Neurocomputing 157 (2015) 243–255

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Vessel traffic flow forecasting by *R*SVR with chaotic cloud simulated annealing genetic algorithm and KPCA

Ming-Wei Li, Duan-Feng Han *, Wen-long Wang

College of Shipbuilding Engineering, Harbin Engineering University, Harbin 150001, Heilongjiang, China

ARTICLE INFO

Article history:
Received 26 October 2014
Received in revised form
13 December 2014
Accepted 7 January 2015
Communicated by Wei Chiang Hong
Available online 20 January 2015

Keywords:
Vessel traffic flow forecasting
Chaotic mapping
Cloud model
Genetic algorithm (GA)
Simulated annealing (SA)
Support vector regression (SVR)

ABSTRACT

The prediction of vessel traffic flow is complicated, its accuracy is influenced by uncertain socio-economic factors, especially by the singular points existed in the statistical data. Recently, the robust ν-support vector regression model (*R*SVR) has been successfully employed to solve non-linear regression and time-series problems with the singular points. This paper will firstly propose a novel hybrid algorithm, namely chaotic cloud simulated annealing genetic algorithm (C^{cat}CSAGA) for optimizing the parameters of *R*SVR, to improve the performance of vessel traffic flow prediction. In which, the proposed C^{cat}CSAGA employs cat mapping to carefully expand variable searching space, to overcome premature local optimum, and uses cloud model efficiently to search a better solution in a small neighborhood of the current optimal solution, to improve the search efficiency. Secondly, the kernel principal component analysis (KPCA) algorithm is adopted to determine the final input vectors from the candidate input variables. Finally, a numerical example of vessel traffic flow and its influence factors data from Tianjin are employed to test the forecasting performance of the proposed K*R*SVR-C^{cat}CSAGA model.

1. Introduction

Vessel traffic flow prediction is fundamental to development planning of national shipping and distribution of regional economic coordination. It also has important guiding significance to port layout planning, cooperation development of regional shipping economic, construction and renovation of waterway, and layout planning of national water rescue.

Numerous various forecasting approaches have been developed for vessel traffic flow prediction. In the conventional quantitative forecasting approaches, the autoregressive moving integrated moving average (ARIMA) models [1] and its improved models (more complicate and with high-accurate-capability) [2] and Kalman filtering model [3] are the most popular and practical time series forecasting model. They are often applied to forecast the series when data are inadequate to construct econometric, or when knowledge of the structure of forecasting models is limited. Time series models are simple in calculation and fast in speed, and are likely to outperform other models in some cases, specially, in short-term forecasting [4,5]. However, time-series forecasting models fail to reflect other related factors of the predicting series, hence, they fail to obtain the accurately forecasting result, when the predicting sequences are affected by the related factors to a large extent.

Artificial neural network (ANN) is primarily based on a model of emulating the processing of human neurological system identify related spatial and temporal characteristics from the historical data patterns (particularly for nonlinear and dynamic evolutions), therefore, they can approximate any level of complexity and do not need prior knowledge of problem solving. Since the vessel traffic flow prediction is too complex to be solved by a single linear statistical algorithm, ANN should be considered as an alternative for solving vessel traffic flow forecasting problem. Due to superior performance to approximate any degree of complexity and without prior knowledge of problem solving, ANN models have been widely applied in traffic flow forecasting [6–8]. Even though ANN-based forecasting models can approximate to any function, particularly nonlinear functions, ANN models have difficulties in the non-convex problem of network training errors and explaining black-box operations, and are easy to trap in local minima [9,10], ANN models have time-consuming training procedures and subjectivity in selecting an ANN model architecture [11]. Additionally, the training of ANN model requires large amount of training samples, while vessel traffic flow and related impact indicators have limited datum.

With the structure risk minimization criterion, support vector regression (SVR) has overcome the inherent defects of the ANN model [12]. SVR-based models [13] have been widely employed to receive higher forecasting accuracy in many fields, such as electric

* Corresponding author. Tel.: +86 18845615108; fax: +86 451 82568317.
E-mail address: dfhan163@163.com (D.-F. Han).

http://dx.doi.org/10.1016/j.neucom.2015.01.010
0925-2312/

load forecasting [14–20], atmospheric science forecasting [21–24], financial time series forecasting [25–29], tourist arrival prediction [30,31] and traffic flow prediction [32–34].

The conclusions of these researches all indicate that the selection of the parameters in an SVR model play a critical role in forecasting accuracy improvement [35]. Although, some recommendations on appropriate setting of SVR parameters have been given in the literature [36], however, these suggestions do not contemporaneously consider the interactive effects among parameters. In order to obtain more appropriate parameters of SVR model, authors have conducted a variety of systematical researches [37–41] by applying various evolutionary algorithms (particle swarm optimization and genetic algorithms) to determine suitable parameter values. In which, all SVR models with parameters determined by different evolutionary algorithms are superior to other competitive forecasting models (ARIMA and ANNs etc.), however, based on the analysis of the research results, these employed algorithms almost have their theoretical drawbacks, such as population diversity decline, convergence time consuming, easy to fall into local optimum. Therefore, authors start another trials by hybridizing chaotic sequence and cloud model with evolutionary algorithms to overcome these shortcomings [42,43]. To continue testing the superiorities of these hybridized chaotic sequence and cloud model with evolutionary algorithms, this paper tries to use the chaotic sequence and cloud model to improve the SAGA, for obtaining more appropriate parameters of SVR model.

GA is a simulated evolution optimization algorithm, in which, new individuals are generated by selection, crossover, and mutation operators. Based on special binary coding process, GA is able to solve some specified problems which are not easily to be solved by traditional algorithms. Therefore, it has been widely used in function optimization, neural network training, pattern classification, fuzzy control system and other engineering fields nowadays [44]. In previous papers [45], GA can empirically obtain a few best fitted offsprings from the whole population, however, due to population diversity decreases after some generations, it might lead to premature convergence. Simulated annealing (SA) is a generic probabilistic search technique that simulates the material physical process of heating and controlled cooling [46]. SA attempts to replace the current state by a random move in each step. The new state can be accepted with a probability that depends both on the temperature and difference between the corresponding functional values. Thus, SA has the ability to get more optimal solution [47]. However, in previous paper [48], SA is time consuming in annealing process. To speed up the search time and improve premature convergence, it is deserved to establish some effective approach to overcome these drawbacks of GA and SA. Hybridization of GA with SA (SAGA) algorithm is an innovative attempt by employing the superior capability of SA algorithm to produce better solution, and by applying the mutation process of GA to reduce the convergence time. Due to the ergodicity of the cat mapping and the randomness and the stable orientation of the cloud model, authors have employed cat mapping and cloud model to improve the PSO [42]. Application results show that the introduction of chaotic sequence and cloud model enrich the diversity of population and reduce the convergence time. To continue testing the superiorities of these hybridized chaotic sequence and cloud model with evolutionary algorithms, improve the search performance of SAGA, receive more appropriate parameters, this paper tries to use the chaotic sequence and cloud model to modify SAGA, by applying chaotic sequence to carefully expand variable searching space, let variable travel ergodically over the searching space, and employing the cloud model to efficiently search a better solution in a small neighborhood of the current optimal solution. Then, a novel hybrid algorithm, namely chaotic cloud simulated annealing genetic algorithm (C^{cat}CSAGA), is proposed, expecting to obtain more suitable parameter combination of SVR model.

The mixed noise existed in forecasting sequence data and related impact factors datum will largely affect the final prediction results, especially on sensitive SVR model. Considering mixed noise of normal distribution, high amplitude values and singular point features in datum of prediction sequence and related impact factors, a robust loss function [49] is designed and a new support vector regression (namely *R*SVR) is obtained. Application results show that the *R*SVR model can effectively suppress noise and lead to better prediction results [48]. Therefore, to deal well with the mixed noise in vessel traffic flow sequence and its related impact factors, this paper adopts the *R*SVR model to improve robustness and accuracy of vessel traffic flow prediction.

The prediction of vessel traffic flow is a complex nonlinear dynamic procedure. The vessel traffic flow is affected by numerous factors, such as gross domestic product, per capita gross national product, total imports and exports, passenger throughput etc. In order to ensure forecasting accuracy, related factors should be selected as the input vector of *R*SVR model, however, high dimensions of input vector will not only increase the operation time consuming, but also reduce the forecasting accuracy of *R*SVR model, therefore, it is very necessary to reduce the dimension of the relevant factors. As one of the most classic dimension reduction method, KPCA is a nonlinear extension algorithm of linear PCA, it extracts principal component by adopting the nonlinear method, maps the input space to a high-dimensional space through some implicit way, and realizes the PCA in the feature space. KPCA has received intense attention and been widely used in many fields, such as, face recognition [50], image recognition [51], spectral dimension reduction analysis [52]. In order to save computing time consuming and improve the model prediction accuracy, this paper employs KPCA to analyze generation mechanism of vessel traffic flow and determine the input vector of forecasting model.

Consequently, the K*R*SVR-C^{cat}CSAGA model, hybridizing *R*SVR with C^{cat}CSAGA and KPCA algorithm, is established, to enhance forecasting accurate level of vessel traffic flow. Five other competitive forecasting approaches, the KSVR-C^{cat}CSAGA, P*R*SVR-C^{cat}CSAGA, K*R*SVR-$C^{logistic}$SAGA, KGRNN and ARIMA models are employed to compare the forecasting accuracy.

The remainder rest of this paper is organized as follows. The fundamental principle and formulation of *R*SVR, the principle and calculation steps of KPCA which is employed to determine the input vector of the *R*SVR model and the C^{cat}CSAGA which is used to select the parameters of the SVR model are presented in Section 2. Section 3 describes the proposed K*R*SVR-C^{cat}CSAGA vessel traffic flow forecasting scheme. A numerical example is provided in Section 4. The conclusions are shown in Section 5.

2. Methodology

2.1. Robust support vector regression (RSVR)

Suppose the training set $T=\{(x_1,y_1),\ldots,(x_i,y_i),\ldots,(x_N,y_N)\}$, where $x_i\in R^n$ is n-dimensional input variable, and $y_i\in R^n$ is the corresponding output value, $i=1,2,\ldots,N$. Through a nonlinear mapping function, $\phi(x)=\{\phi(x_1),\phi(x_2),\ldots,\phi(x_i)\}$, the SVR model maps the sample into a high dimensional feature space, R^m, in which the optimal decision function is constructed as follows,

$$f(x)=\omega\times\phi(x)+b \tag{1}$$

where ω, $\phi(x)$ is m dimensional vector, $b\in R$ is threshold, "$\cdot$" is the dot-product in the feature space. SVR uses structural risk minimization to compute Formula (1), the core idea of structural risk