TIDAL CURRENT SHORT-TERM PREDICTION BASED ON SUPPORT VECTOR REGRESSION

Yang Guozhen¹, Wang Haifeng², Qian Hui³ and Fang Jianming³

¹ Institute of Electrical Engineering, Chinese Academy of Sciences; University of Chinese Academy of Sciences, Beijing 100190, China;
² Institute of Electrical Engineering, Chinese Academy of Sciences, Beijing 100190, China;
³ Zhejiang Chunan Stage Grid Corporation of China, Zhejiang 311700, China.

Email: abandonclock@gmail.com

Abstract: The traditional method of short-term tidal current prediction, harmonic method, typically needs more than 18 years of history records. The method in the article uses univariate feature selection and F-test to reduce the dimension of the data fed to support vector regressor, which reduces the need of history records to less than a year. Model parameters are selected by grid searching and cross-validation. History records from two datasets are used to build prediction models, spanning 3 months and 1 year respectively. Mean average errors of both datasets after normalizing are less than 0.05.

1. Introduction

With the consumption of fossil fuels raising more awareness of environment protection and sustainable development, hydrokinetic power generation technology is burgeoning over last few years. In short-term scale, tidal current has some randomness which affects the stability of the power grid with hydrokinetic generators. Short-term prediction of tidal current meet the requirements of economic dispatching of power grid.

The traditional method of predicting tidal current in short-term scale is harmonic method. Harmonic method assumes that the tidal current speed time series is consisted of several sinusoidal components. By extracting each component and summing up the prediction of each component, the prediction of the whole time series is achieved.

Despite the simplicity of harmonic method, it has some shortcomings. First, due to sea surface wind, geostrophic current and shape of basins or channels, there is randomness in tidal current speed time series, which makes it not strictly sinusoidal. Harmonic method cannot predict the nonsinusoidal distortion[1]. Second, harmonic method typically uses more than 18 years of history records to extract all the sinusoidal components[1][2]. However, in practice the volume of available history records is usually much smaller. History records deficiency leads to error if using harmonic method. Meanwhile, projects implementing hydrokinetic power generation are mostly in experiment state, which implies fewer history records can be accessed.

Many articles are proposed to solve these problems. Some articles uses Kalman filter reduce the volume of history records when calculating the parameters of harmonic components[3][4]. Some articles uses artificial neural network to build a prediction model[5-9].
The support vector machine generally need fewer samples and has better generalization. It’s been widely applied in time series prediction[10-12]. This article proposes a method of building data-driven model based on support vector regression (SVR) to predict tidal current speed using fewer history records while preserving good accuracy.

2. Support vector regression

This article uses SVR to build prediction models. SVR is a method of making regression using support vector machine. The prediction model is denoted as \( f(X) \), where the input vector \( X \) is consisted of several history data at certain time. The output \( f(X) \) is just the prediction of tidal current speed. Suppose there is a training set \( \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\} \), where \( X_i \in \mathbb{R}^d \), \( y_i \in \mathbb{R} \), \( 1 \leq i \leq n \), and \( d \) denotes the dimension of feature vectors. By building a model named \( \varepsilon \)-SVR model, a function \( f(X) \) is obtained, which deviates \( \varepsilon \) at most from target values \( y_i \) for all training samples, meanwhile maintaining the most simplicity of the model to achieve the most affordable generalization. When the error between the output and the target value is less than \( \varepsilon \), the \( \varepsilon \)-SVR model doesn’t try to decease the error. Instead it only simplifies itself under less-than-\( \varepsilon \) constraints.

Suppose the model \( f(X) \) is a linear function, denoted as in (1).

\[
f(X) = \langle w, X \rangle + b
\]

The corresponding optimization problem of \( \varepsilon \)-SVR model is shown in (2) and (3).

\[
\arg\min_{w \in \mathbb{R}^d} \frac{1}{2} ||w||^2
\]
\[
\text{Subjected to } \begin{cases}
(\langle w, X_i \rangle + b) - y_i \leq \varepsilon \\
 y_i - (\langle w, X_i \rangle + b) \leq \varepsilon
\end{cases}
\]

The optimization problem may be infeasible. Under such cases, two slack variables, \( \xi_i \) and \( \xi_i^+ \), are introduced to allow more error so that the problem becomes feasible. The \( \varepsilon \) margin is not strict anymore since more error is affordable. The optimization problem of soft-margin \( \varepsilon \)-SVR model is shown in (4) and (5).

\[
\arg\min_{w \in \mathbb{R}^d} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^+) \quad (4)
\]
\[
\text{Subject to } \begin{cases}
(\langle w, X_i \rangle + b) - y_i \leq \varepsilon + \xi_i \\
y_i - (\langle w, X_i \rangle + b) \leq \varepsilon + \xi_i^+
\end{cases}
\]

The model parameter \( C \) is the penalty parameter, which is larger than 0.

The optimization problem in (4) and (5) is called the prime problem. To solve the prime problem, a Lagrange function is constructed by summing the objective function and all constrains multiplied by Lagrange multipliers. By applying Karush-Kuhn-Tucker conditions (KKT conditions) to the Lagrange function, the dual problem is obtained, which is shown in (6) and (7).

\[
\arg\max_{\alpha_i, \alpha_i^+} -\frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^+) (\alpha_j - \alpha_j^+) (X_i, X_j) - \varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^+) + \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^+). \quad (6)
\]
\[
\text{Subject to } \begin{cases}
\sum_{i=1}^{n} (\alpha_i - \alpha_i^+) = 0 \\
\alpha_i \in [0, C] \\
\alpha_i^+ \in [0, C]
\end{cases} \quad (7)
\]

The dual problem is independent of separate feature vectors but only dependent of inner product of feature vectors. By applying KKT conditions to the Lagrange function, the weight vector \( w \) turns out in form shown in (8).

\[
w = \sum_{i=1}^{n} (\alpha_i - \alpha_i^+) X_i \quad (8)
\]
When $|f(X_i - y_i)| \leq \epsilon$, $\alpha_i$ and $\alpha'_i$ are all zero, which means corresponding feature vectors make no contribution to the weight vector and feature vectors lying in the less-than-$\epsilon$ region make no contribution to the model. Only those lying outside the less-than-$\epsilon$ region have influence on the model. Those feature vectors are called support vectors.

The model $f(X)$ assumed to be a linear model to derive the equations all above. When feature vectors and target values are non-linearly related, they need to be transformed into a higher dimension space, in which they are approximately linearly related. In the transformed feature space, the model building procedure stays same as shown above.

But the transformation caused great trouble when calculating the inner product of transformed feature vectors, due to the high dimensionality, which leads to infeasibility of the whole problem[13]. Function kernels help to calculate the inner product of the transformed feature vectors implicitly at much less cost.

Polynomial kernel, radial basis function kernel (RBF kernel) and sigmoid kernel are all widely used kernels. RBF kernel is mostly used in non-linear model regression. RBF kernel is shown in (9).

$$K(X, X') = \exp(-\gamma \|X - X'\|^2)$$  \hspace{1cm} (9)

3. Model building

The tidal current speed vector is decomposed into east component and north component. And the models of the two components are built separately.

3.1 feature selection

Feature vector input into the model is a main adjustable variable. Since the volume of records is usually limited, feature vectors become sparser in higher dimension feature space, which leads to low prediction accuracy[14]. Features highly correlated with target values should be selected to reduce feature vector’s dimension.

This article used univariate analysis to choose features. By calculating Pearson product-moment correlation coefficient (CORR), the strength of linear correlation between features and target value is acquired. Then by calculating F-ratio and corresponding p-value, certain features will be chosen. Though this method is usually used in feature selection of linear models, RBF kernel is introduced to transform the non-linearity into linearity. So it’s feasible to use this method for feature selection in non-linear cases[15].

The calculation of CORR is shown in (10). It is bounded in range $[-1, 1]$, where zero indicates no correlation, positive values indicate positive correlated and vice versa.

$$\text{CORR} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$  \hspace{1cm} (10)

Fig. 1 shows CORR of samples between the data to be predicted and each history data from 96 hours ago to 24 hours ago. CORR varies nearly sinusoidally, which is apparently related to the semidiurnal tide cycle and diurnal tide cycle. There are several features highly related to the data to be predicted. By calculating CORR, the range of history records from which the features are selected is determined.

In order to make sure features selected correlates to the population from which the history data are sampled, features are selected by F-test. F-ratio is calculated by (11).

$$F = \frac{\text{CORR}^2(n-2)}{(1-\text{CORR}^2)}$$  \hspace{1cm} (11)

where $n$ is the number of samples.

The null hypothesis that the feature is not correlated with the data to be predicted is valid when CORR equals 0. Otherwise, the alternative hypothesis that the feature is correlated to the data to be predicted is accepted.
By calculating p-value, the null hypothesis is judged to be accepted or not. p-value is obtained by the survival function of F-distribution as in (12).

\[ p(f) = 1 - F(f) \] (12)

where \( p(f) \) is the p-value when F-ratio equals \( f \) and \( F(f) \) is the cumulative distribution function of F-distribution \( F(f; 1, n - 2) \).

Fig. 2 shows the F-ratio and corresponding p-value between the data to be predicted and each history data from 96 hours ago to 24 hours ago. Most p-values are much less than 0.05, due to the volume of records. Features with high p-values are deleted from feature vectors to reduce dimension.

3.2 data pre-processing

After choosing features, the data set is divided into a training set \( \{ (X_1, y_1), (X_2, y_2), ..., (X_n, y_n) \} \) and a test set \( \{ (X_{n+1}, y_{n+1}), (X_{n+2}, y_{n+2}), ..., (X_{n+m}, y_{n+m}) \} \). Generally, the volume of training set is larger than test set.

For simplicity of data manipulation and accelerating the optimization, all data are rescaled into a certain range, typically from -1 to 1.

3.3 Parameter tuning

SVR with RBF kernel has three parameters adjustable, the penalty parameter \( C \), the margin parameter \( \varepsilon \) and the RBF kernel parameter \( \gamma \).
The parameter C mainly controls the strength of penalty when a feature vector doesn’t lie in the less-than-ε region. When \( |f(X_i - y_i)| \leq \varepsilon \), the sum of slack variables \( \xi_i + \xi_i' \) turns zero so the parameter C has no effect on the objective function. When \( |f(X_i) - y_i| \geq \varepsilon \), the slack variables turn positive and the parameter C controls how large the slack variables can be. The model fits training set better if the parameter C grows larger. But a larger C may cause overfit situation in which the model fits history data great but has poor predicting accuracy.

The parameter ε controls the extent to which the deviation is affordable. More points lie in the less-than-ε region while ε grows larger, which leads to fewer support vectors. From (8), only support vectors contribute to the model. A larger ε leads to a simpler model, which fits poorer on history data but may have better generalization. A smaller ε means better fit on history data but may lead to overfit.

The RBF parameter γ controls influence of support vectors in transformed feature space. Smaller γ means support vectors would influence more neighboring feature vectors, which simplifies the model. While larger γ makes support vectors only focus on feature vectors closely located, which leads to better fit on history data.

Choosing optimized parameters is basically a tradeoff between generalization and variation pattern recognition. The model would have terrible predicting accuracy with parameters not being tuned well. This article uses grid search method and cross validation to find relatively acceptable parameters. Grid search traverse a bounded subset of parameter space to find the optimized parameters. Cross validation is used to rule out overfitting cases for better generalization. By evaluating error of case, a relatively good model is picked out.

3.4 Error evaluation
This article uses root-mean-squared error (RMSE), mean absolute error (MAE) and coefficient of determination \( r^2 \) as error evaluating criteria.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - y_i| \quad (13)
\]

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (f(x_i) - y_i)^2}{n}} \quad (14)
\]

\[
r^2 = CORR^2 \quad (15)
\]

4. Result and performance analysis
This article uses data from experiment site of Institute of Electrical Engineering, Chinese Academy of Sciences (dataset I) and open data of site cb0102 from U.S. National Oceanic and Atmospheric Administration (dataset II).

For concern of the volume of available data, calculation overhead and prediction accuracy, this article selects 10% features from history data ranged from 96 hours to 24 hours before the data to be predicted. The grid searches in Grid searches parameter C in range \( \{C_i - 3 \leq \log_{10} C_i \leq 3, C_i \in \mathbb{Z}\} \), parameter ε in range \( \{0.05, 0.1\} \) and parameter γ in range \( \{y_i - 3 \leq \log_{10} y_i \leq 3, y_i \in \mathbb{Z}\} \) with 5-fold cross validation.

Prediction error between actual values and prediction values is evaluated after model is built. Table 1 shows RMSE, MAE and \( r^2 \) of the error. Actual values and prediction values are plotted in Fig.3 and Fig.4. One reference shows two series are strongly correlated when 0.6 ≤ CORR < 0.8 and very strongly correlated when CORR ≥ 0.8[16]. Another reference shows two series are strongly correlated when CORR ≥ 0.5[17]. No matter which criteria is chosen, the prediction result and the actual data are strongly correlated at least.
Table 1. Dataset I Prediction error after normalization

| Component      | RMSE  | MAE   | $r^2$  | CORR  |
|----------------|-------|-------|--------|-------|
| East component | 0.0618| 0.0492| 0.7722 | 0.8931|
| North component| 0.0564| 0.0441| 0.2661 | 0.5158|

Table 2. Dataset II Prediction error after normalization

| Component      | RMSE  | MAE   | $r^2$  | CORR  |
|----------------|-------|-------|--------|-------|
| East component | 0.0626| 0.0496| 0.8005 | 0.8947|
| North component| 0.0534| 0.0423| 0.6028 | 0.7764|

Figure 3. Prediction result of dataset I

$\tau^r$ of dataset II is little bit larger than data I’s. This is caused by the difference of volume of available data. Dataset I contains history data of about 3 months. Dataset II contains history data of about 1 year which is much more. Model built with more data has better prediction accuracy.

From Table.1 and Table.2, despite RMSE and MAE of east component and north component are nearly equal, $r^r$ of east component is larger than north component’s. The prediction accuracy of the whole east component series is better than north component series’. The north component has worse predictability than the east component, which can be revealed by calculating autocorrelation of each component.

Autocorrelation shows correlation between the latest data and each history data in a series. It’s calculated as in (16).

$$ R(\tau) = \frac{\langle (X - \mu)(X - \mu) \rangle}{\sigma^2} \tag{16} $$

where $\mu$ and $\sigma^2$ are average and variance of the series respectively. $\tau$ is the difference of indexes of latest data and history data.
Figure 5. Autocorrelation of each component series in dataset II

Fig. 5 shows the autocorrelation of dataset II’s east component and north component. It’s worth pointing out that the east component has better autocorrelation than the north component, though they have close prediction accuracy. The east component has much better autocorrelation, which lies more symmetrically about x-axis and has larger amplitude, than the north component. All these suggest that the east component varies more regular so there’s larger possible that a future value can be deduced by history values.

5. Conclusion
Support vector regression and univariate feature selection successfully reduce the demand of history records. The datasets used for building prediction model in this article contains history records spanning 3 months and 1 year respectively, which is much less than the records of 18 years required by the harmonic method. The mean average error of two datasets are both less than 0.05 after normalization. The tidal current short-term prediction method is this article is feasible and accurate.

References
[1] Steward Robert 2009 Introduction to physical oceanography (Florida: Orange Grove Texts Plus)
[2] Consoli Sergio, Recupero Diego Reforgiato, Zavarella Vanni 2014 A survey on tidal analysis and forecasting methods for Tsunami detection. Science of Tsunami Hazards 33(1) 1-56
[3] Kalman Rudolph Emil 1960 A New Approach to Linear Filtering and Prediction Problems Journal of Basic Engineering 82(1) 35–45
[4] Mizumura Kazumasa 1984 Applications of Kalman Filter to Oceanic Data Journal of Waterway, Port, Coastal, and Ocean Engineering 110(3) 334–343
[5] Yen Pei-Hwa, Jan Chyan-Deng, Lee Youe-Ping, et al 1996 Application of Kalman filter to short-term tide level prediction Journal of waterway, port, coastal, and ocean engineering 122(5) 226–231
[6] Li Ming-chang, Liang Shu-xiu, Sun Zhao-chen 2007 Application of artificial neural networks to tide forecasting Journal of Dalian University of Technology 1 101-105
[7] Huang Wenrui, Murray Catherine 2003 Application of an artificial neural network to predict tidal currents in an inlet ERDC/CHL CHETN-IV-58
[8] Lee Tsong-Lin 2004 Back-propagation neural network for long-term tidal predictions Ocean Engineering 31(2) 225–238
[9] Tsai Ching-Piao, Lee Tsong-Lin 1999 Back-propagation neural network in tidal-level forecasting Journal of Waterway, Port, Coastal, and Ocean Engineering 125(4) 195–202
[10] Zhang Hua, Zeng Jie 2010 Wind speed Forecasting model study based on support vector machine Acta Energiae Solaris Sinica 7 928-932
[11] Wang Ben, Leng Bei-xue, Zhang Xi-hai, et al 2011 Application Profiles of Support Vector
Machine in Short-term Load Forecasting *Proceedings of the Chinese Society of Universities for Electric Power System and its Automation* (4) 115-121

[12] Pai Ping-Feng, Lin Kuo-Ping, Lin Chi-Shen 2010 Time series forecasting by a seasonal support vector regression model *Expert Systems with Applications* 37(6) 4261–4265

[13] Smola Alex, Schölkopf Bernhard 2004 A tutorial on support vector regression *Statistics and computing* 14(3) 199–222

[14] Bishop Christopher 2006 *Pattern recognition and machine learning* (Berlin: Springer)

[15] Mao K. Z 2002 RBF neural network center selection based on Fisher ratio class separability measure *IEEE Transactions on Neural Networks* 13(5) 1211–1217.

[16] Evans James 1996 *Straightforward statistics for the behavioral sciences* (California: Brooks/Cole)

[17] Cohen Lawrence 1988 *Life events and psychological functioning: Theoretical and methodological issues* (New York: Sage Publications)