random process in which the final output of a system at a particular time depends on the system and the output generated by that state. The basic symbols used are explained below.

$\{O_1, O_2, O_3,..., O_n\}$:Observation Sequence; $Q\{q_1, q_2, ... ,q_n\}$:State Sequence;

π:Initialization Matrix; A:Transition Matrix; B:Emission Matrix/Function;

$\lambda(A, B, \pi)$:Model of the System.

Once the value of the Probability of the Observation sequence has been calculated, the value of the Log-Likelihood can be calculated.The solution to calculating the state sequence $Q\{q_1, q_2, ... ,q_n\}$, which is most likely or optimal, is given by the Viterbi Algorithm and the solution to adjusting the model parameters $\lambda(A, B, \pi)\lambda(A, B, \pi)$ is given by the BaumWelch Algorithm.Once we have the solutions to the above problems, the tools to shape based clusting of HMMs are available[5].

B. Support Vector Machine

SVM is a supervised learning method for classification and regression. The method is widely used in pattern recognition and data mining[6].

Consider a dataset $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i \in R^N (i = 1,2,3, ..., n)$ and $y_i \in \{-1,1\}$ is the class label of x_i. A hyperplane in the feature space can be described as $w * x + b = 0$. For linearly inseparable cases, there is no such a hyperplane that is able to classify every training sample correctly[6]. However the optimization idea can be generalized by introducing the concept of soft margin. The final decision function is

$$f(x) = sgn\{\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x) + b^*\} \qquad f(x) = sgn\{\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x) + b^*\}$$

(1)

HYBRID ALGORITHM OF HMM AND SVM

The time series used in this work as an example was generated by the chaotic Mackey-Glass differential equation defined by Eq.(2) below. This equation demonstrates chaotic behaviour when τ > 17; and higher values of τ yield higher dimensional chaos[8]. In this work a value of τ = 30 was employed and Fig.1 illustrates the first 1200 points of this series using an initial condition of y(0)= 1.0. The time series data was obtained by Eq.(2).

$$\frac{dy(t)}{dt} = \frac{0.2y(t-\tau)}{1+y^{10}(t-\tau)} - 0.1y(t) \qquad (2)$$

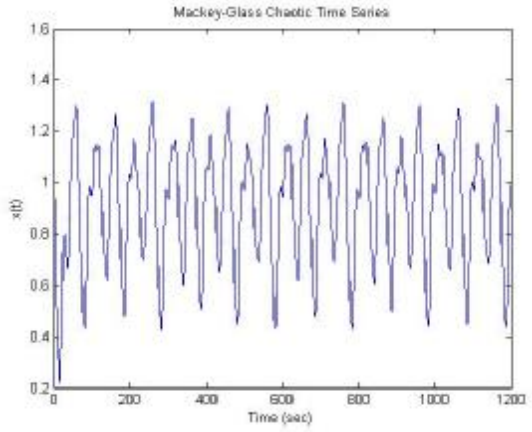

Fig 1. Section of the Mackey-Glass chaotic time series.

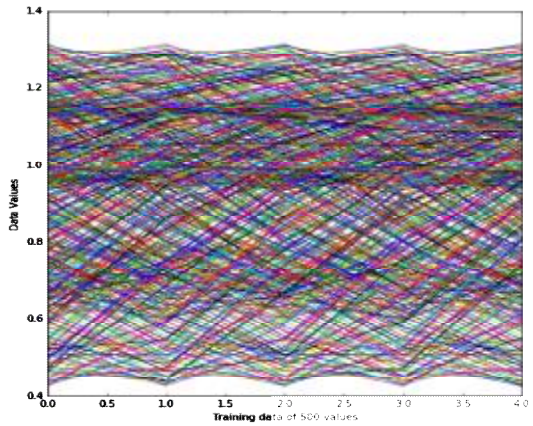

Fig 2. the data before shape-based clustering

The prediction of future values of this series is a benchmark problem which has been considered by a number of researchers[9]. The problem can be formulated as given values y(t-m), y(t-m+1),… , y(t-1); determine y(t-1+n), where m and n are fixed positive integers and t is the series index. There have been several different approaches to this problem such as used by Wang and Mendel[4] where m=8 and n=1 was selected(i.e. using 8 previous values of the series to predict the current value in the series). Here we have taken m=4 and n=1.

For the Mackey Glass Time series, the dataset consists of five elements, four predictor variables, and one dependent variable. Here we have taken 486 data points as training data and the next 486 data points as test data to show the efficiency of the method.The performance criterion for similarities for input data used here is known as Log-Likelihood.Log-Likelihood is defined as the Log of the value of the Probability of observation sequence given the HMM model.

First of all the value of the parameters of the HMM are randomly initialized. This includes initialization of π, A, B. The HMM is trained using the Baum Welch algorithm for the entire input dataset. Once the HMM has been trained, the Forward Algorithm is used to compute the value of $P(O|\lambda)$ which can then be used to calculate the Log-Likelihood of each dataset. By looking closely at the input datasets, it can be easily observed that data points which have similar values of Log-Likelihood have similar shape.

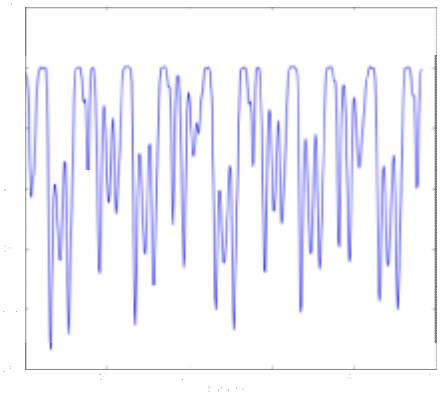

Fig 3. Log-Likelihood of Mackey-Glass data

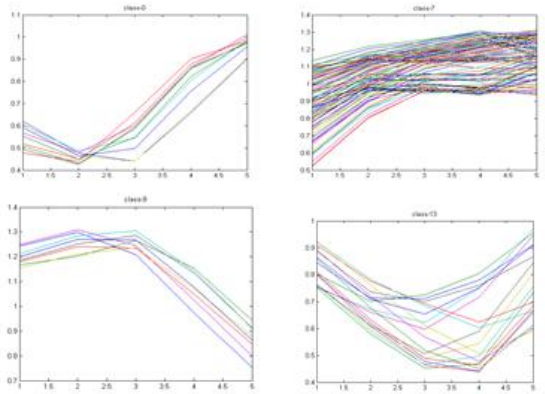

Fig 4. the data before shape-based clustering

Our aim is to convert this data into classes depending upon their shape. We first arrange all the datasets in increasing order of Log-Likelihood by using a standard sorting algorithm and then create classes of data by following a simple procedure. By carefully observing the datasets and their corresponding Log-Likelihoods, It can be easily observed that after a certain value of Log-Likelihood, the shape of the input changes and hence cannot be classified into the same batch.For the Mackey-Glass data, the maximum size of the class was 90 and the maximum Log-Likelihood difference between two consecutive batches was 37. Fig.2 and Fig.4 shows the data before and after shape-based clustering.

Once the data has been arranged into classes based on their shape, the next step involves the training of SVM to predict the results. Here the shape based clustering is done such that it becomes easier and accurate for a SVM to analyze the data and predict outputs. Fig.5 shows a flowchart of hybrid HMM and SVM for forecasting.

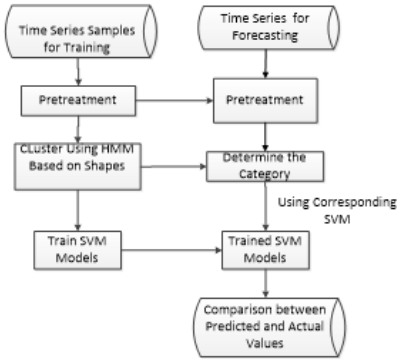

Fig.5 The steps for forecasting using hybrid HMM and SVM

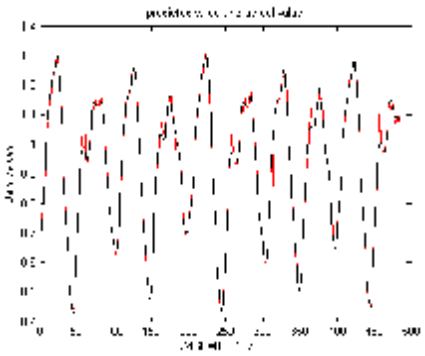

Fig.6 predicted values and actual values

The results of the predicted values are compared with the actual values,which is shown in Figures 6. This illustrates that the HMM-SVM model has high prediction accuracy with relatively small training samples used.The Root Mean Square Error(RMSE) for Mackey-Glass of this method is 0.00507886. The correlation coefficient between real and predicted values is 0.9968.

SIMULATION AND DISCUSSION

As shown above,then we apply this hybrid algorithm of HMM and SVM into runoff prediction. we examined the data obtained from the daily and monthly runoff of the St. Louis hydrological station located in the upper catchment of Mississippi River. The daily runoff data of St. Louis hydrological station,consisting of 365 daily records(January 1st,2015 to December 31th,2015),and the monthly runoff data ranging from October 1932 to July 2016,are used in this study. The dataset was split up into two parts:training and testing, Training data were used exclusively for model development and testing data were used to measure the performance of the model on untrained data.

The performance of the models in forecasting daily and monthly runoff during training and testing are evaluated by using the root mean squared error(RMSE),which are widely used for evaluating the results of time series forecasting. The RMSE are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{obs,i} - y_{model,i})^2} \quad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{obs,i} - y_{model,i})^2} \tag{3}$$

Where $y_{obs,i}$ is the observed value, $y_{model,i}$ $y_{model,i}$is the forecast value,and n is the number of data points.For the St. Louis Station, input combinations based on preceding daily and monthly runoff are evaluated to estimate current runoff value, Q(t-6),Q(t-5),Q(t-4),Q(t-3),Q(t-2) and Q(t-1) and Q(t-1) as inputs and the discharge Q(t) for the current day or month as the output.

Take the Radial Basis Function(RBF) as the kernel function,the best parameters sets of C(penalty factor) and g(RBF parameter) were achieved by cross validation. The daily and monthly predicted runoff obtained by HMM-SVM model and the corresponding observed values are compared in Fig 7. and Fig 8.

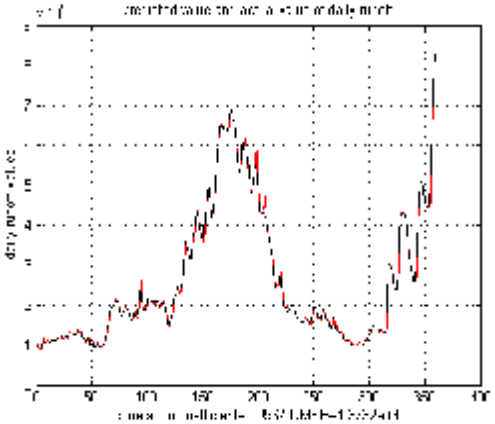

Fig.7 comparison diagram between actual measured

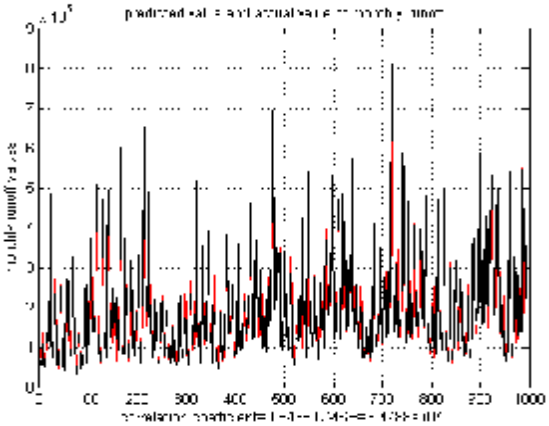

Fig.8 comparison diagram between actual measured

value and predictor for daily flow forecasting value and predictor for monthly flow forecasting

The prediction results show that there is a strong correlation between the real data and the forecast data and the runoff time sequence predicting model based on HMM-SVM has high prediction accuracy.

CONCLUSIONS

In this paper we have proposed a novel combination of experts technique for prediction of time series. This involves a Hidden Markov Model which performs shape based clustering for input data on the basis of their similarities in shape. Using the SVM for each class, we predict the output with a high degree of accuracy and it is clearly visible from the results that our method outperforms other past approaches for the prediction of the standard Mackey Glass Time series.

As shown above, we also experiment with daily and mouthly runoff time sequence.The prediction results show that the runoff time sequence predicting model based on HMM-SVM has high prediction accuracy and can meet the prediction requirement and thus is an effective prediction method. In practical work, newly added samples into the training can continuously improve the predicting model.

Acknowledgements

This work was supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (Grant No. 2014BAL05B02).The authors would like to thank all the reviewers and editors for their helpful comments and suggestions on improving the quality of this paper.

References

[1]S. Srinivasulu and A. Jain, "River flow prediction using an integrated approach," Journal of Hydrologic Engineering, vol. 14, pp. 75-83, 2009.

[2]J. Adamowski, "Development of a short-term river flood forecasting method based on wavelet analysis,"Geophysical Research Abstracts, pp. 09556, 2007.

[3]G. Henkelman, G.Johannesson and H. Jónsson, in: Theoretical Methods in Condencsed Phase Chemistry, edited by S.D. Schwartz, volume 5 of Progress in Theoretical Chemistry and Physics, chapter, 10, Kluwer Academic Publishers (2000).

[4]A.B. Poritz, “Hidden Markov models: a guided tour,”in Proc. of ICASSP, 1988, pp. 7-13.

[5]Yariv Ephraim and Neri Merhav,”Hidden Markov Processes,”IEEE Transactions on information theory, vol. 48, no. 6, June 2002.

[6]N.Cristianini,J.Shawe-Taylor.An Introduction to Support Vector Machines and other Kernel-based Learning Methods.London:Cambridge University Press,2000.

[7]M. Karamouz, A. Ahmadi, A. Moridi Probabilistic reservoir operation using Bayesian stochastic model and support vector machine .Advances in Water Resources vol. 32, pp.1588-1600,

[8]M.C. Mackey, L. Glass, Oscillation and chaos in physiological control systems, Science 197(1977) 287-289.

[9] G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day,Oakland, CA, 1976.