International Journal of Innovative Technologies
WWW.IJITECH.ORG
ISSN 2321-8665
Vol.03,Issue.09,
October-2015,
Pages:1640-1643

SST Gap Filling: A Case Study Based On Big Data

S. DEVIKUMARI[1], J. PAVAN KUMAR[2], B. SANGAMITHRA[3], THATIMAKULA. SUDHA[4]
[1]PG Scholar, Dept of CSE, Sri Padmavati Mahila Visva Vidhyalayam School of Engineering and Technology, Tirupati, AP, India, E-mail: sivvala.devi@gmail.com.
[2]Project SCIENTIST B, DMG, INCOIS, MoES, E-mail: pavankumar.j@incois.gov.in.
[3]Assistant Professor, Dept of CSE, Sri Padmavati Mahila Visva Vidhyalayam School of Engineering and Technology, Tirupati, AP, India, E-mail: mithra1209@gmail.com.
[4] HOD (I/C), Dept of CSE & IT, Sri Padmavati Mahila Visva Vidhyalayam School of Engineering and Technology, Tirupati, AP, India, E-mail: thatimakula_sudha@yahoo.com.

Abstract: Large scale satellite data are generated continuously by multiple sensors in daily communications. Forecast on such data possesses high significance for analyzing the behaviors of huge amounts of data. However, the natural properties of satellite data present three non-trivial challenges: large data scale leads it difficult to keep both efficiency and accuracy; similar data increases the system load; and noise in the data set is also an important influence factor of the processing result. To resolve the above problems, we can work efficiently with the neural networks on large data sets. Data is divided into separated segments, and learned by a same network structure. Then all weights from the set of networks are integrated and renovate the conventional back propagation neural network to the next layer using big data techniques. A Hadoop based framework called HBNN (i.e. Hadoop-based Back propagation Neural Network) is proposed to process forecast on large-scale SST data. It utilizes a diversity-based algorithm to decrease the computational loads. Extensive experiments on gigabyte of realistic SST data are performed on a Big Data platform and the results show that HBNN is characterized by greater efficiency, good scalability and anti-noise.

Keywords: Neural Networks, Sea Surface Temperature, Matlab, Big Data.

I. INTRODUCTION

The dynamics of most global process under investigation have a long time range. That properly requires good data sets. Data here are represented by means of historical time of temperature values. Since the recovery itself serves not only to restore important lost information but also to process retrieved data well, So, to further solve such problems which has impact on the whole application are of time series analysis. The goal of time series analysis is to ascertain and describe the nature of sources that produce signals as rule, the information consisting in a time series frequently contains different processes with different scales of coherence. The Fourier transform is global and describes the overall regularity of signals, but it is not well adapted for finding the location and distribution of singularities. By decomposing signals into elementary building blocks, which are well localized both in space and time, the wavelet transform can characterize the local regularity of signals. Real record is contaminated with noise which is rarely addictive. Gaussian or white. Often, noisy data are generated by nonlinear chaotic processes, which can produce time series having periodic, quasi periodic and chaotic patterns. All of these factors lead to complex, nonlinear, and non-stationary time series. The investigation of such data is not trivial.

However, almost all methods of time series analysis, whether traditional linear or nonlinear, must assume some kind of stationary have been proposed in the literature. If it is absent, one could expect that the differentiation of time series can remove non-stationary in the mean. Another way is to divide data into segments over which the process is essentially stationary and then use the wavelet scale spectrum to estimate the parameters of the time series. Algorithms for estimating these quantities are available. If we cannot assume the existence of underlying low dimension dynamics, we can use the Neural Network time series. The Hadoop-based Back Propagation Neural Network (linear or nonlinear) stationary zero-mean process can be decomposed into the sum of two non-correlated components: deterministic and non deterministic. Besides the short length of historical time series, another substantial problem remains: how to obtain correct result about phenomenon in a time series if there are no fragments of data as it is typical for temperature data. Non--equidistant data distorts even ordinary statistic characteristics.

II. RECOVERING MISSING DATA BY USING HADOOP BASED BACK PROPAGATION NEURAL NETWORK

In this section, we discuss a new non-linear approach to the problem of data recovery. That only one allows obtaining plausible values of missed data. However, the testing of time series with artificial gaps has shown a good result. This method of modeling the data with gaps by using a sequence of curves has been developed. The method is a generalization of iterative construction of singular expansion of matrices with gaps.

III. EXISTING SYSTEM

Traditional methods used for filling of data gaps are not effective for non-stationary and non-linear time series data .In case of missing data and when its location is random there is no solution is provided.

IV. PROPOSED SYSTEM

The SST gap filling (missing data) problem can be solved through Hadoop based Neural Networks. Reason to choose neural networks is due to its massive parallel processing behavior and to handle large data sets we use Map Reduce functionality, hence time consumption is reduced, increase in effectiveness, processing speed increases. Data loss, is a substantial problem for obtaining reliable results from computing. We suggest an approach to solve the above mentioned problem i.e., recovery of missing data in time series using artificial neural networks. The time series data is reconstructed using missing data points obtained from the neural networks model.

V. DATA PROCESSING WITH HIVE

Hive is a database technology that can define databases and tables to analyze unstructured data. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy .We used Hive database to format unstructured data.

A. Hive Queries

Query Snippets:

- create table SST_DATA table (col_value STRING);
- LOAD DATA INPATH '/home/hadoop/training/hive/SSTdata.csv' OVERWRITE INTO TABLE SST_DATA ;
- create table SSTdata (U1 string, U2 string, WV string, CLW string, RR string);
- insert overwrite table SSTdata SELECT
  - regexp_extract (col_value, '^(?:([^,]*)\,?){1}', 1) U1 string,
  - regexp_extract (col_value, '^(?:([^,]*)\,?){2}', 1) U2 string,
  - regexp_extract (col_value, '^(?:([^,]*)\,?){3}', 1) WV string,
  - regexp_extract (col_value, '^(?:([^,]*)\,?){4}', 1) CLW string,
  - regexp_extract (col_value, '^(?:([^,]*)\,?){5}', 1) RR string,
  - regexp_extract (col_value, '^(?:([^,]*)\,?){6}', 1) SSTdata string from SST_DATAtable;

VI. METHODOLOGY

ANN is a massively parallel-distributed processor that has a natural propensity for storing the experimental knowledge and making it available for further use. It resembles the human brain whose speed and efficiency has been always fascinating to researchers for quite a long time. The quest to understand these processes and to solve the associated problems has led to the development of ANN technique. Neural networks essentially involve a nonlinear modeling approach that provides a fairly networks essentially involve a nonlinear modeling approach that provides a fairly accurate universal approximation to any function. Its power comes from the parallel processing of the information from data. No prior assumption of the model is required in the model building process. Instead, the network model largely determined by the characteristics of the data. Single or multiple hidden layer feed forward network is the most widely used model from time series modeling and forecasting. The back propagation network (BPN) is one of the neural network algorithm which is formalized by parker, (1986), Lippmann (1987) and Rumelhart &McClelland (1986) etc. It has been extensively used for inversion, prediction that consists of two passes: a forward pass and a backward pass. In the forward pass the input is applied to input layer and its effect is propagated through network, layer by layer.

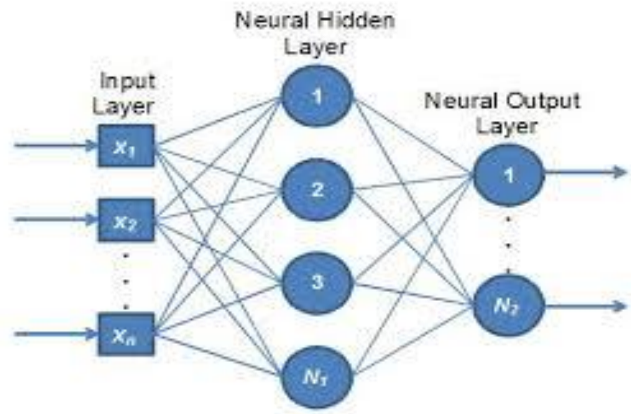

Fig.1. Neural Network Architecture.

The net effect is computed as the weighted sum of squared deviation of the output of the neurons of the previous layer. The sum of squared deviation of the output from the target value at the nodes of the output layer defines the error signal that is to be propagated back to previous layers such the parameters are adjusted to minimize the error in further computations.

Back propagation (BP) algorithm is used for training artificial neural networks. Training usually carried out by iterative updating of weights based on the error signal. The negative gradient of a mean squared error function is commonly used. In the output layer, the error signal is the difference between the described and actual output layer, the error signal is the difference between the desired and actual output values, multiplied by the slope of a sigmoid activation function. Then the error signal is back-propagated to the lower layers. BP is a descent algorithm, which attempts to minimize the error for each epoch. The weights of the network are adjusted by the algorithm such that the error.

VII. NEURAL NETWORK ARCHITECTURE

In information technology, a neural network is a system of programs and data structures that approximates the operation of the human brain. A neural network usually involves a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory. Typically, a neural network is initially “trained” or fed large amount of data and rules about data relationships. Neural Networks represent a class of nonparametric adaptive models and the important issues are to evaluate the performance of the model, here Fig.1 demonstrates the architecture of Neural Network. Performance is achieved by splitting the input data into three sets: training, testing and validation. The parameters (SST, U1, U2, WV, CLW, and RR) of the network are computing using the training datasets. The data learning stopped when the error is minimized and the network is evaluated to the input data from the testing set.

VIII. RESULTS AND DISCUSSIONS

Collecting and preparing sample data is the first step in designing ANN models. The temperature data for a period of 15-year from 1997 to 2012 are collected. After data collection, three data preprocessing procedures are conducted to train the ANNs more efficiently. These procedures are: (1) To solve the problem of missing data, (2) normalize data and (3) randomize data. The missing data are replaced by the average of neighboring values during the same day. Training data samples are randomized by using the function "randperm". This function returns a random permutation of the 5649 integers (training samples) while the order of columns is kept unchanged (SST, U1, U2, WV, CLW, RR). Next, the input variables (SST, U1, U2, WV, CLW, and RR) and output variable (SST) are specified for training and testing. The first five columns represent the inputs while the last column gives the output data (target). The total number of samples is 5649 and in this, training samples are 3955 while the testing samples are 847 and validation sample are 847.Artificial Neural Network technique is used to forecast the SST data based on the previous observed SST data. We considered the thirteen years of SST from 1997 t0 2012 and developed the model.

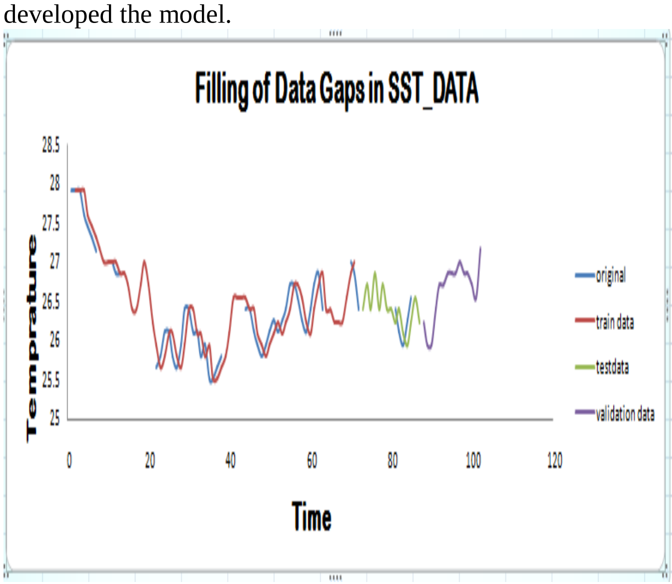

Fig.2. The filling of data gaps in SST data represents the Blue curve. Original data as training (Red curve) from 1 to 70, testing (Green curve) is from 71 to 85 and validation (violet curve) is from 85 to 100.

The model consists of one variable as input parameter and obtained one output variable. Out of thirteen years data we considered 9 years data as the training data to obtain the model and then used 2 years data for testing purpose and 2 years data for validation. We considered the sample is 100, in this case taken training samples are 70 while the testing samples are 15 and validation samples are 15. The model results are show in fig.2 with training, testing and validation data. The blue lines in the figure1 show the original SST data, red line is for training, green line is for testing and violet line is for validation. We considered the sample of 300 values in total , in this case for training are 210 values and the testing values of 45 and validation values of 45. The model results are show in fig.3 with training, testing and validation data. The blue lines in the figure show the original SST data, red line is for training, green line is for testing and violet line is for validation.

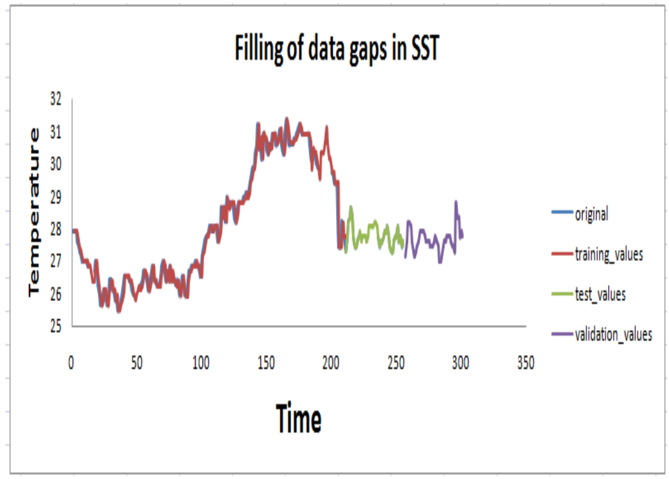

Fig.3. The filling of data gaps in SST data represents the (Blue curve) is the original data as training (Red curve) from 1 to 210, as testing (Green curve) is from 211 to 256 and validation (violet curve) is from 257 to 300.

We considered the sample data of 1000 values , in this case taken training values of 700 while the testing values of 150 and a validation values of 150. The model results are show in fig.4 with training, testing and validation data. The blue lines in the figure show the original SST data, red line is for training, green line is for testing and violet line is for validation.

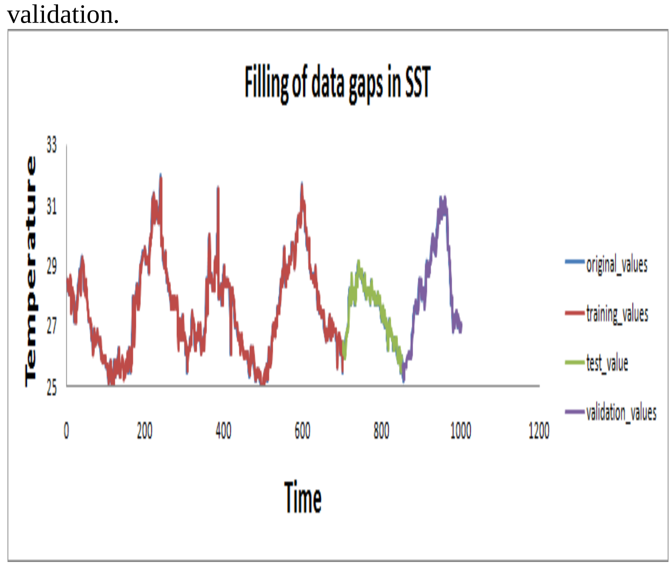

Fig.4. The filling of data gaps in SST data represents the (Blue curve) is the original data as training (Red curve) from 1to 700, as testing (Green curve) is from 701 to 850 and validation (violet curve) is from 851 to 1000.

IX. CONCLUSION

The ANN model developed with five input parameters i.e. (SST, U1, U2, WV, CLW, RR) and the previous SST data is considered as the input to forecast further SST data. The SST data at different times with different intervals is used to develop the model. The obtained values are well matched with the original values. Results indicate ANN technique is useful and well constrained in forecasting the SST values. The main objective of this study is to retrieve missing values in the measured sea surface temperature time series data. The proposed Hadoop based neural networks (HBNN) is well trained and tested in filling the SST data gaps with possible error estimation and correction. Our results show that the efficacy of the estimation procedure and thus the reliability of the estimated missing values are dependent on a number of factors.

X. ACKNOWLEDGEMENTS

The authors are grateful to the Director, INCOIS, Hyderabad for his kind permission and the support of Venkat Seshu Reddam, Sci ;D', INCOIS helped us to publish this research work.

XI. REFERENCES

[1]. “Using MATLAB to Develop Artificial Neural Network Models for Predicting Global Solar Radiation in Al Ain City – UAE “-Maitha H. Al Shamisi, Ali H. Assi and Hassan A. N. Hejase United Arab Emirates University United Arab Emirates.

[2]. “Big Data Analy tics U sing Neural network” - Chetan Sharma, 4-1-2014.

[3]. “Modular Approach to Big Data using Neural Networks” - Animesh Dutta, 4-1-2013.

[4]. “Recovering data gaps through neural network methods”-A. Gorban and A. Rossiev-Institute of Computational Modeling, Krasnoyarsk, Russia-N. Makarenko and Y. Kuandykov -Institute of Mathematics, Almaty, Kazakhstan.

[5]. Dergachev-Ioffe Physical and Technical Institute, St. Petersburg, Russia.

[6]. Schalkoff, R.J. (1997). Artificial neural network. New York: McGraw-Hill.

[7]. Fausett, L.V.(1994). Fundamentals of neural networks: architectures, algorithms, and applications. Englewood Cliffs, NJ; Delhi: Prentice-Hall ; Dorling Kindersley.

[8]. Hassoun, M.H. (1995). Fundamentals of the artificial of artificial neural networks. Cambridge, Mass.: MIT Press.

[9]. Kwon, S.J. (2011). Artificial neural networks. New York: Nova Science Publishers.

XI. FIGURE CAPTIONS

- Figure1: Neural network architecture.
- Figure2: The filling of data gaps SST data with 100 values.
- Figure3: The filling of data gaps SST data with 300 values.
- Figure4: The filling of data gaps SST data with 1000 values.